

Dragana M. Ćorić
University of Novi Sad
Faculty of Law Novi Sad
D.Coric@pf.uns.ac.rs
ORCID ID: 0000-0001-5485-5176

AI ETHICS AND ETHICS IN AI – SOME CRUCIAL DIFFERENCES AND DEVELOPMENT POSSIBILITIES*

Abstract: *AI ethics could be defined as the set of guiding principles that all actors using or developing AI should or must use in order to ensure artificial intelligence technology is developed and used responsibly and wisely. This means taking a safe, secure, humane, and environmentally friendly approach to artificial intelligence (in the further text: AI) and its using, as well as using of results given by AI.*

If we take AI Ethics, shortly explained as earlier, as external ethical guidance for users of AI, should there be also – internal ethical guidance for the AI itself, as set of internal moral values that will lead AI in its development and application in everyday use? The questions we are dealing in this paper are: What is the ethical minimum, as a set of ethical principles that every user or person who participates in the development of AI must have, in order for the use of AI to be safe and wise? On what values should this external ethical minimum rest, and does it, once established, have to remain the same or can it be improved and add/subtract values, in accordance with the exploration of the possibilities provided by AI itself? And is it possible for AI to develop a separate consciousness, with the help of its creators, which will protect it from unmoral attempts by users with bad intentions?

Keywords: *AI, ethics, internal morality, external morality.*

* The paper is a result of scientific research done during the project “Legal Tradition and New Legal Challenges”, which is financed by The Faculty of Law in Novi Sad.

THE IMITATION GAME

When in 1950 he asked the question “CAN machines think”¹, Alan Turing probably could not imagine that in the 21st century we will have to establish rules HOW machines can and must think. The work of machines, first of all computers, is an “imitation game”² of human thinking. Why is it a problem for us that these same machines, when imitating us, their creators- hide data, lie, distort data or completely invent them³, when the machines just imitate us? Don’t people themselves do that, as their defense mechanism, which protects them from “attacks” from the outside, but also confirms their image of themselves as great, good and the best? Why would we deprive machines that imitate us of such a defense mechanism? Is it really a defense mechanism or something that corrupts the core of human existence, and in a way, the role of machines taken as human imitation?

In his article entitled “*Computing Machinery and Intelligence*”, Turing also considered the issue of machine learning- that machines must, at some point, begin to learn what, when and how to do, and that a system of rewards and punishments could be applied to them as well, as a way of learning through conditioning, that is also applied to humans. In the case of machines, the above would apply to the “student” (machine) and the use of certain symbols, sent through “unemotional” channels. When something is wrong or poorly done, the “punishment” would be, for example, deleting what was done, while the reward would be – continuation of work. This type of learning, Turing further believed, would lead to the machine itself developing logical thinking, conditioned by experience from previous situations in which it was punished or rewarded for a certain activity.⁴

At some point, it will be important to distinguish what exactly the machine did and what the man did. If the game of imitation (imitating man’s way of thinking) would be very successful, humanity will get, as Turing believed, either a powerful ally or a worst enemy. That is why it become very important to develop both ethical frameworks – external (AI users and those who develop it) and internal (AI itself)⁵.

¹ Alan M. Turing, “Computing Machinery and Intelligence”, *Mind, New Series*, Vol. 59, No. 236 (Oct., 1950), pp. 433-460, Published by: Oxford University Press on behalf of the Mind Association.

² An expression taken from Turing’s article, previously mentioned.

³ This refers to numerous cases of “hallucinations” Chat GPT – fabrication of court precedents, judges’ opinions and even persons who participated in those trials and their statements and opinions. Lawyers who used Chat GPT irresponsibly and unethically were eventually caught and punished with fines or even revocation of their law license. More on that in the paper: Драгана Ћорић, “Вештачка интелигенција и адвокатура – сарадници, савезници или непријатељи”, Мајско саветовање “Савремено право у ери дигитализације и одрживог развоја”, Крагујевац, 2025, 67-81.

⁴ A.Turing, 435-437.

⁵ For example, there are certain criminal acts in the domestic legislation, where the abuse of information and communication technologies constitutes a more severe form of a criminal act. Serbian Criminal Code (Кривични законик, *Службени гласник Републике Србије* бр. 85/2005-

We can say, that Turing predicted the main issues we face in the 21st century: the question of HOW machines should think and work, how can **we differ the work of humans from the work of AI** and where should we stop using AI and try to live the common life as before-without the help of machines.

Following Turing's work, in 1955 a group of scientists proposed conducting an experiment, under strictly controlled conditions, that would last two months and during which the scientists would investigate ways of learning that the machine could later imitate⁶. This so called *Dartmouth experiment* (or Dartmouth Workshop), conducted in 1956, as one of the longest brainstorming sessions, has led to the creation of the term “**artificial intelligence**”, and specific way of thinking and reasoning⁷, later applied in the framework of machine learning. A machine that would learn in such a way was named *artificial intelligence, as opposed to human (non-artificial, natural) intelligence* that creates learning rules for the machine.

Over time, the tasks that were put before the machines, which they had to learn to solve, became more complex, and therefore the definitions of artificial intelligence became more comprehensive. Gradually, we have reached the intersection of two activities: need to develop “consciousness” or “(reasoning) mind” (which is a prerequisite for the ability to understand), and the *capability to compute* (which is the strict implementation of commands without thinking about whether those commands are correct, legal, good, bad, etc.). In this intersection we find basis for equal development of AI ethics and ethics in AI, which will be discussed further in the paper.

DEVELOPING ETHICAL RULES FOR THE USE OF ARTIFICIAL INTELLIGENCE

Digital technologies are taking over the world of people⁸. Although their development during the 20th century seemed slow, burdened by numerous prejudices

30, 88/2005-51 (исправка), 107/2005-171 (исправка), 72/2009-53, 111/2009-36, 121/2012-3, 104/2013-3, 108/2014-3, 94/2016-7, 35/2019-3, 94/2024-390) in its Art.112 defines computer data, computer network computer programme and computer virus; Art. 298-305. deal with criminal offenses against the security of computer data.

⁶ John McCarthy et alia, “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”, 1955, available on : <https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>, link retrieved: 20.10.2025.

⁷ Bringsjord, Selmer and Naveen Sundar Govindarajulu, „Artificial Intelligence“, *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), (eds.Edward N. Zalta & Uri Nodelman), available on: <https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence>, 20.10.2025.

⁸ “People worry that computers will get too smart and take over the world, but the real problem is that they’re too stupid and they’ve already taken over the world“, Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York, 2015, 18. Domingos states that we are living in the age of algorithms. Whatever we do

and doubts, the last two decades have shown an unprecedented rise. The computers, mobile phones, tablets, numerous applications for communication that connect people from different continents, numerous programs for specific activities sovereignly rule our everyday life. If we had doubts before, the pandemic situation in 2020 convinced us how important the development of digital technologies is to us. Communication, education, the economy continued to function in a somewhat modified form, in circumstances that could have completely blocked all life in a large number of countries in the world. On the other hand, digital technologies and their use in medicine have faster detected and isolated the cause of the pandemic and shortened the process of obtaining vaccines. However, five years later we are still struggling to get back into the offline world, as we were before the pandemics. The online world showed us all its advantages: reducing business costs, possibility of constant communication and the possibility that when processes cannot be carried out live, they can be continued online. Returning to what was before is impossible. So we have to work out new mechanisms for our stay in circumstances that are created for us more by digital technologies directed by the will of their creators, than by us as individuals.

Digital technologies were developed as systems to help and support people in their daily work, with the aim of saving time and less engagement of human resources while optimizing results with as little expenditure as possible. People, deprived of endless working hours, would have more time to organize their private life, create a family, have fun, and their hobbies and would return more to themselves. Therefore, the purpose of digital technologies was for people to manage them, and thereby make their work and life easier.

On the other hand, the basic urge of scientists is to reach with their research where many of their predecessors could not even think they would ever reach. In that unconscious race for scientific power and prestige, during the process of conquering new scientific knowledge and developing new ways to use existing knowledge, everything has changed. Instead of people being the exclusive bearers of the development of digital technologies, the opposite seems to be happening – digital technologies are starting to use human resources for their own development, presenting their demands to us.

For example, we considered that by releasing Chat GPT as an AI system, one of the main challenges, set from the beginning of human civilization, was solved: we created someone who knows everything. But the emergence of Chat GPT has opened up the question of responsible use of information stored in this system, precisely by people. Until Chat GPT existed, we didn't even think about the fact

today, some algorithm is probably included in it: in our case, choosing music on Youtube, in our mobile phone, paying our bills electronically or voting. What we are striving for is the master algorithm, which solves everything-faster, cleaner, effortless.

that someone's personal data would remain permanently stored in some system; that they could be abused in various ways and spread further without the consent or knowledge of the owner of that personal data; nor that system, which knows everything, at some point could lie and create false information⁹. So, AI has set before us new challenges that we have to solve with the fact that they are difficult to predict. One of those challenges is the just mentioned creation of false and distorted information: who could actually expect that artificial intelligence, in this case Chat GPT, would behave like any human being in a similar situation – without a special program for such a way of thinking and reacting to stressful situations being directly inserted into it? So, previously mentioned is telling us that we must know “what we should do with these systems, what the systems themselves should do, what risks they involve, and how we can control all these”¹⁰.

A lot of authors have tied AI exclusively to machine learning. But, it turns out that AI is much more of that. We find that AI is:

*“the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings”*¹¹;

*“AI is the field devoted to building artifacts that are intelligent, where ‘intelligent’ is operationalized through intelligence tests (such as the Wechsler Adult Intelligence Scale), and other tests of mental ability (including, e.g., tests of mechanical ability, creativity, and so on)”*¹²;

“‘AI system’ means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment,

⁹ The protection of personal data is no longer simple and is not solved by a simple recommendation “don’t give personal data to anyone you don’t know”, because there is no longer enough good protection, every firewall can be broken, every database hacked, and someone’s personal data used in such a way that it becomes increasingly difficult to prove that AI and its irresponsible use is not involved. Kanan Naghiyev, “Chat GPT From a Data Protection Perspective”, Baku State University Law Review, Vol. 10:1, 2024, 1-34. Also, important document is: European data protection Board, “Report of the work undertaken by the Chat GPT Taskforce”, 23 May 2024, available on : https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf, 20.10.2025.

¹⁰ Vincent C. Müller, “Ethics of Artificial Intelligence and Robotics”, *The Stanford Encyclopedia of Philosophy (Fall 2025 Edition)*, (eds. Edward N. Zalta & Uri Nodelman), <https://plato.stanford.edu/archives/fall2025/entries/ethics-ai/>, 20.10.2025.

¹¹ *Britannica*, internet encyclopedia, available on: <https://www.britannica.com/technology/artificial-intelligence>, 20.10.2025. It explains also that “*The term is frequently applied to the project of developing systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience*”, but despite the huge development, “there are as yet no programs that can match full human flexibility over wider domains or in tasks requiring much everyday knowledge”.

¹² Newell, N., 1973, “You Can’t Play 20 Questions with Nature and Win: Projective Comments on the Papers of this Symposium”, in *Visual Information Processing*, W. Chase, ed., New York, NY: Academic Press, pp. 283–308. Cited as mentioned in article of Bringsjord and Sundar Govindarajulu.

*and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*¹³.

As we can see, mentioned definitions deal with certain aspects of AI and its development, not always covering everything – because the development of AI is constant, the structure of AI is becoming more complex and its application is expanding into fields in which it had not been before.

Some authors define key ethical issues, arising by expanding the use of AI, such as: privacy and manipulation, opacity and bias, human-robot interaction, employments, and the effects of autonomy¹⁴. On the other hand, when AI systems are considered as subjects, it is important to consider issues of machine ethics and artificial moral agency, as well as the question of “singularity” of superintelligence – once the AI become super-intelligent and invincible. Due to the number and scope of the debates on development and use of AI, we limit ourselves to the consideration of the basic characteristics of the relationship between ethics and artificial intelligence. Therefore, in the rest of the article, we explore only the relationship between AI on the one side, and morality and ethics on the other. **Their relationship is twofold: we distinguish external ethics and morality in the use of AI**, which is more human ethics and morality that leads to responsible and smart use of AI, and **internal ethics and morality within AI itself**, as a set of moral rules that should be developed and implemented within AI itself and which, when human moral compasses do not work well, will be a braking mechanism for people and prevent them from doing immoral things. In the following, we will call these two phenomena AI Ethics and Ethics in AI and will consider them in separate chapters.

AI ETHICS -POSSIBLE DEFINITION, VALUES AND PRINCIPLES TO RELY ON

Usually, AI Ethics is defined as “the moral principles that guide companies toward responsible, fair development and use of AI”¹⁵ which means “taking a safe, secure, humane, and environmentally friendly approach to AI”¹⁶. In this definition, as actors are entitled academics (Researchers and professors), government, inter-

¹³ *Regulation (EU) 2024/1689, or AI Act*. European Commission issued then *The Guidelines on the AI system definition*, to explain the practical application of the legal concept, as defined in The AI Act. Some provisions have started to apply from the 2nd February 2025, and the rest will be applicable from the second part of 2026.

¹⁴ Ibid.

¹⁵ “AI Ethics: What It Is, Why It Matters, and More “, *Written by Coursera Staff*, updated on Sep 16, 2025, available on: <https://www.coursera.org/articles/ai-ethics>, 20.10.2025.

¹⁶ Ibid.

governmental entities (the United Nations, UNESCO and the World Bank are responsible for raising awareness and drafting agreements for AI ethics globally), non-profit organizations and private companies. UNESCO, as one of the mentioned actors, brought the most comprehensive document that covers the widest possible list of principles and values that everyone must be guided by in the application of AI, entitled ***Recommendation on the Ethics of Artificial Intelligence*** (in further text: *Recommendation*). Adopted on November 23, 2021, it was accepted by 194 countries, which makes it the first global document in this sphere of UNESCO's activities¹⁷.

This is not UNESCO's first activity regarding the issue of use of AI. ***Preliminary Study on the Ethics of Artificial Intelligence***, prepared in 2019 by the UNESCO World Commission on Ethics of Scientific Knowledge and Technology (COMEST) was first milestone on this path¹⁸. It depicted raising concern for implications of use of AI in culture and cultural diversity, education, scientific knowledge, and communication and information, as well as for issues of gender equality, preserving peace and protection of environment¹⁹. The Recommendation is natural the result derived from *Preliminary Study* and numerous other documents of UNESCO itself and other organizations with global influence, and deserves more detailed presentation.

The Recommendation rests on the belief that

*“globally accepted ethical standards for AI technologies, in full respect of international law, in particular human rights law, can play a key role in developing AI-related norms across the globe”*²⁰, and that

*“AI technologies can bring important benefits, but that achieving them can also amplify tension around innovation, asymmetric access to knowledge and technologies, including the digital and civic literacy deficit that limits the public's ability to engage in topics related to AI, as well as barriers to access to information and gaps in capacity, human and institutional capacities, barriers to access to technological innovation, and a lack of adequate physical and digital infrastructure and regulatory frameworks”*²¹.

This means that UNESCO has fully recognized the transformative role of AI²², which is creating a new chapter of the information society, and decided to pay due

¹⁷ United Nations Educational, Scientific and Cultural Organization, *Recommendation on the Ethics of Artificial Intelligence*, 2022, available on: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>, 20.10.2025.

¹⁸ UNESCO & World Commission on the Ethics of Scientific Knowledge and Technology, *Preliminary study on the ethics of artificial intelligence*, Paris, 2019.

¹⁹ Ibid.

²⁰ Preamble of the Recommendation.

²¹ Ibid.

²² Although it had good intention at the beginning, AI and its rapid development are changing inevitable everything we know about people, societies, communication, education, environment, etc. Human rights are more than ever under the threat of their violation, in ways we couldn't even

attention to it, integrating all the relevant factors that must previously exist so that AI can be used in a safe and responsible manner.

When drafting this document, UNESCO had in mind some crucial things, such as: the character of AI systems, which are information-processing technologies, who are tasked with integrating models and algorithms in order to produce a capacity to learn and to perform cognitive tasks leading to certain outcomes, such as prediction of future events and decision-making; ethical questions regarding AI which should be raised in all stages of the AI system life cycle; and that those ethical issue are influencing on different policies regarding employment and labour, social interaction, health care, education, media, access to information, digital divide, and other.”In the long term, AI systems could challenge humans’ special sense of experience and agency, raising additional concerns about, inter alia, human self-understanding, social, cultural and environmental interaction, autonomy, agency, worth and dignity. AI reshapes all crucial areas of our life, especially: education, science, cultural identity and diversity, communication and information”, etc.²³. We are all involved in the life cycle of AI development, and vice versa, AI is increasingly affecting the human life cycle. Only the synergy of both worlds, as UNESCO correctly recognized, with mutual knowledge and understanding can contribute to equal development and will benefit both parties.

So, the need for ethical guidance to all AI actors, including the public and private sectors, is strongly expressed in this document. Through defining values and principles on which AI ethics must rely on, is possible to respond more quickly and easily to many challenges and questions posed by the development and application of AI in various spheres of life, and to establish rules that will be simple, easily applicable and widely accepted.

According to UNESCO (and the Recommendation), **the values that AI ethics should rest upon** are²⁴:

- ***Respect, protection and promotion of human rights and fundamental freedoms and human dignity.*** No human being or human community should be harmed or subordinated, whether physically, economically, socially, politically, culturally or mentally during any phase of the life cycle of AI systems. People can interact with AI, gain some help and support, but must not never be subordinated to AI, in a way that their dignity or human rights would be violated or abused²⁵.

imagine a decade ago. More on this in: Ana Čović, “Transformative impact of Chat GPT program on education: protection of intellectual property rights and human rights in the conditions of accelerated digitalization, in: ”Education through the COVID-19 pandemic”, Vol. 2: Socio-Humanistic Aspects, 2024, 659–672.

²³ Ibid.

²⁴ All values listed as in the *Recommendation*.

²⁵ Creating video clips, voice messages and other content using AI for the purpose of fraud, blackmail or violation of someone’s privacy, dignity and reputation have become criminal of-

- ***Environment and ecosystem flourishing.*** Even the greatest benefit that we as humanity can get from the development of AI must not be a factor in jeopardizing ecological stability and sustainability, and not to threaten the even more fragile structure of the ecosystem, which is already struggling with numerous challenges.
- ***Ensuring diversity and inclusiveness.*** AI must be available to all under the same conditions, without any discrimination on basis of race, color, descent, gender, age, language, religion, political opinion, national origin, ethnic origin, social origin, economic or social condition of birth, or disability and any other grounds.
- ***Living in peaceful, just and interconnected societies.*** The value of peace is fully understood only when one goes through a period of war, unrest and conflict. Also, the loudest cry for justice comes from those who have felt injustice and unequal access to their rights, as well as the need to live in a community and the need to belong to a certain circle of people is most strongly expressed in people who at some point in their lives have felt loneliness. AI therefore needs to play an integrative role, correct injustices and create conditions for coexistence of all diversities.

The Recommendation also defined **principles on which new AI society must be based on:**

- ***Proportionality and Do No Harm.*** Means that AI system should not infringe any core values or violate human rights in any way, and that it's use must be proportional to achieve a given legitimate aim²⁶;
- ***Safety and security.*** Prevention of any safety or security risks is obligated, especially when it comes to privacy of individuals;
- ***Sustainability, understood as a synergistic realization of all aims in human, social, cultural, economic and environmental dimensions.*** Without any discrimination and having in mind all aspect of modern life. It means that use of AI and the benefits we get from using AI are greater than the costs invested in its development and use. The sustainability of the system also depends on the proper allocation of all available resources;

fenses in a large number of countries. Proving the truth in this case is left to digital forensics experts who determine whether some content is real or created by AI. This is known as *deepfake*, we list some works about this topic: Gauri Jaiswal et alia, "Deepfake Technology: The Threat of AI-Generated Misinformation", *International Journal of Research Publication and Reviews*, Vol (6), Issue (3), 2025, 930-934; Sami Alanazi and Seemal Asif, "Understanding Deepfakes: A Comprehensive Analysis of Creation, Generation, and Detection", *Artificial Intelligence and Social Computing*, Vol. 72, 2023, 216–224.,

²⁶ This principle is derived from the value of *Respect, protection and promotion of human rights and fundamental freedoms and human dignity*.

- ***Fairness and non-discrimination.*** The benefits of AI technologies are available and accessible to all, without limitations²⁷.
- ***Right to Privacy, and Data Protection.*** Design and implementation of AI systems should be expressed in a way to ensure that personal information is protected throughout the life cycle of the AI system.
- ***Human oversight and determination.*** AI and its use must be human centered, prioritizing human needs, values, ethics, and well-being, and augmenting human capabilities, also understandable for real, average people;
- ***Transparency and explainability.*** Full information about the participation of AI in all spheres of public life must be always present, leaving a clear choice for each person whether AI will be used or not, and explaining clearly and precisely every step and the consequences of using AI, in a way that is easily understandable to the average citizen.
- ***Responsibility and accountability.*** Refers to the respect, protection and promotion of human rights and fundamental freedoms, and the protection of the environment and ecosystems.
- ***Awareness and literacy.*** All members of society can take informed decisions about their use of AI systems and be protected from undue influence. This principle derives from the principle of transparency and explainability.
- ***Multi-stakeholder and adaptive governance and collaboration.*** Constitutes obligation of states to respect international law and national sovereignty in the use of data, and to cooperate with each other on the implementation plan of all relevant documents.

The comprehensiveness of those principles, on which the new world created with the use of AI should be based, represents the determination not only of UNESCO but also of the entire international community to seriously deal with the challenges that AI brings. The principles are actually strategic goals, which are equally legal and ethical by nature – they represent the basis for the adoption of umbrella international and national regulations as well as ethical guidelines for the behavior of different categories of AI user.

The next important document is ***Ethics Guidelines for Trustworthy Artificial Intelligence*** which was presented in 2019. It stated that trustworthy AI should be: ***lawful*** – respecting all applicable laws and regulations; ***ethical*** – respecting ethical principles and values; and ***robust*** – both from a technical perspective while taking into account its social environment²⁸ upon which they are based in the EU Charter.²⁹

²⁷ More about this in: Yuval Shany, “A Right to a Human-to-Human Interaction | Ethics in AI”, *Institute for Ethics in AI*, Oxford, 2024. Available on: <https://www.oxford-aiethics.ox.ac.uk/blog/right-human-human-interaction>, 20.10.2025.

²⁸ High-Level Expert Group on Artificial Intelligence (European Commission), “Ethics Guidelines for trustworthy AI, B-1049 Brussels, Document made public on 8 April 2019, 2.

²⁹ Ibid, 15-18.

Unlike the *Recommendation*, which has more specific values and principles on which the ethical use of AI should rest, *Ethical guidelines* recognizes only four **principles for making AI trustworthy: respect for human autonomy, prevention of harm, fairness and explicability**. The rest falls into the category of “**key requirements for Trustworthy AI**”: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being and accountability (which are recognized in Recommendations mostly as principles of ethical use of AI).

Specificity of this document is creation of **Assessment list**- lists of questions which should be asked when investigating whether AI is trustworthy. For example, when investigating whether requirement of diversity, nondiscrimination and fairness is fulfilled, there should be asked questions such as: did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design; did you assess and acknowledge the possible limitations stemming from the composition of the used data sets; did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases?³⁰ Following the logic of the answers, it is easier to conclude whether AI is used due to imposed requirement, or some adjustments or corrections must be done.

Important to mention is **The Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024**³¹. This Regulation entered into force on the twentieth day following that of its publication in the Official Journal of the European Union, but it shall apply from 2 August 2026. The Regulation is the second global document dealing with use of AI and calling for the ethical application of AI as it is “a tool for people, with the ultimate aim of increasing human well-being”. Art.1 of the Regulation states that aim of this document is “to improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union and supporting innovation”, and sets a lot of rules for ethical use of AI in EU common market, as well as for markets in member states.

In another ethical guide, under the name **Understanding AI Ethics and Safety** developed by Dr. David Leslie at The Alan Turing Institute’s Public Policy Programme, was made strong emphasize on responsible design and implementation

³⁰ Ibid, 32.

³¹ The regulation is known also as AI Act, which is laying down rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). The introductory part of this act has 180 articles, and the pure legal part has 113 rather long articles, and 13 annexes.

of AI systems in the public sector. As for Values that should AI rest upon, they identified only three: **Support, Underwrite, Motivate**. The realization of these values creates an environment in which individual dignity and autonomy can be fully respected, while fostering sincere and inclusive interpersonal relationships; also to the wellbeing of all stakeholders, to promotion of social values and public interest should be given special attention.

Common to all presented documents is a clear vision and intention to regulate the development and application of AI in a responsible and sustainable manner. Also, all documents emphasize the need for an approach focused on people's well-being, which means preventing AI abuses and human rights violations. As the time of application of The Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 is approaching, we expect the intensification of work on the creation of new documents and guidelines for the legal and ethical application of AI.

ESTABLISHING THE MORAL COMPASS INSIDE AI – IS IT POSSIBLE?

In order to avoid the dystopian scenario where robots that have completely taken over the world and use people as their slaves, which was shown in the film franchise Matrix, we may say there is the need for the development of moral and legal awareness within AI itself. The more autonomous AI systems are, the more dangerous our life can be³². Leaving decision-making to artificial intelligence slows down the development of human thinking, and feeds artificial intelligence with experiences that at some point can make it more powerful than its creators. Reducing learning, delegating numerous processes to artificial intelligence, with the original intention of freeing people from numerous tasks and burdens, inevitably leads to laziness and leisure.

In his latest video, posted on Youtube, Johannes Himmelreich³³ states that “Some of ethical concerns we associate with AI are actually problems that have been there all along”. AI is just kind of magnifying glass which more underlines most of the problems we already have, but are not brave enough to cope with them directly, such as lies of others, cheating or else. Also, he states that AI would be much like ourselves. So, it seems that creating ethical AI, would be “a hell of a task”, because we will try to create something that is better and greater from us and somehow perfect.

³² R. Picard rightly claims that “the greater the freedom of a machine, the more it will need moral standards”. Rosalind Picard, *Affective Computing*, Cambridge, MA and London, MIT Press., 1997, 19.

³³ Prof. Johannes Himmelreich, *What are the ethical concerns of AI?*, <https://www.youtube.com/watch?v=j6fH5XuSQu8>, 20.10.2025.

But, “we are not ideal beings!”, declares Susan Anderson, one of the creators of machine ethics³⁴. Since we are not ideal, do we even have the possibility to create something better than ourselves, as Anderson further suggests? We know the saying that the student can surpass the teacher – of course, when the teacher allows him/her to do so. But what if the teacher does not know which perspectives he should open to his student, and in which ways to support the student’s development? Having in mind that, most of the researches dealing with AI in general suggest the human centered approach to use and development of AI, it seems that this teacher(human society) will never allow the student (AI) to develop and become a better than the teacher himself.

So what should be an ethical minimum embedded within AI itself? Is the recognition of that ethical minimum possible by applying standardized commands built into the AI system, or is it necessary to do something more? It seems that the best way to live consciousness within AI is by imitating the way it is done in the human mind. But how does conscience arise in people? We must bear in mind the fact that morality, as a type of social norms, rests on the understanding that someone is more good or bad, i.e. moral or immoral. In the same way, the understanding of good and bad is changeable- something good today will become bad tomorrow and even a criminal act, while something bad and deviant can become socially desirable. That deviated moral compass complicates situation between reasonable actors- people, individuals³⁵; we can only imagine what confusion would be created inside machines that operate based on the analysis of countless language patterns and mathematical formulas.

People themselves sometimes find it difficult to navigate such changes of socially acceptable, moral behavior/how would AI, a rigid creation that rests on commands, put an end to these changes with a number of sterile explanations? In the end, the very understanding of the need for the existence of machine ethics, as its authors S. Anderson and M. Anderson state, at the beginning of the 21st century seemed like science fiction, and after a quarter of a century it became an absolute necessity³⁶. “I don’t think we should build all human values into autonomous machines, since humans are prone to unethical behavior”, says S. Anderson³⁷. So as we install human values and our way of thinking into machines, we probably will install ways of avoidance of acting, according to those rules and ways

³⁴ M. Anderson et alia, “Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson”, *Conatus* 6, no. 1, 2021, 177-202.

³⁵ Celia Moore, Francesca Gino, “Ethically Adrift: How Others Pull Our Moral Compass from True North, and How We Can Fix It”, *Research in Organizational Behavior*, Volume 33, 2013, 53-77.

³⁶ M. Anderson, et alia, 179.

³⁷ Ibid.

to perfectly logically and ethically justify that avoidance. Life, more precisely, the maintenance of life is the greatest value; won't a person reject all other values and principles in order to maintain his own life? How to incorporate such a way of reasoning in AI, knowing all the flaws in that way of thinking? And should it be installed at all?

It seems that we still haven't cleared up whether in the future we need more machines that will have a completely objectified way of reasoning, without any prejudice or discrimination on any basis, or will we need an additional braking mechanism for all human actions that cross the border of legal or moral, and that will stop a person when no one else can? In the second case, that braking mechanism must have a developed conscience, an understanding of good and bad, as well as legal awareness, which must have all legal systems and laws in the world stored in it – because AI users come from all over the world and AI must be ready to warn each user equally, in accordance with their legal system or cultural heritage.

We are actually asking for a counter service: by establishing the rules for the ethical development and use of AI, we ask that AI behave ethically towards us, in an equal way. Only recent legal acts, with the intention of global application, and which we mentioned earlier in our paper, have turned our promise that we will be good when using AI into binding norms, the violation of which is followed by an appropriate penalty, either national, international or supranational law. Fortunately, "what machines are good at (better than humans) is digesting a lot of data quickly: discovering connections, etc."³⁸. The machines memorize a lot more than we are; they find patterns and create strategies faster than the ordinary human mind can; their performance of tasks more efficiently and saves human resources. If they develop a conscience, and begin to reason adaptively about what is good or bad, they may even become better than us in that regard. And then that dystopia scenario from the film *Matrix* will become real – the rulers (people) will be enslaved by someone whom they hoped would remain subservient to them, but who only behaved like a man in certain circumstances- when he had the opportunity to rule, he used it, without any reservations, having previously suppressed any moral consideration towards those who previously subjugated him.

The previous brief discussion raised a number of questions. One thing is clear, AI also needs an internal moral compass. In this way, it would better imitate the human way of thinking. Furthermore, in situations where a person cannot do it, it would prevent a person from doing something immoral, which results from the determination of AI as human-centered technology.

³⁸ Ibid, 201.

REFERENCES

Literature:

- “AI Ethics: What It Is, Why It Matters, and More“, *Written by Coursera Staff*, updated on Sep 16, 2025, available on: <https://www.coursera.org/articles/ai-ethics>, 20.10.2025.
- Alanazi, S.; Asif, S., “Understanding Deepfakes: A Comprehensive Analysis of Creation, Generation, and Detection“, *Artificial Intelligence and Social Computing*, Vol. 72, 2023, 216–224
- Anderson M. et alia, “Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson“, *Conatus* 6, no. 1, 2021, 177-202.
- Bringsjord, S. ; Sundar Govindarajulu, N. “Artificial Intelligence“, *The Stanford Encyclopedia of Philosophy (Fall 2024 Edition)*, (eds. Edward N. Zalta & Uri Nodelman), available on: <https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence>, 20.10.2025.
- Britannica, internet encyclopedia, <https://www.britannica.com/technology/artificial-intelligence>, 20.10.2025.
- Domingos, P., *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York, 2015
- European data protection Board, “Report of the work undertaken by the Chat GPT Taskforce“, https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf, 20.10.2025.
- Jaiswal, G. alia, “Deepfake Technology: The Threat of AI-Generated Misinformation“, *International Journal of Research Publication and Reviews*, Vol (6), Issue (3), 2025, 930-934
- McCarthy, J. et alia, “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence“, 1955, : <https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>, 20.10.2025.
- Moore, C., Gino, F., “Ethically Adrift: How Others Pull Our Moral Compass from True North, and How We Can Fix It“, *Research in Organizational Behavior*, Volume 33, 2013, 53-77.
- Müller, V.C., “Ethics of Artificial Intelligence and Robotics“, *The Stanford Encyclopedia of Philosophy (Fall 2025 Edition)*, (eds. Edward N. Zalta & Uri Nodelman), <https://plato.stanford.edu/archives/fall2025/entries/ethics-ai/>, 20.10.2025.
- Naghiyev, K., “Chat GPT From a Data Protection Perspective“, *Baku State University Law Review*, Vol. 10:1, 2024, 1-34.
- Newell, N., 1973, “You Can’t Play 20 Questions with Nature and Win: Projective Comments on the Papers of this Symposium“, in *Visual Information Processing*, W. Chase, ed., New York, NY: Academic Press, pp. 283–308. Cited as mentioned in article of Bringsjord and Sundar Govindarajulu
- Picard, .R., *Affective Computing*, Cambridge, MA and London, MIT Press., 1997
- Turing, A.M., “Computing Machinery and Intelligence“, *Mind, New Series*, Vol. 59, No. 236 (Oct., 1950), pp. 433-460, Published by: Oxford University Press on behalf of the Mind Association.
- Ђорић, Д., “Вештачка интелигенција и адвокатура – сарадници, савезници или непријатељи“, *Мајско савештовање “Савремено право у ери дигитализације и одрживог развоја“*, Крагујевац, 2025, 67-81.

Himmelreich, J., *What are the ethical concerns of AI?*, <https://www.youtube.com/watch?v=j6fH5XuSQu8>, 20.10.2025.

Čović, A., “Transformative impact of Chat GPT program on education: protection of intellectual property rights and human rights in the conditions of accelerated digitalization”, in: “*Education through the COVID-19 pandemic*”, Vol. 2: *Socio-Humanistic Aspects*, 2024, 659–672.

Shany, Y., “A Right to a Human-to-Human Interaction | Ethics in AI”, *Institute for Ethics in AI*, Oxford, 2024. <https://www.oxford-aiethics.ox.ac.uk/blog/right-human-human-interaction>, 20.10.2025.

Legal acts and other documents

High-Level Expert Group on Artificial Intelligence (European Commission), “Ethics Guidelines for trustworthy AI, B-1049 Brussels, Document made public on 8 April 2019

Кривични законик, *Службени њласник Републике Србије* бр. 85/2005-30, 88/2005-51 (исправка), 107/2005-171 (исправка), 72/2009-53, 111/2009-36, 121/2012-3, 104/2013-3, 108/2014-3, 94/2016-7, 35/2019-3, 94/2024-390) .

Regulation (EU) 2024/1689

United Nations Educational, Scientific and Cultural Organization, *Recommendation on the Ethics of Artificial Intelligence*, 2022: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>, 20.10.2025.

UNESCO & World Commission on the Ethics of Scientific Knowledge and Technology, *Preliminary study on the ethics of artificial intelligence*, Paris, 2019.

Драјана М. Ђорић
Универзитет у Новом Саду
Правни факултет у Новом Саду
D.Coric@pf.uns.ac.rs
ORCID ID: 0000-0001-5485-5176

Етика (употребе) вештачке интелигенције и етика у вештачкој интелигенцији – неке кључне разлике и могућности развоја

Сажетак: Етика вештачке интелигенције може се дефинисати као скуп водећих принципа које сви актери који користе или развијају вештачку интелигенцију треба или морају да користе како би осигурали да се технолозија вештачке интелигенције развија и користи одговорно и мудро. То значи заузимање безбедној, сигурној, хуманој и еколошки прихватљивој пракси у вештачкој интелигенцији (у даљем тексту: ВИ) и њеном коришћењу, као и коришћењу резултата које даје ВИ.

Ако ВИ етику, укратко објашњену као раније, узмемо као скупа етичке смернице за кориснике ВИ, да ли би требало да постоје и – интерне етичке смернице за саму ВИ, као скуп интерних моралних вредности које ће водити ВИ у њеном развоју и примени у свакодневној употреби? Питања којима се бавимо у овом раду су: Шта је етички минимум, као скуп етичких принципа које сваки корисник или особа која учествује у развоју ВИ мора имати, да би употреба ВИ била безбедна и мудра? На којим вредностима треба да почива овај скупа етички минимум и да ли он, једном када се успостави, мора да остане исти или се може побољшати и додати/одузети вредности, у складу са испољивањем могућности које пружа сама ВИ? И да ли је могуће да ВИ, уз помоћ својих стваралаца, развије посебну свест која ће је заштитити од неморалних покушаја корисника са лошим намерама?

Кључне речи: ВИ, етика, унутрашњи морал, спољашњи морал.

Датум пријема рада: 22. 2. 2026.

Датум прихватања рада: 24. 4. 2026.