UDC 340.1 + 341.1
DOI 10.37772/2309-9275-2024-2(23)-3
JEL Classification: K 10, K 33, O 19

THE EMERGENCE OF THE HUMAN RIGHT TO SAFE ARTIFICIAL INTELLIGENCE

Andrii HACHKEVYCH, ORCID 0000-0002-8494-1937[1]

[1]*Associate Professor, Lviv Polytechnic National University, Lviv, Ukraine*

Corresponding author: Andrii Hachkevych, andrii.o.hachkevych@lpnu.ua

Abstract. Several international agreements on artificial intelligence were adopted last year. There has also been a growing focus on the connections between new technologies and the legal sphere, evidenced by many publications. We may assume that specific rights related to artificial intelligence are emerging alongside digital rights. This article presents the author's perspective on the right to safe AI. Minimizing risks and ensuring that humans maintain control over technology lie at the root of this right. The author explores the various human needs connected to AI systems. Furthermore, he investigates how the right to safe AI may differ from existing human rights established in international law. The proposed new right aims to ensure that the risks associated with developing and using AI systems are expected to be kept to an unacceptable minimum. The findings of this study could be valuable for further exploration of human rights issues stemming from the rise of new technologies. The article briefly defines the human right to safe AI as "the right to safe use of AI systems".

Keywords: right to safe AI; high-risk AI systems; human rights; AI existential risk; UN General Assembly Resolution A/78/L.49.

Author contributions

The author prepared the article independently. The author independently selected the literature, analyzed it and formulated conclusions

Disclosure statement

The author does not have any competing financial, professional, or personal interests in relation to others.

INTRODUCTION

Due to the significant influence of artificial intelligence and other new technologies on society, we should rethink the concept of human rights. Under human rights, we understand objectively defined and tend to develop essential opportunities required for a fulfilling human life.

The law ensures that individuals can meet their needs in the era of artificial intelligence, providing legal protections to mitigate potential harm from technological advancements. This includes adding new rights, such as the right to safe AI.

Following the recent adoption of several international agreements, including UN General Assembly Resolution A/78/L.49, and considering a growing focus on the connections between new technologies and the legal sphere, evidenced by a large number of publications and existing sets of AI ethical principles, this article presents our vision on the scope of the right to safe AI.

THEORETICAL FRAMEWORK

Artificial intelligence is one of the most discussed phenomena especially from the perspective of governance and regulation recently. This involves understanding how the government and other stakeholders should influence the development and use of AI systems. The decisions of policy-makers

play an important role in ensuring that technological progress is conducted responsibly, avoiding negative impacts on human rights, and mitigating existential risks.

Modern legal research has revealed a distinct area focusing on the nature of AI systems. This area does not emphasize artificial intelligence's progressive problem-solving abilities but evaluates its acceptability for use and the legitimacy of its development.

When determining acceptable and legitimate artificial intelligence ideals, we often refer to principles such as transparency, safety, security, reliability, accountability, governability, and trustworthiness. These principles underpin ethical standards and are embodied in the first international agreements on artificial intelligence adopted in 2024.

Among these ideals, the safety of artificial intelligence stands out as a key prerequisite for recognizing a new right: the right to safe AI.

Hendrycks' book is one of the most recent AI safety and ethics studies. It aims to provide a comprehensive approach to understanding AI risk. The author examines the risks inherent in artificial intelligence based on the source-related approach (malicious use, AI race, organizational skills, Rogue AI). A part of the book is devoted to the safety aspects of artificial intelligence: single-agent safety, safety engineering, and complex systems. The book is also important for solving the problem of AI governance, as the author explores some of the existing approaches (Hendrycks, 2024).

Another book discusses the concept of AI assurance, which aims to make the development and application of artificial intelligence more valid, explainable, fair, and ethical. The authors examine how to ensure various AI methods, including machine learning, natural language processing, and predictive analytics (Batarseh & Freeman, 2023). «Ethics of Artificial Intelligence,» edited by Lara and Deckers, addresses the main ethical issues arising from using AI in various fields. Several chapters are dedicated to the issue of AI boundaries, which is very close to the need for control discussed in our article (Lara & Deckers, 2023).

Shneiderman proposes a list of recommendations to bridge the gap between widely ethical principles of Human-centered AI and practical steps for effective governance, promoting safety culture through business management strategies (Shneiderman, 2020). Köse poses a question in the title of the article «Are We Safe Enough in the Future of Artificial Intelligence?» which he considers to explain dystopian scenarios of an AI-engaged future and moral dilemmas relevant to developing good AI systems (Köse, 2018).

In this way, we can see that safe AI is not only about the present and observing human rights but also about possible catastrophic predictions of the future, some of which are indeed wildly exaggerated.

We can also mention some publications on human rights in the context of artificial intelligence and new technologies.

Aizenberg and van den Hoven attempt to present a toolkit for translating fundamental human rights into context-dependent design requirements while promoting transparent, explainable, and fair AI (Aizenberg & van den Hoven, 2020). Risse sets out an agenda for artificial intelligence and human rights that consists of short-, medium-, and long-term issues, the latter being that «humans may have to live with machines that are intellectually and possibly morally superior» (Risse, 2019). Studying the impact of AI systems on human rights, Gordon suggests that states should develop and adopt a special convention for the rights of AI systems. This convention, if implemented, could significantly influence international and national law, shaping new norms and frameworks for AI and human rights (Gordon, 2023). The group of authors, with the participation of Balcerzak and Kapelańska-Pręgowska, examines the international governance of artificial intelligence, looking for accountable and responsible AI consistent with human rights. These authors focus on the efforts of the United Nations, the Council of Europe, and the European Union regarding actual and potential human rights issues caused by AI technologies (Balcerzak & Kapelańska-Pręgowska, 2024).

These and other works emphasize the need for more thorough research on safe AI and its connection with human rights. The results of such research could significantly influence international and national law shortly by establishing new norms.

METHODOLOGY

The research methodology is based on the author's hypothesis that the right to safe AI might be recognized as a distinct human right. To test this hypothesis, the author employs a method of generalization to identify key features of human rights, which will serve as guidelines to determine the extent to which each feature applies to this emerging right.

This systematic approach has been chosen to present the author's perspective on the nature of the right to safe AI, which has been largely influenced by the adoption of recent agreements at the levels of the UN, the EU, the G7, and the Council of Europe.

RESULTS

The pivotal question that sparked the author's curiosity in the present research was: *What conditions signify the emergence of a new right?*

The answer to this question is deeply rooted in the concept of human rights. It hinges on the various interpretations, where an inevitable interdependence exists between those that hinge on the category of a claim (1) as the foundation for the legal content of the right, e.g., "have a justified claim against others that they act in a fitting manner" (Kass, 1993, P. 34), and the emphasis on the inherent nature of these rights to all people (2), "they are rights one possesses, not by virtue of some special status, but simply as a human being" (Wellmann, 1997, P. 15). Renteln defines the classic definition as follows: "a human right is a right that is universal and held by all persons," further quoting Cranston's understanding: "A human right by definition is a universal moral right, something that all men, everywhere, at all times ought to have, something of which no one can be deprived without a grave affront to justice, something that is due to every human being simply because he is human" (Renteln, 1988, P. 347). Tiedemann, in turn, underscores that human rights are those rights whose violation leads to an inhuman state of the bearer of this right (Tiedemann, 2012).

We summarize the key features of human rights to determine the conditions for stating that a new right has emerged.

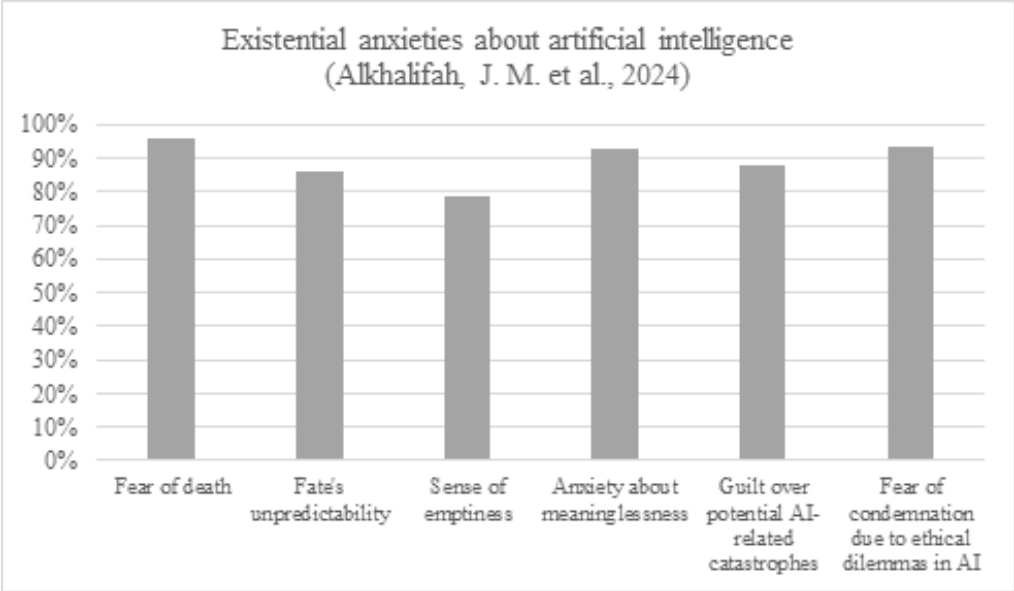
1. Common to all people, which can be considered an axiom in theory. Everyone in the world possesses human rights.

2. Driven by the necessity (or desire) to live. Human rights are needed to fulfill one's needs, although the scope of those needs may vary.

3. The existence of legally binding recognition transforms a theoretical stance into a legal norm. This transformation is often achieved by international treaties, national constitutions, and domestic laws, which acknowledge human rights and provide mechanisms for their enforcement.

Safe AI links to several human needs. First among these is the human-centered desire to control all processes that occur. Not all processes are suitable for such control; for example, humans are powerless in front of natural disasters, although they can try to prevent them or minimize their negative consequences. At the same time, when considering artificial intelligence and other new technologies, it is up to humans to further improve them and determine how and for what purpose they serve. Thus, technological progress in this respect manifests itself differently than the forces of nature; although humans do not fully control the former, it is controlled to a certain extent. And this limit should be highlighted and taken into account.

The potential for AI systems to spiral out of control is a source of anxious uncertainty. This uncertainty will not lead to catastrophic outcomes in the most optimistic scenario. However, it could even pose an existential threat in the worst case. A survey of about 300 respondents revealed existential anxieties related to artificial intelligence (Alkhalifah, J. M. et al., 2024). The implications of this uncertainty are underscored by headlines such as "Our final invention: Artificial intelligence and the end of the human era" (Barrat, 2023).



The need for control is noticeable in two forms: first, where control ensures that the use of AI systems aligns with human expectations and leads to satisfactory results, from instructing Siri to perform a task to executing a flawless text translation (beneficial AI).

Second, regarding human control over non-human intelligence, where the lack of control might lead to existential risk.

In addition to control, people may have many other needs related to artificial intelligence:

– the assurance that the information provided by AI systems can be trusted: whether it is a video of a public figure's statement, a Chat-GPT response to a professional query, or any other AI-generated content, a deed of trust is one of the most important,

– a proper assessment of their knowledge and skills, as well as the work they have done, which is associated with the need for job stability (this does not include specialists in information technology or other areas that are projected to be in demand),

– the confidence that the information entered and the data output remain confidential, no one will have access to them, and the technology will not harm devices or turn off other programs,

– the awareness that decisions concerning their interests or even fate are fair and impartial: some time ago, Amazon used an AI tool for the automatic processing of resumes of job applicants as an experiment, which demonstrated technology bias against women (Dastin, 2022),

– understanding whether communication is with a human or a non-human assistant and whether a human or artificial intelligence is behind diagnoses and recommendations in healthcare.

In a broad sense, all of these needs are somewhere related to human safety, for which the following threats may arise: job losses due to the automation of manufacturing processes, a computer program rejecting a social security application, personal data being posted online, or devices used for personal purposes failing (as for the latter threats, we should not forget that artificial intelligence is also subject to the threats that are typical of technologies in general). Safe AI, in a broad sense, can be explained in the light of minimized risks comparable to the observance of fundamental human rights. Developing and using AI systems is safe if they do not violate human rights.

At the same time, in a narrow sense, safe AI refers to AI systems designed and deployed in a way that does not harm humanity, particularly in an existential dimension.

When discussing the right to safe AI, we should consider both the broad and narrow senses of the term.

In fairness, we can agree with Algan, who sees the process of past development of human rights as consisting of several stages: emergence in theory, incorporation into legal form, and internationalization

of the protection, furthering the protection of existing rights at all levels, and extension of the list of human rights (Algan, 2004, P. 123). Each human right, including those related to artificial intelligence, has a similar evolution.

The right to safe AI is at the beginning of its development. Its foundational aspects include theoretical provisions—some of which are authored by the scholars mentioned above—as well as the ethical principles of artificial intelligence, which are also the result of reflection. These principles significantly influence the application of AI systems and their prior development. Furthermore, AI ethics has been shaping the perception of artificial intelligence, particularly regarding governing and regulating.

The role of safe AI deserves to be discussed within some of the known frameworks of ethical principles.

A. UNESCO Recommendation on the Ethics of Artificial Intelligence contains "Safety and security" in the list of AI principles (UNESCO, 2021). These two values are combined, and their explanation highlights safety risks, mainly unwanted harms and vulnerabilities to attack as security risks. Additionally, some means of ensuring this principle are explained: "Safe and secure AI will be enabled by the development of sustainable, privacy-protective data access frameworks that foster better training and validation of AI models utilizing quality data."

B. Asilomar Priciples (Future of Life Institute, 2017). In this list, safety encompasses security and should be maintained throughout the entire operational lifetime of the system. If applicable and feasible, safety needs to be verified. Interestingly, a separate section of the Asilomar Principles, titled «Longer-term Issues,» emphasizes that risks associated with AI systems—particularly catastrophic and existential risks—should be anticipated and mitigated.

C. US Department of Defense Principles of Artificial Intelligence Ethics (Todd Lopez, 2020) Among the five key principles designed for practitioners, the principle of «reliability» is included. It implies that AI systems should have explicit, well-defined uses, leveraging their available capabilities and ensuring safety, security, and effectiveness throughout their life cycle. Another principle of governability leans towards the safe operation of AI systems. It describes the need to design and develop AI systems that can detect and prevent harmful consequences and to disable or deactivate systems that demonstrate unintended behavior.

We will focus on the UN General Assembly Resolution among the recently adopted agreements, reflecting the tendency of legally binding recognition as a key feature of a new human right.

The UN General Assembly resolution, "Seizing the Opportunities of Safe, Secure, and Trustworthy Artificial Intelligence Systems for Sustainable Development," was adopted without a vote during its 78th session in 2024 (United Nations General Assembly, 2024). This resolution marks a significant milestone and builds on the series of resolutions from 2022-2023, which focused on the observance of human rights, particularly privacy, in the use of information and communication technologies and the light of sustainable development (United Nations General Assembly, 2022; United Nations General Assembly, 2023a; United Nations General Assembly, 2023b; United Nations General Assembly, 2023c).

The 2024 UN resolution is one of the first to specifically address AI governance at the level of international law and approach it to the concept of sustainable development. It highlights the UN system's key accomplishments in addressing AI issues, especially the work of the International Telecommunication Union, the United Nations Educational, Scientific and Cultural Organization, the Human Rights Council, and the United Nations High Commissioner for Human Rights. Additionally, it mentions the Secretary-General's initiative to establish a High-level Advisory Body on Artificial Intelligence, known for publishing important results in its report following the adoption of this resolution (United Nations, 2024).

The resolution emphasizes safe, secure, and trustworthy AI systems for non-military use, considering that they should be human-centered, reliable, explainable, ethical, inclusive, promote human rights and international law, privacy-preserving, sustainable development-oriented, and responsible. AI

systems are expected to contribute to achieving all 17 UN Sustainable Development Goals, facilitating digital transformation, strengthening peace, and overcoming digital divides within countries.

The resolution also highlights the importance of developing and supporting effective governance and regulatory frameworks for AI systems. States and various stakeholders—including the private sector, international and regional organizations, civil society, media, academia, research institutions, and individuals—should collaborate in their respective roles and responsibilities (a solidarity approach is adopted to achieve the desired outcomes).

Notably, the resolution asserts the obligation of member states and other stakeholders to refrain from or cease using AI systems that cannot operate in compliance with international human rights law or that pose undue risks to human rights.

In support of this resolution, Camila Harris, the representative of the state it was proposed by, remarked upon its adoption that artificial intelligence must be adopted and advanced in a way that protects everyone from potential harm and ensures everyone can enjoy its benefits (The White House, 2024).

DISCUSSION

The identified features of human rights allow for the confirmation or rejection of the hypothesis regarding the emergence of a new human right. These rights are understood as legally enshrined opportunities essential for meeting human needs.

While brief designations describe human rights, their normative content—outlined in a few sentences—possesses a depth and breadth that far exceeds those sentences.

In proposing the «right to safe AI» designation for this new right, we understand there is room for discussion on its name and nature. Additional research in this domain may also cover the potential for associated rights to «trustworthy AI,» «secure AI,» and «reliable AI.»

Moreover, the relationship of this proposed right to safe AI to other established and universally recognized human rights—particularly those found in the International Bill of Human Rights—is worth investigating. It brings to mind the ongoing discourse surrounding digital rights and their position within the human rights system. The discussions center on the question of whether these digital rights represent new (fourth-generation) rights or updated measures for realizing existing rights.

On the one hand, human needs are safeguarded by other rights—such as the rights to information and labor—and are encompassed within the concept of safe AI, while existential risk relates to the right to life. Meanwhile, there are some concerns about the ability of existing rights to ensure that individuals can effectively protect their interests and obtain compensation for harm arising from the development and use of AI systems. Ultimately, the emergence of this new right indicates its growing importance and suggests that its enforcement may be rendered more effective.

Undeniably, safe AI aligns with human needs through its applications, which do not infringe upon human rights and do not precipitate adverse consequences for individuals. Additionally, it promotes the kind of development that does not pose threats to humanity or large groups of the population.

Safe AI encompasses nearly all ethical principles related to artificial intelligence, asserting that technological advancements should provide benefits rather than harm. The focus on risks narrows the concept of the right to safe AI compared to other potential human rights concerning AI, emphasizing the importance of using AI systems with minimal risks.

CONCLUSIONS

The growing impact of artificial intelligence and modern technologies on society is enhancing the concept of human rights. The study results confirm the hypothesis that the right to safe AI might be recognized as a distinct human right. The provisions of the UN General Assembly resolution cited as

one of the «pieces of evidence» indicate that this right is expected to become universally recognized in some time and will not be just a proposal of the author of this or any other article.

In addition to the UN, important agreements have been reached within the EU (the AI Act, which differentiates AI systems by risk levels, thereby guaranteeing safe AI), the G7 (the Hiroshima AI Process initiated important restrictions on artificial intelligence for companies), and the Council of Europe, which adopted the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law last year.

The right to safe AI could be explained as the right to safe use of AI systems. Its meaning lies in guaranteeing that *no harm will be caused and that no other undesirable consequences will occur due to such use, and the development of AI systems should not threaten humanity.* Such concerns, however, often contribute to widespread scepticism regarding emerging technologies, frequently expressed by intellectuals through initiatives like the Bletchley Declaration («AI should be designed, developed, deployed, and used, in a manner that is safe, in such a way as to be human-centric, trustworthy and responsible") and the open letter "Pause Giant AI Experiments" ("Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources») (UK Government, 2023; Future of Life Institute, 2023).

REFERENCES

Aizenberg, E. & Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720949566

Algan, B. (2004). Rethinking «Third Generation» Human Rights. *Ankara Law Review, 1*(1), 121–155.

Alkhalifah, J. M., Bedaiwi, A. M., Shaikh, N., Seddiq, W. & Meo, S. A. (2024). Existential anxiety about artificial intelligence (AI) – is it the end of humanity era or a new chapter in the human revolution: Questionnaire-based observational study. *Frontiers in Psychiatry, 15*, art. 1368122. https://doi.org/10.3389/fpsyt.2024.1368122

Balcerzak, M., and Kapelańska-Pręgowska, J. (2024). *Artificial Intelligence and International Human Rights Law.* Elgar Publishing.

Barrat, J. (2023). *Our final invention: Artificial intelligence and the end of the human era*. Quercus Editions Ltd.

Batarseh, F., and Freeman, L. (2023). *AI Assurance: Towards Trustworthy, Explainable, Safe, and Ethical AI*. Academic Press.

Dastin, J. (2022). Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women. In K. Martin (Ed.). *Ethics of Data and Analytics: Concepts and Cases* (pp. 296–299). Auerbach Publications.

Future of Life Institute. (2017). *Asilomar AI Principles*. https://futureoflife.org/open-letter/ai-principles/

Future of Life Institute. (2023). *Pause Giant AI Experiments*: An Open Letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Gordon, J. (2023). *The Impact of Artificial Intelligence on Human Rights Legislation. A Plea for an AI Convention*. Palgrave Macmillan.

Hendrycks, D. (2024). *Introduction to AI Safety, Ethics and Society*. Taylor & Francis.

Kass, L. (1993). Is There a Right to Die? *The Hastings Center Report, 23*(1), 34–43.

Kцse, U. (2018). Are We Safe Enough in the Future of Artificial Intelligence? A Discussion on Machine Ethics and Artificial Intelligence Safety. *Broad Research in Artificial Intelligence and Neuroscience*, *9*(2), 184–197.

Lara, F. & Deckers, J. (Eds.). (2023). *Ethics of Artificial Intelligence*. Springer.

Renteln, A. (1988). The Concept of Human Rights. *Anthropos, 83*(4/6), 343–364.

Risse, M. (2019). Human Rights and Artificial Intelligence: An Urgently Needed Agenda. *Human Rights Quarterly*, *41*(1), 16. https://dx.doi.org/10.1353/hrq.2019.0000

Shneiderman, B. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems, 10*(4), Article 26, 31. https://doi.org/10.1145/3419764

The White House. (2024). *Statement from Vice President Harris on the UN General Assembly Resolution on Artificial Intelligence*. https://www.whitehouse.gov/briefing-room/statements-releases/2024/03/21/statement-from-vice-president-harris-on-the-un-general-assembly-resolution-on-artificial-intelligence/

Tiedemann, P. (2012). Is there a human right to life? *Annual Review of Law and Ethics, 20*, 345–360.
Todd Lopez, C. (2020). *US Department of Defense Principles of Artificial Intelligence Ethics*. https://www.defense.gov/News/News-Stories/article/article/2094085/dod-adopts-5-principles-of-artificial-intelligence-ethics/
UK Government. (2023). *The Bletchley Declaration by Countries Attending the AI Safety Summit*, 1-2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000381137
United Nations General Assembly. (2022). *The right to privacy in the digital age*: Resolution adopted by the General Assembly on 15 December 2022. https://documents.un.org/doc/undoc/gen/n22/762/14/pdf/n2276214.pdf
United Nations General Assembly. (2023a). *Impact of rapid technological change on the achievement of the Sustainable Development Goals and targets*: Resolution adopted by the General Assembly on 25 July 2023. https://documents.un.org/doc/undoc/gen/n23/227/47/pdf/n2322747.pdf
United Nations General Assembly. (2023b). *Promotion and protection of human rights in the context of digital technologies*: Resolution adopted by the General Assembly on 19 December 2023. https://documents.un.org/doc/undoc/gen/n23/422/28/pdf/n2342228.pdf
United Nations General Assembly. (2023c). *Information and communications technologies for sustainable development*: Resolution adopted by the General Assembly on 19 December 2023. https://documents.un.org/doc/undoc/gen/n23/418/14/pdf/n2341814.pdf
United Nations General Assembly. (2024). *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development: draft resolution*. https://documents.un.org/doc/undoc/ltd/n24/065/92/pdf/n2406592.pdf
United Nations. (2024). *Governing AI for Humanity*: Final Report. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf
Wellman, C. (1997). *An Approach to Rights: Studies in the Philosophy of Law and Morals*. Springer Science & Business Media.

ВИНИКНЕННЯ ПРАВА ЛЮДИНИ НА БЕЗПЕЧНЕ ВИКОРИСТАННЯ ШТУЧНОГО ІНТЕЛЕКТУ

Анотація. Після прийняття за останній рік низки міжнародних угод щодо штучного інтелекту, насамперед резолюції Генеральної Асамблеї ООН про безпечні, захищені та надійні системи штучного інтелекту для сталого розвитку (A/78/L.49), а також внаслідок посилення уваги до взаємозв'язку новітніх технологій та правової сфери – з'явилась велика кількість публікацій, автори яких комплексно підходять до вивчення питання управління штучним інтелектом та формування правового забезпечення використання штучного інтелекту. Зміни, які відбуваються, показують, що поряд із цифровими правами з'являються окремі права, пов'язані зі штучним інтелектом (на надійний, захищений, відповідальний, а також безпечний штучний інтелект). У цій статті представлене авторське бачення права на безпечний штучний інтелект, яке зможе гарантувати, що в результаті його використання шкода не буде завдана і небажані наслідки не виникнуть. В основі даного права лежить мінімізація ризиків і здійснення контролю над технологіями з боку людини. Автор розглядає різні людські потреби, що мають відношення до систем штучного інтелекту, задоволення яких впливає на дотримання прав людини у сфері штучного інтелекту, включаючи потребу в контролі. Крім того, автор досліджує, як право на безпечний штучний інтелект відрізняється від наявних прав людини, встановлених міжнародним правом. У статті зроблена спроба розкрити сенс права на безпечний штучний інтелект, зокрема шляхом аналізу етичних принципів. Очікуваним результатом від впровадження цього права є зведення до прийнятного мінімуму рівня ризиків при розробці та використанні систем штучного інтелекту. Висновки цього дослідження можуть бути цінними для подальшого вивчення проблематики прав людини, спорідненої із розвитком нових технологій. Право людини на безпечний штучний інтелект визначене як «право на безпечне використання систем штучного інтелекту» та має на меті підкреслити, що розробка систем штучного інтелекту не повинна загрожувати людству.

Ключові слова: право на безпечний штучний інтелект; системи штучного інтелекту високого ризику; права людини; екзистенційний ризик ІІІ; Резолюція Генеральної Асамблеї ООН A/78/L.49.

THE EMERGENCE OF THE HUMAN RIGHT TO SAFE ARTIFICIAL INTELLIGENCE

Abstract. Several international agreements on artificial intelligence were adopted last year, first and foremost the UN General Assembly resolution on safe, secure, and trustworthy AI systems for sustainable development (A/78/L.49). There has also been a growing focus on the connections between new technologies and the legal sphere, evidenced by many publications that approach comprehensively the issue of AI governance and establishing a legal framework for AI applications. We may assume that specific rights related to artificial intelligence are emerging alongside digital rights, such as trustworthy AI, secure AI, reliable AI, and safe AI. This article presents the author's perspective on the right to safe AI, guaranteeing that no harm will be caused and no undesirable consequences will occur due to such use. Minimizing risks and ensuring that humans maintain control over technology lie at the root of this right. The author explores the various human needs connected to AI systems and upon which AI-related human rights depend, including the need for control. Furthermore, he investigates how the right to safe AI may differ from existing human rights established in international law (rights to information and labor). The article contains an attempt to search for the meaning of the right to safe AI, primarily through examining some AI ethical guidelines. The proposed new right aims to ensure that the risks associated with developing and using AI systems are expected to be kept to an acceptable minimum. The findings of this study could be valuable for further exploration of human rights issues stemming from the rise of new technologies. The article briefly defines the human right to safe AI as "the right to safe use of AI systems" and stresses that developing AI systems should not threaten humanity.

Keywords: right to safe AI; high-risk AI systems; human rights; AI existential risk; UN General Assembly Resolution A/78/L.49.

Cite this article: Hachkevych, A. (2024). The emergence of the human right to safe artificial intelligence, *Law and innovative Society*, 2 (23), 32-40. doi: https://doi.org/10.37772/2309-9275-2024-2(23)-3.