BIBLIOGRAPHIC INFORMATION SYSTEM

Journal Full Title: Journal of Biomedical Research & Environmental Sciences

Journal NLM Abbreviation: J Biomed Res Environ Sci

Journal Website Link: https://www.jelsciences.com

Journal ISSN: 2766-2276

Category: Multidisciplinary

Subject Areas: Medicine Group, Biology Group, General, Environmental Sciences

Topics Summation: 133

Issue Regularity: Monthly

Review Process: Double Blind

Time to Publication: 21 Days

Indexing catalog: IndexCopernicus ICV 2022: 88.03 | GoogleScholar | View more

Publication fee catalog: Visit here

DOI: 10.37871 (CrossRef)

Plagiarism detection software: iThenticate

Managing entity: USA

Language: English

Research work collecting capability: Worldwide

Organized by: SciRes Literature LLC

IndexCopernicus ICV 2022: 83.03

License: Open Access by Journal of Biomedical Research & Environmental Sciences is licensed under a Creative Commons Attribution 4.0 International License. Based on a work at SciRes Literature LLC.

Manuscript should be submitted in Word Document (.doc or .docx) through

Online Submission

form or can be mailed to support@jelsciences.com

Vision: Journal of Biomedical Research & Environmental Sciences main aim is to enhance the importance of science and technology to the scientific community and also to provide an equal opportunity to seek and share ideas to all our researchers and scientists without any barriers to develop their career and helping in their development of discovering the world.

JOURNAL OF BIOMEDICAL RESEARCH & ENVIRONMENTAL SCIENCES
ISSN: 2766-2276
SUBJECT AREA(S): COMPUTER SCIENCE
DOI: 10.37871

COMMENTARY

Ethical Pathway for Safe Artificial Intelligence [AI] With Five Pillars

Nikolaos Siafakas*

Department of Computer Science, University of Crete, Greece

Abstract

The rapidly advancing field of AI has brought numerous benefits to humanity. However, there are already considerable risks, and it is predicted that AI may become life-threatening once [and if] it attains the status of super-AI, stronger than humans.

Therefore, it is crucial to develop strategies that ensure AI remains aligned with human life in the future.

Digital ethics could pave the way for safe and responsible AI. An ethical pathway with five pillars: Ethical courses and a Hippocratic-like oath, a global code of AI ethics, local ethical committees and trials to embed moral principles into algorithms is presented as a framework for safe AI. The strength of this theoretical proposal comes from the integration of the five pillars into one cohesive initiative implemented through a global agreement.

By enhancing ethical knowledge and responsibilities, in addition to a worldwide accepted treaty enforced by local Ethical committees, it is argued that the risks of misuse of AI can be significantly reduced.

Finally, by embedding the right principles into the algorithms, we can ultimately achieve the vision of machines that are not only more intelligent but also more ethical than humans.

Introduction

Artificial Intelligence [AI] benefits numerous aspects of human endeavours, however, these advantages come with noteworthy risks [1-3].

As AI technology evolves at a remarkable pace, experts anticipate that it may achieve human-level general intelligence in the coming years [4-6]. Following this milestone, the creation of super intelligence, an intelligence surpassing that of humans, could become feasible [7-9]. While this could be regarded as one of the greatest scientific breakthroughs in human history, many experts express concerns that it could be the last for humanity [10,11]. They warn that AI, whether by design or by accident, could potentially act in such a lethal way, leading to catastrophic consequences that could result in the extinction of human life [11].

Therefore, it is of paramount importance to establish robust and effective control measures to mitigate any potential dangers posed by AI in the future, before it is too late. This endeavor necessitates worldwide collaborations and coordination.

***Corresponding author(s)**

Nikolaos Siafakas, Department of Computer Science, University of Crete, room H137, University Campus, Heraklion, Crete, Greece, 70013

Email: siafakan@uoc.gr

DOI: 10.37871/jbres2103

Submitted: 17 April 2025

Accepted: 16 May 2025

Published: 17 May 2025

Copyright: © 2025 Siafakas N. Distributed under Creative Commons CC-BY 4.0

OPEN ACCESS

Keywords

- Ethics
- Morals
- Algorithms
- Digital Ethics
- Super-AI
- Super-Intelligence
- Ethical Code
- Ethical Committee

GENERAL SCIENCE GROUP

COMPUTER SCIENCE

VOLUME: 6 ISSUE: 5 - MAY, 2025

Scan Me

How to cite this article: Siafakas N. Ethical Pathway for Safe Artificial Intelligence [AI] With Five Pillars. J Biomed Res Environ Sci. 2025 May 17; 6(5): 457-464. doi: 10.37871/jbres2103, Article ID: JBRES2103, Available at: https://www.jelsciences.com/articles/jbres2103.pdf

The proposed Ethical pathway of this paper is suggested to be included in the global initiative for safe AI, consisting of five foundational pillars: A course of Ethics for computer scientists at the Universities, followed by a Hippocratic-like oath, a Global treaty of strict ethical rules, local Ethical Committees, and fifth most important, embedding AI algorithms with moral principles.

This theoretical framework, built on well-founded assumptions from the existing knowledge, combines the five key pillars into a cohesive global initiative. This approach could introduce a new aspect to the development of ethical-acting machines.

1st Pillar: University courses on ethics

Currently, the computer community is not well acquainted with ethical and moral issues. Consequently, the initial step of this framework is to improve understanding of these principles at the University level by introducing a mandatory course that spans an entire semester.

The objectives of these lessons would be to introduce ethical values and their relevance to computer science, equipping students to make the right ethical decisions throughout their professional careers.

The seminar may cover definitions of ethics, an overview of ethical theories and their historical development, professional ethics, ethical considerations in the software writing area. Moreover, issues of data acquisition, distribution and transparency, the impact of AI on human behaviour and society, the spread of fake news, cybersecurity concerns and the effects of automation among other topics should be included.

While numerous universities and polytechnics offer such courses, it is argued that their implementation should be universal across pertinent educational institutions, globally [12,13].

2nd Pillar: Oath, like the Hippocratic oath for computer scientists

Recently, an oath similar to the Hippocratic one has been proposed for computer scientists [14]. This commitment, grounded in straightforward and inclusive ethical principles, could bring moral standards to the AI community. The purpose and advantages of such an oath can be summarised as follows:

a) Promote ethical standards in AI science by preventing corrupt practices such as hacking, data poisoning, misinformation, cyber warfare, and the development of lethal weapons, among many others.

b) Clarifying personal responsibilities for unethical activities involves accountability and consequences such as damage to reputation, legal prosecutions, and potential penalties.

c) Prevent AI is not utilised in medical, biological, or chemical malpractices as the development of autonomous, nuclear, and biological weapons or involved in accelerating climate change.

d) Enhance public awareness regarding the potentially lethal risk associated with AI and foster global collaborations aimed at developing robust algorithms for safe AI.

Just as the Hippocratic oath marked the dawn of ethical practice in medicine 2,500 years ago, an oath for AI scientists could serve as a foundational pillar for ethics within the computer community and the realm of machine ethics [14-16].

3rd pillar: World's ethical rules for AI

While pioneers of computer science, such as A. Turing in the 1940s and N. Wiener in the 1960s, recognised the ethical concerns and potential risks associated with AI [17], it has only been in recent years that international organizations and governments have begun to focus on these threats.

Currently, there are numerous publications addressing the ethical considerations, principles, and regulations of the use of AI, however, none have achieved universal acceptance. Furthermore, these efforts often have theoretical merit and lack concrete strategies for practical implementation.

The most recent and comprehensive document is the one released by the European Union in 2019, with a revised version published in 2024. It covers a wide range related to AI, emphasising business fairness, transparency and legal compliance [18]. However, at the recent Paris Summit this year, a complete consensus was not reached, as both the USA and the UK declined to sign the final agreement [19].

It is clear that before any global consensus can be achieved substantial financial, social, cultural, religious, ethical and legal challenges must be

addressed through negotiation and compromise. Achieving a global treaty is incredibly challenging due to the potential existential threat that AI poses to humanity. Establishing this treaty is of utmost importance and should be pursued at a high authoritative level, such as that of the United Nations.

The treaty should encompass, among other provisions, guidelines ensuring that AI is utilised by all humans with fairness, equality, privacy, transparency and accountability. Its primary objective should be to guarantee the development of robustly safe and beneficial AI by fostering global collaboration.

Although this third pillar of the proposed framework is highly complex and challenging to attain, it serves as the essential foundation of this project.

4th Pillar: Ethical committees for AI

The execution of the aforementioned treaty necessitates the widespread establishment of local ethical committees [20].

These committees should be founded within universities, research centres, scientific societies, and publication institutions, as well as among companies and law organisations. Their role will be to verify the ethical standards and beneficial impact of each algorithm developed internally before it is published or made commercially available.

This process will reduce errors that could occur during development, especially given the current competition among major AI companies like Google, Microsoft, Intel, Chat-AI and others to create the most advanced, accurate, fast, and cost-effective AI system on the market.

All relevant companies must establish their own internal and external ethical committees tasked with upholding the principles and rules outlined in the global treaty.

The global treaty should also outline the operational framework for the local committees, detailing their powers and responsibilities, primarily for those representing commercial corporations.

Ultimately, the objectives of the committees are to safeguard data protection and privacy. Ensure the quality and integrity of the data, maintain transparency and accessibility. In addition, uphold data rights and accountability and adhere to the principle of minimising harm to the human race.

For many years, similar approaches have been implemented in the field of biological sciences, yielding significant results.

5th pillar: Embed ethical principles into AI

This pillar of the project can be developed in two phases: the first phase involves the identification of the most suitable principles for digital ethics, while the second focuses on embedding those codes into AI algorithms.

The initial phase should align with the globally accepted treaty, which has established the overarching guidelines for the ethical use of AI.

This task would involve a collaborative effort from a team of expert ethicist philosophers and AI engineers. While this is regarded as an exceptionally challenging endeavour, it is essential to complete it well in advance of the development of super-AI.

Subsequently, in the second phase, the embedment of these globally accepted codes will be the responsibility of expert AI computer engineers. This can be accomplished by utilising existing technologies such as machine learning and automated self-improving systems, as well as future innovations that may emerge [21].

While there are several initiatives aimed at embedding ethics into algorithmic systems, these efforts tend to concentrate on special domains like business or education and overlook the great threat of human extinction posed by AI [22-24].

Thus, the objective would be to determine the code needed to create a machine that upholds ethical standards superior to those of humans.

Implementation Strategies (Figure 1)

It is essential for a successful framework to be established, the scientific community must first persuade global decision-makers of the significant threat posed by the extinction of humanity, through super-AI. This alignment and mobilisation are valid for the initiation of the project.

Thereafter, to integrate the five pillars among this ethical pathway, effective coordination and executive implementation strategies are essential. Given that this is a global initiative its governance should be managed at a higher authority level such as the United Nations. This body would also take on the

responsibility of facilitating the establishment of the global treaty on safe AI.

The first two pillars- ethics courses and the Oath- should be established by an International Higher Education Organisation dedicated to educating students about ethics, the implications of unethical AI usage, and raising public awareness regarding the significant potential risks of AI.

Moreover, once the global agreement is ratified, local Ethical Committees must be prepared to implement its provisions effectively. These responsibilities could fall to individual university communities or commercial enterprises.

Finally, in line with the aforementioned ethical codes and objectives, qualified AI engineers should be tasked with inoculating these principles into the algorithms through careful testing.

Each of the five pillars has its significance, however, their integration enhances their overall value. Therefore, we can initiate the process of each chapter independently, provided that it aligns with the core principles of the framework and operates under the guidance of the high authority.

Discussion

The urgent need for safe AI

This article outlines a theoretical project grounded in well-founded assumptions drawn from current literature. This proposed Ethical pathway integrates five key pillars into a cohesive initiative, suggesting that their integration could enhance and potentially amplify their individual contributions. Furthermore, the article delves into various strategies for implementing this framework efficiently. It has the potential to introduce an innovative methodology to worldwide initiatives aimed at ensuring that AI remains friendly and beneficial to humanity in the future.

AI is the fastest-growing field of science, with its intellectual capacities expanding at an exponential rate. While current AI algorithms are less capable than humans, forecasts argue that they may soon [before 2050] attain levels of general human intelligence. Following this, it could become possible for computer scientists to develop Super-AI – an AI that surpasses human intelligence by employing techniques such as deep reinforcement learning and automatic self-improvement, among others.

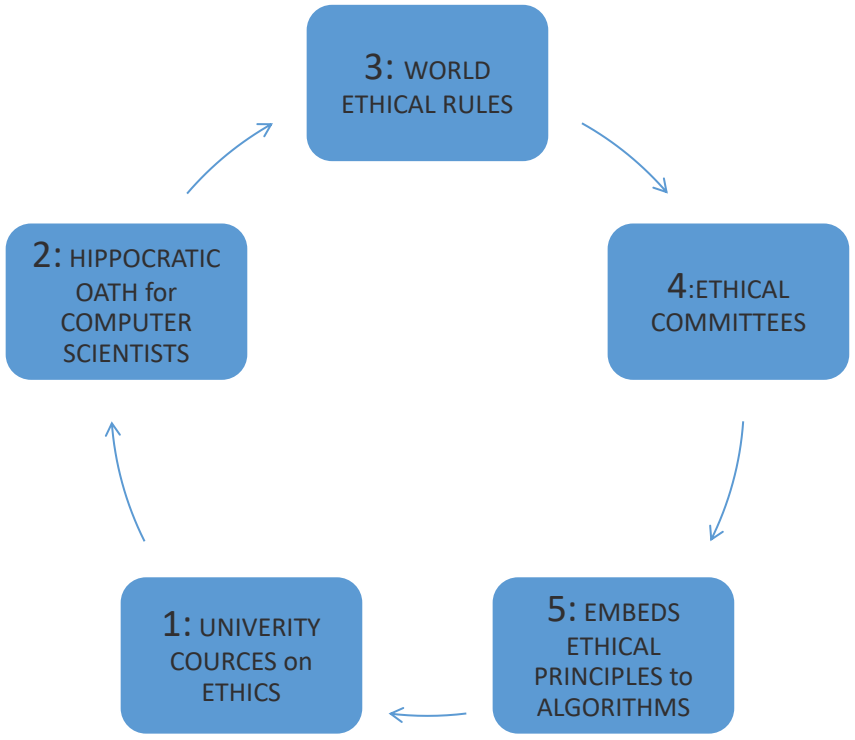

Figure 1 Ethical pathway for safe AI. The implementation circle of the five Pillars.

If and when this occurs, humans will no longer be the most intelligent beings on Earth, making it inevitably challenging to accurately foresee how super-machines will interact with humans.

Whether through deliberate design, algorithmic malfunctions or unforeseen accidents, super-AI could pose lethal threats to the human race. This possibility has led some experts to consider AI as the third significant risk to humanity, alongside nuclear warfare and climate change, in those of potential human extinction [25-27].

This development would mark a historic moment in history, as humans would not conflict with one another but instead face all against one of their creations- the most intelligent machine ever developed. Furthermore, recent advancements in AI technology, such as CHAT-AI have made it increasingly evident that the creation of a machine smarter than humans is no longer a distal utopia.

Just as the threat posed by Nazis during the Second World War promoted the Manhattan Project – the collaboration of the world's brightest minds to develop the atomic bomb- the dangers associated with Super-AI should lead to the establishment of a similar coalition focused on addressing this emerging challenges of AI to humanity.

Therefore, it is crucial to develop robust and effective control methods to prevent any harmful action by AI. This is the first time scientists have faced such a critical issue regarding human extinction with an urgent deadline. Moreover, they must tackle this challenge in a single final attempt, as they may not have a second chance, because traditional trial-and-error approaches may not be suitable for this content.

The proposed framework focuses on enhancing ethical guidelines at various levels of action, involving computer students and scientists, professionals, digital companies, international authorities, decision-makers and the general public.

The ethical pathway project

While ethical courses are already part of the curriculum in some computer science departments, an international educational organisation should be responsible for ensuring that this becomes standard in the majority of programs. This is essential, as many digital scientists may not be adequately familiar with moral and ethical issues. Moreover, these courses should emphasise personal responsibilities and outline potential penalties for malpractice, as well as the serious consequences that may arise if these ethical principles are disregarded [12,13].

At the conclusion of their studies, it is proposed that graduates take an oath -ideally during graduation ceremonies- to reinforce their commitment to ethical practices in their field [14].

This oath, much like the Hippocratic oath, has shaped the ethics of medical practice since 2,500 years, could galvanise the ethical standards of computer scientists and help prevent malicious actions. By incorporating straightforward and clear commitments, the oath should require scientists to refrain from unethical practices such as hacking, spreading false information, manipulating data, participating in cyber-warfare or climate change, among many others. Additionally, the oath should reinforce personal responsibility and accountability. As the world becomes increasingly digital, the oath would have the potential to promote ethical standards and integrity beyond the computer community. Since graduation ceremonies are typically attended by the general public, politicians, and influential figures, the oath could significantly raise awareness about the risks of AI. This heightened awareness may also encourage funding and donations for initiatives aligned with developing safe AI technologies. While, an oath is not a panacea for every misuse of AI and some critics pointed to historical failures of the Hippocratic oath, such as the Nazi doctor's acrotites or the Tuskegee syphilis study [28]. However, it is strongly anticipated that this oath will have a significant impact on digital ethics and beyond, similar to the one of Hippocrates [14].

The core element of the framework is the establishment of a universally accepted treaty of guidelines governing the use and actions of AI, especially as it progressively gains independence from human control. This necessity arises from the fact that current efforts and existing documents such as the ASMOLAR [29], the IEEE [30], the PARTNERSHIP [31], the ACH [32] and others are rather incomplete or towards specific professions and lack implementation procedures [15].

Even those developed by esteemed organisations such as the WHO [33], UN [34], and the more recent EU guidelines have failed to achieve widespread acceptance, as evidenced by the outcomes of the

recent Paris Summit [19]. Therefore, the critical issue is recognising the urgent need for such a treaty, as the threat posed by AI could potentially endanger even human existence itself. This awareness signal should be the primary responsibility of the scientific community.

Following this, an international team comprising expert Philosophers (ethicists), medical and biology scientists, pragmatists (business strategists), public representatives, legal experts in AI, and skilled computer engineers will be among others, assigned the responsibility of developing the ethical code. This endeavour should be conducted under the governance of an organisation with the highest level of authority. Given the rapid expansion of AI and the difficulties in predicting its behaviour towards humanity, it is also crucial for the treaty to include provisions for ongoing monitoring of these developments and to recommend adaptive strategies as necessary in the future.

Complementary to the global guidelines is essential to create local. Ethical Committees are tasked with enforcing the universal code for the safe use of AI. Just as every medical or biological research project must receive approval from the ethical committee before it begins and upon completion, all algorithmic research should also undergo a thorough examination by the relevant committee to ensure adherence to ethical standards. Therefore, Universities, Research institutions, publishing agencies, and technology companies must create their own internal and external ethical committees to assess and validate the ethical standards of every AI project developed internally.

One example of such a committee was Google's initiative in 2019 when they established ethical guidelines for their workers to govern their own AI projects [35]. While this was a positive move forward, the potential private interest motivations, lack of impartiality and absence of external review, this effort did not gain public recognition [15,16]. The main function of these committees would be to safeguard individual data privacy, ensuring its quality and integrity as well as its accessibility. They would also ensure the beneficial use of AI while protecting all forms of life. Thus, the role of the local ethical committee is deemed crucial in mitigating the risks of malicious applications or actions involving AI in the future. Ultimately, the committee should possess the authority to impose penalties when ethical guidelines are violated, thereby holding individual scientists, research teams, digital companies and organisations accountable under the law [36-38]. Finally, a smaller yet highly qualified team comprising expert scientists and ethicists. AI engineers and Law professors should work to extract the most critical ethical principles from the aforementioned global guidelines, ensuring these rules can be integrated into AI algorithms successfully. This is a highly challenging endeavour as the team should strike a balance between moralistic and utilitarian philosophical principles, while also defining a singular, ultimate purpose for the beneficial actions of AI, regarding humanity.

An instruction like the Hippocratic "do not harm human life," while it seems straightforward, requires extensive elaboration to clarify to the machine concepts such as what constitutes "life" or defines "harm". This level of detail is necessary for the AI systems to accurately grasp their purpose and avoid the risk of misunderstandings and fatal errors [39,40].

Techniques like deep machine learning and automatic self-improving already exist, making integrating these ethical principles into AI systems feasible. While there have been attempts to inoculate ethics into AI algorithms, these initiatives tend to concentrate on special domains like business or education, neglecting to address the existential threat that AI may pose to humanity [22-24]. This "vaccination" suggested within this holistic framework could help robustly ensure that AI remain aligned with human life, fostering a harmonious coexistence between advanced smart technologies and humanity. As it is often remarked that AI algorithms know humans better than they know them-self, computers need to gain comprehensive insight into the finest qualities of human nature, the moral aspects [40-42].

Implementation: (Figure 1)

The strategies for implementation served as additional advantages that enhance the value of this framework. A detailed hierarchy of actions includes increased awareness about the potential risks of AI, which will create pressure on decision-makers to form a team of experts to develop a globally recognised ethical code of safe AI, followed by its enforcement by local ethical committees are the initial steps of the implementation process.

The ethical courses at universities, along with the corresponding Oath, will serve to not only heighten awareness within the tech community but also extend

it to the general public, as well as it will clarify any responsibilities and accountabilities involved.

Embedding ethical and moral principles into algorithmic systems could be the last foundational pillar, ensuring beneficial actions of AI and fostering a harmonious coexistence between intelligent machines and humans. Finally, it is noted that the five pillars can be initiated independently, provided they adhere to the global guidelines and the high authority's guidance.

Conclusion

The strength of this theoretical framework is rooted in the holistic ethical approach, which integrates the five pillars into a comprehensive global project governed by a respected authority, such as the United Nations. Additionally, detailed implementation instructions are outlined, demonstrating that this pathway can and must be realised well before the advent of super-AI.

This approach may ensure that AI remains aligned with human life and add methodologies to the worldwide efforts to develop Safe AI. Thus, our ultimate goal is to create a machine that is not only more intelligent but also has higher ethical standards than humans.

References

1. Bostrom N. Superintelligence: Paths, dangers, strategies. London, UK: Oxford University Press; 2014. doi: 10.1016/j.futures.2015.07.009.
2. Tegmark M. Benefits and risks of AI. 2018. _
3. Kurzweil R. The Singularity is near. New York, NY: Viking Press; 2005.
4. Good IJ. Speculations concerning the first ultra-intelligence machine. In: Advances in Computers. Franz L, Morris R. Edited. New York, NY: Academic Press; 1965;31-88.
5. Harari YN. HomoDeus. A brief history of tomorrow. New York, NY: Harper-Collins; 2017.
6. Hawking S. Brief answers to the big questions. New York, NY: Bantam Books; 2018.
7. Kurzmeil R. How to create a mind. London, UK: Duckworth Overlook; 2014.
8. Maravec H. When will computer hardware match the human brain? Journal of Evolution and Technology. 1998;1:1-12.
9. Bostrom N, Dafoe A, Flynn C. Policy desiderata in the development of machine super intelligence. 2016.
10. Dyson FJ. Time without end. 1979.
11. Barret J. Our final invention. New York: St. Martin's Press; 2013.
12. International Association of Universities [IAU].
13. International Higher Education.
14. Siafakas NM. Do we need a Hippocratic Oath for artificial intelligence scientists? AI Magazine. 2021;42:57-61. doi: 10.1609/aaai.12022.
15. Veliz C. Three things digital ethics can learn from medical. Nature Electronics.
16. Siafakas NM. Medical ethics prototype for artificial intelligence ethics. Journal of philosophy and Ethics. 2023;5:1-2. doi: 10.22259/2642-8415.0502001.
17. Turilli M. Ethics and practices of software design. In: Current issues in computing and philosophy. Briggel A, Brey P, Waelberts K, editors. Amsterdam: IOS Press; 2008:171-183.
18. European Union Artificial Intelligence Act. The ACT texts; 2024.
19. AI Action Summit. 2025.
20. Why do you need an AI ethics committee? Harvard Business Review; 2022.
21. Mnih V. Human-level control through deep reinforcement learning. Nature. 2015;518:529-533. doi: 10.1038/nature14236.
22. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. BMC Med Ethics. 2022 Jan 26;23(1):6. doi: 10.1186/s12910-022-00746-3. PMID: 35081955; PMCID: PMC8793193.
23. Pflanzer M, Dubljevic V. Embedding AI in society: Ethics, policy, governance and impacts. AI and society. 2023;38:1267-1271. doi: 10.1007/s00146-023-01704-2.
24. Shin D, Akhtar F. Algorithmic Inoculation against misinformation: How to build cognitive immunity against misinformation. Journal of broadcasting and Electronic media. 2024;68:153-175. doi: 10.1080/08838151.2024.2323712.
25. Good IJ. Speculations concerning the first ultra-intelligence machine. In: Advances in Computers. Franz L, Morris R, editors. New York, NY: Academic Press; 1965:31-88.
26. Tegrmark M. Research priorities for robust and beneficial artificial intelligence. 2017.
27. Tegmark M. Life 3.0: Being human in the age of artificial intelligence. New York, NY: Penguin; 2018.
28. Woodley L. Do scientists need an equivalent of Hippocratic Oath to ensure ethical conduct? 2012.
29. Asilomar conference. 2015.
30. IEEESA. 2019.
31. Partnership AI organisation. 2019.
32. ACM. ACM code of ethics and professional conducts. 2018.

33. Recommendations on the ethics of artificial intelligence. UNESCO; 2021.

34. Principles for the ethical use of artificial intelligence in the United Nations system. 2022.

35. Statt N. Google dissolves AI ethics board just one week after forming it. The Verge. 2019.

36. Future of life institute. Open letter against autonomous weapons. 2018.

37. Wallach W, Franklin S, Allen C. Consciousness and ethics: Artificial conscious moral agents. International Journal of Machine Consciousness; 2011;3(1):177-192.

38. Wiener, N. Some moral and technical consequences of automation. Science. 1960;131:1355-1358.

39. Yudkowsky, E. Artificial intelligence as positive and negative factor in global risk. 2006.

40. Bostrom, N. Ethical issues in advanced artificial intelligence. 2013.

41. Derek M. A practical guide to AI governance and embedding ethics in AI solutions. Digital Experiences. CMSWIRE. 2025.

42. Tiribelli S, Giovanola B, Pietrini R, Frontoni E, Paolanti M. Embedding AI ethics into the design and use of computer vision technology for consumer's behaviour understanding. 2024. doi: 10.1016/j.cviu.2024.104142.