Ethical Pathway for Safe Artificial Intelligence (AI) with Five Pillars

Nikolaos Siafakas *

Posted Date: 30 January 2025

doi: 10.20944/preprints202501.2213.v1

Keywords: ethics; morals; algorithms; digital ethics; super-AI; superintelligence; ethical code; ethical committee

Article

Ethical Pathway for Safe Artificial Intelligence (AI) with Five Pillars

Nikolaos Siafakas

Department of Computer Science, University of Crete, room H137, University Campus, Heraklion, Crete, Greece, 70013; siafakan@uoc.gr

Abstract: Artificial Intelligence (AI), the rapidly advancing science, has been beneficial for humanity so far. However, there are already significant dangers, and it is predicted that it may become life-threatening when (and if) AI reaches the level of super-AI, an AI stronger than humans. Therefore, it is of paramount importance to invent methods that will guard AI to remain friendly to the human race in the future. Digital Ethics may be one way to achieve safe AI. An Ethical pathway with five pillars: ethical courses and a Hippocratic-like oath for computer students, a global code of AI ethics, ethical committees, and trials to embed ethical principles into algorithms is discussed as an additional method for safe AI. Enhancing ethical knowledge and responsibilities of AI scientists in addition to a globally accepted treaty and a proper function of local ethical committees, the danger of mal uses of AI would minimized. Finally by 'vaccinating' AI algorithms with appropriate ethical and moral principles, the dream task of a more clever but also more ethical than the human machine may achieved.

Keywords: ethics; morals; algorithms; digital ethics; super-AI; superintelligence; ethical code; ethical committee

Introduction

Artificial Intelligence (AI) benefits the majority of human activities. However, these benefits are followed by significant potential dangers.[1–3]

As AI advances extremely rapidly, it is predicted that within the next few years, it may reach the level of humans' general intelligence.[4–6] Thereafter, it would be easy to develop a SUPER-AI, an AI well above human intelligence.[7–9]

Although this could be considered the most important scientific achievement ever in human history, several experts fear that it may be the last one for humanity. Scientists, report that AI, intentionally or unintentionally, may act against humanity in such a lethal way that may cause the extinction of human life.[4,5,10,11]

Therefore, is of paramount importance to develop solid and efficient control mechanism(s) to prevent any harmful action of AI in the future, before it is too late. Obviously, this requires global efforts and coordination.

An Ethical pathway is proposed to be included in the global efforts for safe AI, consisting of five pillars: Courses of Ethics at the Universities for computer scientists, followed by a Hippocratic like oath, a Global Treaty of strict ethical rules, Local Ethical Committees, and fifth most important to embed AI machines with ethical and moral principles.

This pathway may add valuable methodologies to the Global efforts for safe AI to produce an ethical-acting machine.

1st Pillar: University Courses on Ethics

Today, the computer community is not well acquainted with ethical and moral principles. Therefore, the first step of this proposal is the enhancement of this knowledge at the University level

 doi:10.20944/preprints202501.2213.v1

through specific, whole-semester mandatory courses. The aim of the course could be to understand ethical principles and their application to computer science and to prepare the students to make ethical and moral decisions in their professional careers.

The course may include definitions of ethics, overviews of ethics theories and their evolutions over the years, professional ethics, ethical considerations in software development, distributions, and transparencies, the significant impacts of AI on human and society's behavior, cybersecurity, and effects of automation, among others.

2nd Pillar: An Oath Like the Hippocratic for Computer Scientists

An oath, similar to the Hippocratic one for medical students, was suggested for AI scientists.[12] This oath using simple and understandable ethical principles introduces moral practices into the AI community.

The aim and benefits of such an oath could summarized:

- Improve the ethical standards in AI science by preventing corrupt practices: hacking, data poisoning, cyber-war, and others.
- Increase the awareness of the potentially lethal dangers of AI.
- Make clear the personal responsibilities for unethical activities, their consequences, and the potential personal penalties.
- Prevent AI involved in medical malpractice and the production of lethal autonomous, nuclear, and biological weapons.
- Increase global collaborations to develop solid systems for safe AI.[12–14]

3rd Pillar: Globally Accepted Ethical Rules for AI

Although the founders of computer science, Allen Turing in the 40s and Wiener in the 60s had noticed the ethical implications and threats of AI, only recently International Organizations and Governments paid attention to those dangers.[15]

Today, there are several publications on ethical issues, principles, and rules for the use of AI but none have reached global acceptance. Moreover, these efforts had more theoretical merit without specifying ways for their implementation. It is obvious, that before any Global acceptance is achieved, significant financial, social, cultural, ethical, and legal issues have to be solved.

The more recent and comprehensive document of AI ethics is the one produced by the European Union in2019 and its revision published in 2024.[16] This publication covers most of the current issues of AI involvement and focuses on business fairness, transparency, and legality.[16]

However, a global treaty urgently needed to be developed at the highest authoritarian level, such as the one of the United Nations, and signed by all its country members.

It is considered that if the majority (if not all) of the countries of the world agree, the chances of the global efforts to develop a safe and solidly friendly AI towards humanity could be more easily accomplished.

The treaty should include rules that AI should be used by all humans with fairness, equality, privacy, transparency, and accountability.

Furthermore, the principal aim of the treaty should be to ensure the safe and beneficial use of AI for humanity, in the future. In that case, the treaty should promote the global collaboration of AI scientists in developing the proper algorithms to ensure a safe and human-friendly AI.

4th Pillar: Ethical Committees for AI

The implementation of the Global Treaty needs local Ethical Committees to be established.

These committees at the universities, Research centers, Scientific Societies, Publication Institutes, Computing Companies, and Organizations will have the task of verifying the ethical standards and merits of each algorithm produced in-house before its publication or becoming commercially available. This procedure minimizes errors that may happen during the race of the major AI

 doi:10.20944/preprints202501.2213.v1

companies (Google, Microsoft, Intel, Chat-AI, and others) to develop, the most advanced, accurate, fast, and cheap AI system of the market. It is obvious, that the above companies have to establish their internal ethical Committees.

Thus, the global treaty should also cover the way that the local ethical committees would function, especially those acting for commercial corporations. Overall, the goals of these committees should be the protection and privacy of the data, its quality and integrity, its transparency and availability, as well as, data rights and accountability.

5th Pillar: Embed ethical principles into AI

Theoretically, this pillar of the pathway could be developed in two stages: one to verify the most proper principles of digital ethics and a second these principles to be technically embedded into machines, by AI engineers.

It is apparent, that the first stage is an extremely difficult one because the task of producing a globally accepted set of ethical rules is enormous. It will have to overcome very strong obstacles concerning philosophical, legal, cultural, religious, and ethical issues.

Although extremely difficult, all these problems have to be solved well on time, with compromises and wisdom.

The second phase of this pillar will be to embed the globally accepted modes of ethics into the functions of AI and its proper algorithms.

This may achieved by expert computer engineers using already existing technologies like machine learning and automatic self-improving techniques or those invented in the future. The task could be a 'machine' more ethical than the humans.

Discussion

In this article, a proposal for safe AI based on Ethical principles is presented. This Ethical pathway consists of five pillars aiming to add a methodology to the global effort to keep AI friendly and beneficial to humans, in the future.

It is known, that AI is the most rapidly advancing science, and its intellectual capacity is increasing exponentially. Although AI today is weaker than human intelligence, it is predicted that in a few years (before 2050) it may reach the level of the human. Thereafter, it would be feasible for computer scientists using machine learning techniques and automatic self-improving ones, to produce a SUPER-AI, an AI with intelligence well above the human.[17–20]

When (and if) this happens humans will no longer be the smartest entities on Earth and it is extremely difficult to predict accurately the way that the super-machines will behave towards humanity. Intentionally or unintentionally SUPER-AI may act lethally against the human race, and this makes some experts consider AI the third potential danger, in addition to nuclear war and climatic change, of extinction of humanity. [5,6,10,11]

It would be the first time in history that humans would not fight humans but confront one of their creations the more intelligent machine.

Thus, it is of paramount importance to invent a solid and efficient control mechanism(s) to prevent any harmful effects of AI. It is the first time that scientists have to solve such a crucial problem of human existence with a very short deadline. In addition, they have to do it in one go, since they may not have a second chance and the classical method of trial and error may not be applicable in this case.

Obviously, such a great treat needs global collaboration, coordination, and strong alliances of scientists using a variety of methodologies to achieve the goal of a friendly-to-human AI. [20]

This proposal is based on enhancing Ethical principles at different levels of actions of AI and concerns computer students and scientists, Digital companies, and international high authority organizations, as well as, decision-makers and the public.[12,14,16]

Although courses of Ethics are already included in the curriculum of several Computer Science Departments in the West, it is suggested that this should be the case for all of them, because today

 doi:10.20944/preprints202501.2213.v1

digital scientists are not well acquainted with moral issues and the potential lethal consequences if those principles are ignored. Moreover, these courses should emphasize the personal ethical responsibilities of each scientist and the potential penalties in case of malpractice.

At the end of their studies, it is proposed that they must take a special oath, preferably during their graduation ceremony.[12]

This oath, like the Hippocratic one for Medicine, will enhance

the morality of future scientists in avoiding malicious practices. In addition, it may increase the awareness of the AI dangers of the public, politicians, and other influential persons who usually attend Graduation Ceremonies and may increase the funding and donations for developing safe AI.() Although an Oath is not a panacea for all misuses of AI and few are arguing against it,[21] it is strongly believed that as the Hippocratic oath shaped not only medical ethics but was the basis for general ethics and significantly affected professional behaviors. The oath for computer scientists is suggested to affect positively humanity since its future is predicted to become more and more digital.[12,14]

The recent public discussions on the potential harms of AI make it apparent that the invention of a smarter machine than humans is no longer a utopia.

Today the existing documents for the ethical use of AI, such as the Asimolar [22], the IEEE[23], the Partnership AI [24], the ACM [25], and others, are more theoretical and luck-specific implementation strategists.[26] Even the most recent, the one produced by the European Union, has more continental application [16].

Since the dangers of AI are potentially global, the efforts to prevent them must be universal and at the highest authoritarian level. It is suggested that this project should be initiated, coordinated, and implemented by the United Nations. An international team of expert Philosophers (Ethicists), Medicals, and Pragmatists (business strategists). Representatives of the UN (public representatives), Law experts and AI Engineers will be given the task to produce such an ethical code. In addition, as AI is rapidly advancing and it is difficult to predict its behavior towards the human race, the treaty should advise for continuous monitoring of its evolution and adaptations of the rules accordingly.

Complimentary to the world treaty for safe AI, must be the establishment of Ethical Committees whose function would be the local implementation of the above universal code of the use of AI. Similarly to the way that every research project in medicine or biology, well before starting, had to be approved by an ethical committee, as well as its final product, all AI algorithms have to be examined with scrutiny for their ethical values by an appropriate ethical committee. University Departments, Research Centers, Publication Agencies, and Computer Companies should establish their committee to verify the ethical standard of each AI algorithm made in-house.[26]

One attempt at such a committee was the one of Google in 2019, by issued rules for their work on AI. [30]Although it was a step in the right direction, due to the potential private interest-driven motivation and luck of impartiality and external reviewing, this effort did not achieve public recognition.[13,26] These committees should protect the privacy of individual data sets, their quality and integrity, its publicly availability, as well as, maintain the beneficial use of AI and protect any form of life on earth. It is considered that the function of these committees is very important in preventing malicious uses of AI such as hike, fake news, and data poisoning cyberwars. Autonomous weapons [27]and more others.

In addition, the committee should have the authority to implement penalties when the ethical rules are not followed making individual scientists, research teams, or digital companies accountable against the Law.[28–31]

Finally, based primarily on the accepted global rules computer scientists will have to embed these principles into all AI algorithms, using the already known techniques such as deep machine learning and automatic self- improving and others that will invented in the future, to 'vaccinate' AI with ethics making them to continue to act beneficially towards humanity. As commonly stated AI knows us better than we know ourselves, let it know the best of human aspects, the ethical and moral behavior.

 doi:10.20944/preprints202501.2213.v1

In conclusion, this ethical pathway could be an additional method in the global effort to produce a safeguard method against any harmful action of Super-AI. The development of all five pillars of the pathway should be initiated as soon as possible to be accomplished well before a Super-AI invention takes place, as our dream task should be the existence of a machine more cleaver but more ethical than humans.

References

1. Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. London,UK: Oxford University Press.
2. Tegmark, M. 2018b. Benefits and risks of AI. http://futurelife.org
3. Kurzweil, R. 2005. *The Singularity is Near*. New York, NY: Viking Press.
4. Good, I. J. 1965. "Speculations concerning the first ultra-intelligence machine." In *Advances in Computers*, edited by L. Franz and R. Morris, 31–88. New York, NY: Academic Press.
5. Harari, Y.N. 2017.*HomoDeus.ABrief History ofTomorrow*.NewYork, NY: Harpet-Collins
6. Hawking, S. 2018. *Brief Answers to the Big Questions*. New York, NY: Bantam Books.
7. Kurzmeil, R. 2014. *How to Create a Mind*. London, UK: Duckworth Overlook.
8. Maravec, H. 1998. "When will computer hardware match the human brain?." *Journal of Evolution and Technology* 1: 1–12
9. Bostrom, N., A. Dafoe, and C. Flynn. 2016. Policy desiderata in the development of machine superintelligence.http://nickbostrom.com/papers/aipolicy.pdf
10. Dyson, F. J. 1979. Time without end. http://blog.regehr.org/extra-files/dyson.pdf
11. Barret J: Our final invention 2013. St. martins press, New York.
12. Siafakas,N.M.2021. "Do we need a Hippocratic Oath for artificial intelligence scientists?"AIMagazine42:57–61.
13. Veliz, C. 2019. "Three things digital ethics can learn from medical
14. Siafakas N.M. 2023. Medical ethics prototype for artificial intelligence ethics. Journal of Philosophy and Ethics: 2023: 5(2): 1-2
15. Turilli, M. 2008. "Ethics and practices of software design." In *CurrentIssues in Computing and Philosophy*, edited by A. Briggel, P. Brey,and K.Waelberts, 171–83. Amsterdam: IOS Press.
16. European Union artificial Intelligence Act. The ACT texts, 2024. Accessed June 15,2024.https://artificialiteligenceact.eu/the-act/
17. Mnih, V. et al. 2015. "Human-level control through deep reinforcement learning." *Nature* 518: 529–33
18. Good, I. J. 1965. "Speculations concerning the first ultra-intelligence machine." In *Advances in Computers*, edited by L. Franz and R.Morris, 31–88. New York, NY: Academic Press.
19. Tegrmark, M. 2017. Research priorities for robust and beneficial artificial intelligence. http://fururelife.org/ai-open-letter/
20. Tegmark, M. 2018a. *Life 3.0: Being Human in theAge of Artificial Intelligence*.New York, NY: Penguin.
21. Woodley, L. 2012. Do scientists need an equivalent of Hippocratic Oath to ensure ethical conduct? http://www.lindau-nobel.org/doscientist-need-an-equivelent-of-the-hippocratic-oconduct/
22. Asilomar conference. 2015. http://tinyurl.com/asilomarAI
23. IEEE A.2019.http://standards.ieee.org/develop/indcom/ec/ead_v1.pdf
24. Partnership AI Organization. 2019. http://www.partnershiponai.org
25. ACM 2018. ACM code of Ethics and Professional Conducts. https://www.acm.org/code-of-ethics.
26. Statt, N. 2019. "Google dissolves AI ethics board just one week after forming it." *The Verge*. https://go.nature.com/2Zg727k
27. Future of Life Institute. 2018. Open letter against autonomous weapons. http://futurelife.org/open-letter-autonomousweapons/
28. Wallach, W., S. Franklin, and C. Allen. 2011. "Consciousness and ethics: Artificial conscious moral agents." *International Journal of Machine Consciousness* 3: 177–92.
29. Wiener, N. 1960. "Some moral and technical consequences of automation." *Science* 131: 1355–8.

 doi:10.20944/preprints202501.2213.v1

30. Yudkowsky, E. 2006. Artificial intelligence as positive and negative factor in global risk. http://inteligence.org/files/AIPosNegfactors.Pdf
31. Bostrom, N. 2013. Ethical issues in advanced artificial intelligence. http://www.nickbostrom.com/ethics/a.i.html