Josip Horvat

THE ETHICS OF ARTIFICIAL INTELLIGENCE

UDC
Essay

Summary

There are many ethical questions relating the issue of developing an intelligent system. There is strong and increasing pressure to raise capabilities of the artificial intelligence at least to the human levelled intelligence as the ultimate goal. This essay describes possible paths of development of the artificial intelligence. It is discussed how this changes will affect our society and challenges that humanity will have to face. Principles, guideways and modern viewpoints are presented and confirmed with the statements of the renowned scientists and experts in the field of the artificial intelligence ethics.

Key words: *ethics; artificial intelligence; moral issues; robotics*

1. Introduction

At the time when advances in computing science has made intelligent systems more and more applicable to our everyday life, we have once again came to the point when we have to reassess our ethical framework and social norms. We are definitely at the turning point in human history. Machines are smarter and smarter, intelligent algorithms are in use in various branches and in engineering science we are already in fourth industrial revolution [1] which is characterized by so-called “Cyber-Physical Systems” (CPS) [2]. Rapid technology development has fundamentally changed our society and economy. Consequences of intelligent systems are various and due to nature of artificial intelligent algorithms they are sometimes unpredictable.

The field of artificial intelligence ethics has existed for a number of years and has recently seen a resurgence of interest. The ethics of artificial intelligence deals with number of different issues such as: ethics in machine learning algorithms, moral status and rights of intelligent machines, superintelligence, roboethics, and moral behaviour of humans as designers of artificially intelligent beings.

1.1 Institutions and researchers

There are many institutions and researchers dealing with artificial intelligence ethics. For example Centre for Study of Existential Risk (CSER) as part of research centre at the University of Cambridge, study possible extinction-level threats posed by present or future technology. On the other hand Future of Humanity Institute (FHI) is part of the Faculty of Philosophy and the Oxford Martin School at the University of Oxford and deals with same effect of future technology development on the human condition. There are also Future of Life Institute (FLI) that is settled in Boston and Machine Intelligence Research Institute (MIRI) that acts as non-profit organization. MIRI's technical agenda states that new formal tools are needed in order to ensure the safe operation of future generations of AI software (friendly artificial intelligence) [3]. As research institution MIRI is dedicated to research of safety issues related to the development of strong artificial intelligence.

Scientists that are working on the field of AI ethics are next. Probably the most famous is Nick Bostrom [4] known for his work on existential risk, the anthropic principle, human enhancement ethics, superintelligence risks, the reversal test, and consequentialism. Nick Bostrom is Swedish philosopher at the University of Oxford and he is currently the founding director of the Future of Humanity Institute at Oxford University. He is author and co- author of many scientific papers, books and articles. Most known books are: Anthropic Bias: Observation Selection Effects in Science and Philosophy, Human Enhancement, Global Catastrophic Risks and Superintelligence: Paths, Dangers, Strategies. Second well known person is Raymond "Ray" Kurzweil [5]. He has written books on health, artificial intelligence (AI), transhumanism, the technological singularity, and futurism. The best known are: The Age of Intelligent Machines, The Age of Spiritual Machines and The Singularity Is Near. Peter Norvig [6] is also one of the most famous scientists dealing with the AI ethics. He is Director of Research at Google Inc. He was Assistant Professor at the University of Southern California and a research faculty member at Berkeley. He has published over fifty papers in various areas of Computer Science and has written number of books including: Artificial Intelligence: A Modern Approach, Paradigms of AI Programming: Case Studies in Common Lisp, Verbmobil: A Translation System for Face-to-Face Dialog, and Intelligent Help Systems for UNIX. There is also a number of scientists dealing with AI ethics, and they are: Steve Omohundro, Stuart J. Russell, Anders Sandberg and Eliezer Yudkowsky. In consideration of Nick Bostrom's work and his science background the following text of this essay will be mostly based on his articles, papers and philosophical reflections.

2. What is Artificial Intelligence?

To completely understand the content of the artificial intelligence ethics, it has to be explained what artificial intelligence really is. Artificial intelligence is academic field of study which studies how to create machines and computer software that are capable of intelligent behaviour. In some definitions it could be found that AI is science and engineering of making intelligent machines, especially intelligent algorithms. By its genuine property it follows that artificial intelligence is intelligence expressed by machines or computer programs. In order to achieve profound understanding of artificial intelligence it has to be explained term of intelligence. According to John McCarthy "Intelligence is the computational part of the ability to achieve goals in the world. Varying kinds and degrees of intelligence occur in people, many animals and some machines."[7] Solid definition of intelligence doesn't exist. "Not yet. The problem is that we cannot yet characterize in general what kinds of computational procedures we want to call intelligent. We understand some of the mechanisms of intelligence and not others."[7] Artificial intelligence as multi-disciplinary field of research is still constantly and actively growing and changing. Applications and possibilities of such algorithms are numerous. There are many different approaches to programing intelligent software. Some of them are based on probabilistic methods such as Bayesian network, Hidden Markov model, Kalman filter, Decision theory and Utility theory. Other are based on the mathematical theory of optimization such as evolutionary algorithms and classifiers where the neural networks are most widely used.

The ultimate goal in the field of making and understanding intelligence is to make computer programs that can solve problems as well as humans, or better. However, modern scientists and researchers are much less ambitious. Modern solutions and approaches gives excellent results in specialized field of computing, decision making or in one single domain. Modern issue of artificial intelligence is generality. This is the field in which humans are superior to computers. As it was written in John McCarthy’s 1971 Turing Award Lecture entitled ``Generality in Artificial Intelligence" [8]. Computers are intrinsically incapable of doing what people do and this problem date since 1930s when mathematical logicians, especially Kurt Gödel and Alan Turing, established that there did not exist algorithms that were guaranteed to solve all problems in certain important mathematical domains. To define more closely computers are still terrible at figuring out what the main verb occurring in figure (Fig. 1).

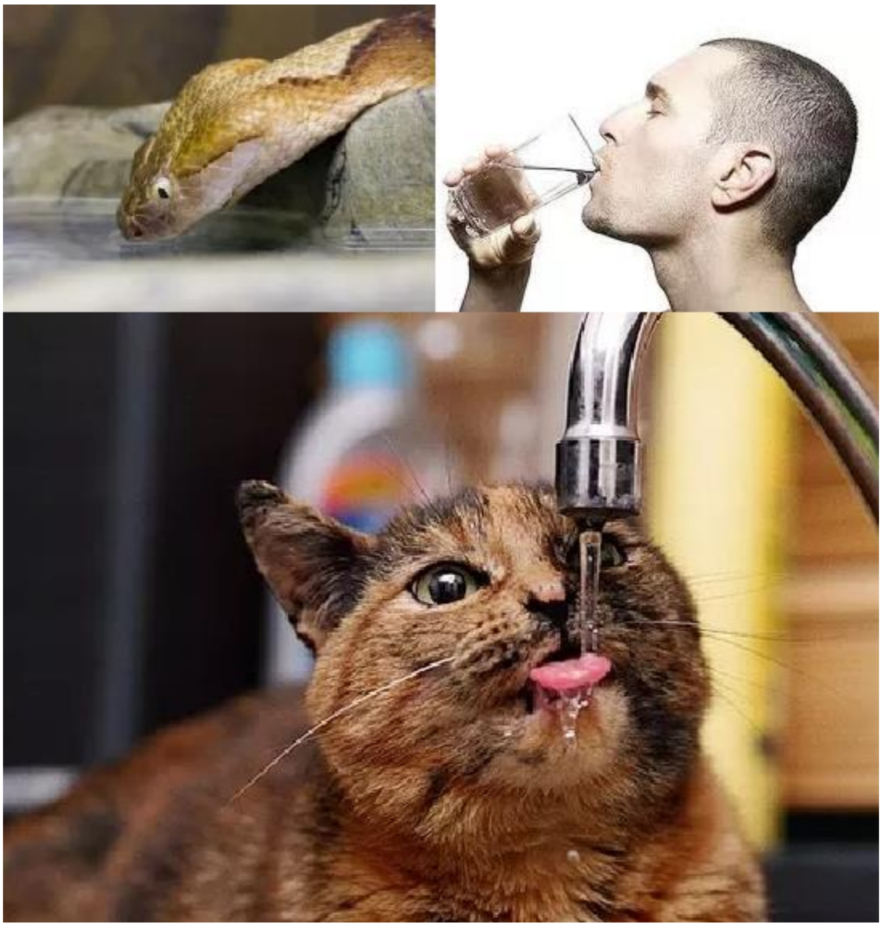

Fig. 1 Common problem in artificial intelligence [9]

As it is said earlier real scientists are less interested in building a machine that possesses human-levelled intelligence. They are rather interested in building systems that can solve problems efficiently and correctly.

Artificial intelligence is still in its early stage. But there are some beautiful examples and impressive application. It has seen the biggest boom thanks to the phenomenon known as big data. For example Google Translate uses text search of massive collections of documents translated from one language into another. Thanks to this big date engines we could say that that are the examples of humanities greatest thinking tools. Since the technology is rapidly progressing the possibility of the artificial general intelligence (AGI) could begin to be realistic.

3. Ethical issues in artificial intelligence

As it was written by Nick Bostrom in Cambridge Handbook of Artificial Intelligence [10] the possibility of creating intelligent machines raises a host of ethical issues. The first and probably the most important thing is the issue of harming people. Then there is an issue of assignment of moral status to intelligent machines. As well there is moral obligation of designers of intelligent systems. In case of creating an AI more intelligent than humans there has to be ensured that they are used for good rather than ill.

3.1 Three Laws of Robotics

From the first appearance of word “robot” in Karel Čapek’s play from 1921 entitled "Rossum's Universal Robots" since today there was lots of ethical issues linked to the intelligent machines. Isaac Asimov, author and creator of the Three Laws of Robotics made the first step. His laws of robotic are inevitable theme in all ethical discussions and they present the foundation of ethics in artificial intelligence. They are as it follows.

1. “A robot may not injure a human being or, through inaction, allow a human being to come to harm.”
2. “A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.”
3. “A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws” [11]

But soon after the creation of three laws of robotics appeared additional laws of robotics because the first three couldn’t cover possible endangering of human species. It was sure that the ethics of the artificial intelligence is much wider field and it can’t be covered with only three laws.

3.2 Ethics in Machine Learning and Other Similar Approaches

Today are machine learning and other similar approaches well known and applied in the practice. These approaches are widely spread and used in problems of robot motion planning, localization, mapping, knowledge classification, object recognition and decision making. Ethical validity of intelligent robotic arm is no more questionable in the term of its design and function because there are not new ethical challenges. But application of AI software which takes cognitive work with society and plays a large role in bringing out socially important decisions inherits the social requirement. In Nick Bostrom’s [12] work there is example with mortgage application for approval where rejected applicant brings a lawsuit against the bank and claims that the application makes decisions based on the colour of the skin. Depending on the type of the application algorithm it could be almost impossible to find out is this statement true or false. With this in mind it is obvious that AI algorithm should be transparent. If the software is used for socially important decisions should be predictable too. Predictability of legal systems is important feature because its role is to provide predictable environment in which citizens can optimize their own lives. So contracts can be written knowing how they will be executed. Robustness against manipulation is maybe the most important criterion and as Bostrom said, nearly the criterion.

In the conclusion of this section it will be cited Nick Bostrom’s [12] list of different criterion that should be a guiding light towards the better and ethically suitable AI software. *“Responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream with helpless frustration: all criteria that apply to humans performing social functions; all criteria that must be considered in an algorithm intended to* *replace human judgment of social functions; all criteria that may not appear in a journal of machine learning considering how an algorithm scales up to more computers. This list of criteria is by no means exhaustive, but it serves as a small sample of what an increasingly computerized society should be thinking about."*

3.3 Artificial General Intelligence

As it was written in the second chapter intelligent software shows excellent characteristic in one single domain. When they are applied to some general activities they tend to be unusable. For many researchers general intelligence is still among the long-term goals. However the possibility of creating general intelligence raises philosophical issues about the nature of the mind and the ethics of creating artificial beings. Machines with general intelligence (strong AI) that could successfully perform any intellectual task as a human being are qualitatively different class of problem than the AI operating in one single domain as it is written in chapter above. As was written by Bostrom [12] *"But our adapted brain has grown powerful enough to be significantly more generally applicable; to let us foresee the consequences of millions of different actions across domains, and exert our preferences over final outcomes."* With that in mind, it is really hard to define ethical suitability and safety of an AGI. It is impossible to foreseen the possible decisions of the software that would operate across a number of domains and work on unforeseen problem. Considering such application it is obligation to make a trustworthy assurance that will guarantee the AGI safety. There are many questions about ethics of such software and for sure it not an easy topic. Although there isn't defined laws and obligations we can establish some general guide lines proposed by Bostrom [12].

- "The local, specific behaviour of the AI may not be predictable apart from its safety, even if the programmers do everything right;
- Verifying the safety of the system becomes a greater challenge because we must verify what the system is trying to do, rather than being able to verify the system's safe behaviour in all operating contexts;
- Ethical cognition itself must be taken as a subject matter of engineering" [12].

3.4 Moral status

Moral status is term mostly connected with human species. If we take in consideration the Francis Kamm's definition of moral status that "X has moral status = because X counts morally in its own right, it is permissible/impermissible to do things to it for its own sake." [12] than an AI system could easily be candidate for moral status. For example, today AI systems do not possess moral status. Moral constraints are grounded only in the responsibilities of designers of the software. According to Bostrom, two criteria are commonly linked to moral status: sentience and sapience (or personhood) and they could be defined as it follows.

> "Sentience: the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer"
>
> "Sapience: a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent" [12]

Following this two definition's Bostrom states that one common view is that many animals have qualia and therefore have some moral status, but only the humans have sapience.

Therefore this could suggest that an AI system could have moral status if it possesses an ability to feel pain. As it is wrong to inflict pain to a mouse or cat the same would hold for any sentient AI system. If intelligent system would have kind of sapience similar to the normal human than it would have full moral status. In chapter “Machines with Moral Status” [12] Nick Bostrom adds two more principles. First principles entitled “Principle of Substrate Non-Discrimination” says “If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.” Second principle entitled “Principle of Ontogeny Non-Discrimination” says “If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status.” This principles do not imply that computer could be conscious or be equal and have same functionality as a human being. But at the end as Bostrom says, it makes no moral difference whether a being is made of silicon or carbon or whether its brain uses semi-conductors or neurotransmitters. On the other hand if the principle of non-discrimination is accepted a number of ethical questions arises. Because of the different properties of the human and artificial minds we must consider how this changes would affect the moral status of artificial minds.

3.5 Superintelligence

Superintelligence is by the esteemed definition, a system capable of understanding its own design and regarding to this fact, capable of redesigning itself or creating successor system, more intelligent, which could redesign itself again to become more intelligent repeating this characteristic in positive feedback cycle becoming better and better. According to another definition superintelligence is any intellect that vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills [13]. It is logical that the existence of such system would bring many ethical problems quite different from those arising in current automation and information systems. To the Stephen Hawking [14] "The development of full artificial intelligence could spell the end of the human race." As it is said earlier three laws of robotic are incapable of designing a functional ethical framework. By definition intelligent entities have the cleverness to easily overcome such barriers. Taking this in consideration, superintelligence could have much greater consequences than all the activities of the human race which practically reshaped the globe by carving mountains, taming rivers, building skyscrapers, etc. To define the characteristics of this kind of intelligence Bostrom made a number of statements.

- Superintelligence may be the last invention humans ever need to make.
- Technological progress in all other fields will be accelerated by the arrival of advanced artificial intelligence.
- Superintelligence will lead to more advanced superintelligence.
- Artificial minds can be easily copied.
- Emergence of superintelligence may be sudden.
- Artificial intellects are potentially autonomous agents.
- Artificial intellects need not have humanlike motives.
- Artificial intellects may not have humanlike psyches.

Conclusion is that thinking that emergence of superintelligence can be predicted by extrapolation history of other technological breakthrough or that the nature and behaviours of artificial intellects would necessarily resemble those of human or other animal minds should be taken with precaution and big suspicion. On the other hand superintelligence could outstrip humans and become superior to human capabilities of ethical behaviour. It could have all the correct answers that can be brought by reasoning and weighting up evidence in order to make ethically better decisions. As it this field, so in others, like long-term planning a superintelligent machine could outperform human in every point.

The catastrophic script is described in book “Superintelligence” where Nick Bostrom provides a detailed argument for the threat that artificial intelligence may prove to mankind. He states that an AI system can achieve world dominance in only a few phases. Such behaviour of an AI would depend upon the capacity to match human intelligence, evolve and develop the intelligence of its own. Bostrom find this script very possible although much scientific evidence proves otherwise. For the end of this chapter it will be cited a statement based on the VOX TECHNLOGY article [14] to ensure us , even for a moment that this script is most likely impossible to happen. “*No matter how advanced a machine is, it cannot rely on itself, or another machine, for repairs and assistance. Humanity, therefore, is necessary for the preservation of artificial intelligence. A supercomputer that extinguished humanity would be extinguishing itself. Such a computer, with above-human intelligence, would undoubtedly realize this, and keep humanity alive to ensure the perpetuation of its kind. Some argue that other machines could be built to aid these computers, but those machines in turn would need repairs. The creation of self-service machines is impossible for the foreseeable future, making it impossible for artificial intelligence to exist without humanity.”* [16]

4. Conclusion

Several fields of artificial intelligence where described in this essay and discussed from ethical point of view.

Development of an AI system, even single domain one, presents a great responsibility for ethical creators in the terms of ethical suitability. If an intelligent system makes socially important decisions by ethical users it must be transparent, predictable and robust against manipulation. If not, the predictability of a general intelligence software could be ethically compromised by practice with heavy consequences. In that case there must be an ethical creator obligation to make ethical and trustworthy assurance that will guarantee the AGI safety and reliability with respect to governing ethical requirements. And, by the way, ethical rules need further investigations in terms of AI development, implementation and application.

There is a possibility that some future AI systems might be candidates for having some form of moral status. If the humankind is serious in developing an advanced intelligent systems it could count as person. Then humanity has to comprehend the structure of ethical questions regarding AI in the social and natural context of possible conflicts of interests at least in the way it understands the complex structure of chess.

Further advancing and making conditions for development of superintelligence, humanity will be most probably confronting unseen circumstances. Although this script sounds fictional it is likely to be developed sooner or later. If that will happen, a new ethical framework will be established and ethical perspective will have to change into a new one. Compared to a present-day perspective, ancient and people of today most probably will not be seen as ethically perfect by future civilizations, (if perfect ethic could exists at all?).

REFERENCES

[1] Jaap Bloem, Menno van Doorn, Sander Duivestein, David Excoffier, René Maas, Erik van Ommeren; “The Fourth Industrial Revolution -Things to Tighten the Link Between IT and OT”; Sogeti VINT 2014; http://www.fr.sogeti.com/globalassets/global/downloads/reports/vint-research-3-the-fourth-industrial-revolution Retrieved 10.02.2016.

[2] The TerraSwarm Research Center; “Cyber-Physical Systems”; http://cyberphysicalsystems.org/ Retrieved 10.02.2016.

[3] Soares, Nate; Fallenstein, Benja (2015). "Aligning Superintelligence with Human Interests: A Technical Research Agenda" (PDF). In Miller, James; Yampolskiy, Roman; Armstrong, Stuart; et al. The Technological Singularity: Managing the Journey. Springer https://intelligence.org/files/TechnicalAgenda.pdf Retrieved 10.02.2016.

[4] http://www.nickbostrom.com/ Retrieved 10.02.2016.

[5] http://www.kurzweilai.net/ray-kurzweil-biography Retrieved 10.02.2016.

[6] https://en.wikipedia.org/wiki/Peter_Norvig Retrieved 10.02.2016.

[7] http://www-formal.stanford.edu/jmc/whatisai/node1.html Retrieved 10.02.2016.

[8] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.3043&rep=rep1&type=pdf Retrieved 10.02.2016.

[9] https://www.quora.com/How-intelligent-is-AI-now Retrieved 10.02.2016.

[10] Keith Frankish, The Open University, Milton Keynes; William M. Ramsey, University of Nevada, Las Vegas; “The Cambridge Handbook of Artificial Intelligence”; July 2014

[11] Asimov, Isaac; “I, Robot”; (1950);

[12] http://www.nickbostrom.com/ethics/artificial-intelligence.pdf Retrieved 10.02.2016.

[13] Bostrom, N. (1998). "How Long Before Superintelligence?" International Journal of Futures Studies, 2. http://www.nickbostrom.com/superintelligence.html Retrieved 10.02.2016.

[14] "Stephen Hawking warns artificial intelligence could end mankind - BBC News".BBC News. Retrieved 10.02.2016.

[15] "Will artificial intelligence destroy humanity? Here are 5 reasons not to worry.".*Vox*. Retrieved 10.02.2016.

[16] https://en.wikipedia.org/wiki/Artificial_intelligence#cite_note-205 Retrieved 10.02.2016.

Defined by Josip Horvat: 02.01.2016.

Delivered to Kalman Žiha: 13.01.2016

Josip Horvat, josip.horvat91@gmail.com
Klimaoprema d.d.
Gradna 78A, HR 10430 Samobor