Ethics and Safety in the Future of Artificial Intelligence: Remarkable Issues

Utku Kose*‡, Ibrahim Arda Cankaya*, Tuncay Yigit*

*Dept. of Computer Engineering, Faculty of Engineering, Suleyman Demirel University, Suleyman Demirel University, Dept. of Computer Engineering, Faculty of Engineering, E9 Block, West Campus, 32260, Isparta, Turkey

(utkukose@sdu.edu.tr , ardacankaya@sdu.edu.tr , tuncayyigit@sdu.edu.tr)

‡ Corresponding Author; Utku Kose, Suleyman Demirel University, Dept. of Computer Engineering, Faculty of Engineering, E9 Block, Z-23, West Campus, 32260, Isparta, Turkey, Tel: +90246 211 13 91, Fax: +90246 237 08 59, utkukose@sdu.edu.tr

This work was presented at the IDITAI Workshop 2017, which was held at Nisantasi University, Turkey on 28th December 2017.

Received: 18.05.2018 Accepted: 30.06.2018

Abstract- The humankind is currently experiencing a life supported often with intelligent systems designed and developed based on the foundations of Artificial Intelligence. It is clear that this scientific field is one of key elements for shaping better future for us. But there are also some anxieties regarding possible ethical and safety related issues that may arise because of intense use of powerful Artificial Intelligence oriented systems. In this context, objective of this paper is to provide a look at to some remarkable issues about ethics and safety within the future of Artificial Intelligence. After focusing on currently wide-discussed issues, the paper also comes with some possible solution suggestions for achieving a better Artificial Intelligence supported future with no or less issues on ethics and safety.

Keywords- artificial intelligence; future of artificial intelligence; ethical artificial intelligence; artificial intelligence safety; machine ethics

1. Introduction

Artificial Intelligence has an important and effective role on transforming the current and the future state of our life. That’s not surprisingly but rise of this field has a remarkable story behind, considering many developments appeared on the background. It is remarkable that especially revolutionary developments within computer, electronics and communication technologies have enabled scientific minds to improve momentum of innovative and practical technological developments and in this way, it has become more possible to think about real forms of previously imagined, science-fiction oriented technological products. Eventually, Artificial Intelligence has taken an increasing active role in humankind’s modern life.

Today, we can see – observe effects of Artificial Intelligence based systems almost all fields [1-3] In this context, it has already become a common think to see any intelligent mechanism in our cell phones, machines at home, and even cars. From a general perspective, it is possible to indicate that the Artificial Intelligence is making our life more practical and easier. Thanks to different approaches, methods, and techniques of Artificial Intelligence, it is now not impossible to solve advanced, complex problems or spend more time for solving such problems via traditional solution ways. Because of that, Artificial Intelligence seems promising a good, better future life for the whole humankind. But on the other side of the medallion, there are already serious discussions on a dystopian future of Artificial Intelligence with many important issues associated with ethical and safety oriented issues. Yet there is not any serious state of a worse future captured by intelligent machines but in at least a theoretical manner, some issues are widely discussed and being tried to be solved by researchers.

Considering the explanations provided so far, objective of this paper is to provide a look at to some remarkable issues about ethics and safety within the future of Artificial Intelligence. Because that is an important research interest currently, there is also an improving literature, which will probably be deeper and deeper in time because of the multidisciplinary aspects considered about Artificial Intelligence when ethical and safety oriented subjects are thought. So, that paper is believed to be an alternative contribution to the associated literature. After focusing on currently wide-discussed issues, the paper also comes with some possible solution suggestions for achieving a better Artificial Intelligence supported future with no or less issues on ethics and safety.

Based on its subject – objectives, the remaining content of the paper is organized as follows: The next section is devoted to a brief explanation of research interests having relations with the future of Artificial Intelligence. After that section, the third section explains the most remarkable issues when ethics and safety are thought in the context of Artificial Intelligence and the future. Following that section, the fourth section introduces some possible solution suggestions and finally the paper is ended by discussions on conclusions and some future work plans.

2. Future Of Artificial Intelligence and the Related Research Interests

When we think about the future of Artificial Intelligence, we should know some trendy research interests to understand what is currently discussed widely in the associated literature. That's critical to derive ideas about good or bad sides of intelligent systems of the future and design – develop possible solution ways in this manner. So, this section is about the related research interests that should be known essentially.

2.1. Machine Ethics – Ethical Artificial Intelligence

Machine Ethics is one of the most important issues considered when dealing with the future of Artificial Intelligence. In the literature, another concept: Ethical Artificial Intelligence is also used instead of Machine Ethics [4, 5]. As general, Machine Ethics is focused on research works trying to find appropriate answers for the problem scope: "the consequences of the behaviours, which are shown by machines to humans and other machines" [6]. In detail, research works done in this scope are based on the idea of defining ethical rules to prevent from any possible dangerous – harmful results (especially for the humankind) caused by intelligent systems [5, 7-9]. Because there is the situation of understanding what is ethical or not and what can be considered as ethical or safe by intelligent machines, this research interest is always keep in touch with different fields like sociology, psychology, and even education. On the other hand, the technical side is clearly connected with the essential fields of Artificial Intelligence, like mathematics, logics, electronics, and computer software – hardware…etc.

2.2. Artificial Intelligence Safety

Machine Ethics is about finding the ethical ways of realizing intelligent systems. But if we move from a dystopian fact that intelligent systems can always have some potential dangerous features - functions, then another research interest should be formed. At this point, the research interest of Artificial Intelligence Safety is focused on especially unsafe scenarios in which intelligent systems may have active role, so it deals with possible solutions to eliminate unsafe potentials or at least make intelligent systems controllable by us [10, 11]. Majority of research works in this manner are for designing and developing solutions for making intelligent systems safer or more controllable in case of any undesired situation. So, there is another puzzle to be solved here as which kind of situations are undesired and which kind of currently observed situations can lead us to think about something going actually wrong.

2.3. *Technological Singularity*

As a more utopian idea, which is also widely accepted by the scientific audience, Technological Singularity is briefly a hypothesis, which is based on the idea that the Artificial Intelligence and its products will overcome human intelligence in the future and eventually, civilizations and human nature will be transformed radically into different forms [12, 13]. This hypothesis is mostly connected with a better future supported by intelligent systems and accepts the ideas of improvements in a new world in which intelligent systems are some-how the ruler but not so dangerous for the humankind and other living organisms in an existential manner.

2.4. *Superintelligence*

Superintelligence is another research interest, which is attracting researchers' interest widely. Briefly, the concept of Superintelligence is based on the idea of an intelligence type making intelligent machines to "surpass human brain in general intelligence" [14]. Because of that, research works oriented in this interest are based on the systems having better intelligence form rather than humans including even the most intelligent one living. In this context, even systems that are able to solve problems that cannot be solved by humans are directly included under the scope of this research interest.

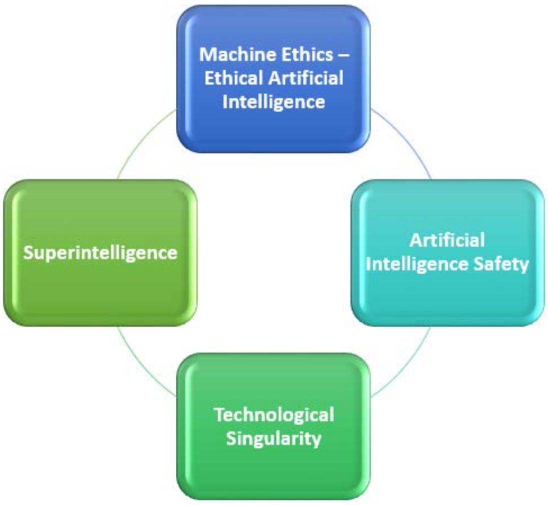

Machine Ethics – Ethical Artificial Intelligence
Artificial Intelligence Safety
Technological Singularity
Superintelligence

Figure 1. Some research interests related to future of Artificial Intelligence.

All the mentioned research interests are some-how associated with the ideas, predictions, improvements and developments done regarding the future of Artificial Intelligence. When the subject is examined in detail, there are also some wide scope research areas such as Existential Risks [15] and these research areas include also Artificial Intelligence in their topics because of its great influence in different fields and potential in directing the future of humankind, life and even universe.

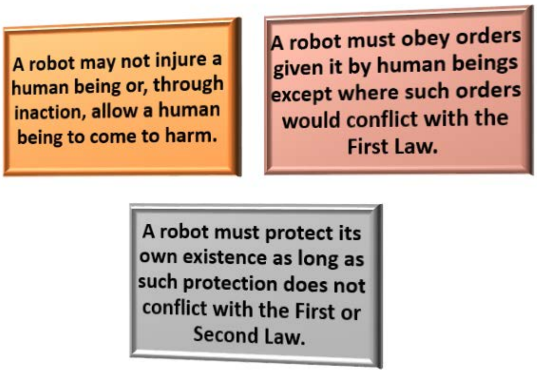
Considering the future of Artificial Intelligence, the next section briefly focuses on some remarkable ethical and safety oriented issues caused by Artificial Intelligence.

3. Ethical and Safety Oriented Remarkable Issues Caused By Artificial Intelligence

Some of remarkable issues associated with ethical and safety oriented factors caused by Artificial Intelligence can be explained - discussed briefly as follows:

3.1. Laws for Intelligent Systems

Because of the idea that any intelligent, autonomous system can behave dangerously after a while following many learning phases, it has been a commonly discussed problem to design some laws – rules, which the related systems should obey. That idea actually has some origins to the three laws expressed by Asimov in his famous science-fiction novels [16]. At this point, it is important to achieve some emergency oriented mechanisms to stop or at least direct robots – intelligent systems when it is observed that they are behaving in a not desired way. Because of the butter-fly effect, some previously gathered information from experiences or expert knowledge provided by humans may cause very dangerous situations at the end. The idea of providing laws at this point tries to find answers for the questions like 'Which laws should be applied for intelligent systems?', 'How can we design an algorithmic structure covering the designed laws for intelligent systems?', 'How can we know the provided laws are enough for meeting with all undesired situations and eliminating them well?'.



A robot may not injure a human being or, through inaction, allow a human being to come to harm.

A robot must obey orders given it by human beings except where such orders would conflict with the First Law.

A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

3.2. Moral Dilemmas

Moral dilemmas have always been important for people from different aspects of sociology, psychology, education and many other social sciences. But when intelligent machines, which can behave like a human, are taken into consideration, this research subject becomes a more important issue for having idea about future of Artificial Intelligence. As a famous example, accident dilemma considers the decision making mechanism for self-driving cars and focuses on the question of which scenario should be followed in case of any fatal accident including many members to be evaluated for who should be death, injured or be saved at the end of the accident [17]. Some other examples regarding moral dilemmas may include: 'Which patient should be treated by an intelligent (doctor) system in case of there are many patients having high level priorities at the same time.', 'At which level should a person be punished by an intelligent (judge) system if he / she killed anyone in order to save himself / herself or decides to achieve a justice for everyone being affected by the same factor dangerously or unjustly.', or 'How to behave in case of the situation mentioned by Montaigne: "one person's profit always involves another person's loss [18].'

3.3. Jobs to be done by Intelligent Systems

A both ethical and safety issue related to employment of Artificial Intelligence in the daily life is connected with the anxious among people that their jobs will be taken over by intelligent systems – machines, which are faster, more effective, efficient without even being tired. At this point, there is a serious discussion that how people will be in an economic situation if they lose their jobs and how it will be safe to give critical tasks to intelligent systems, which can be hacked and re-programmed by evil-minded sources (people, other intelligent systems – machines...etc.) [19].

3.4. *Hacked Intelligent Systems*

In addition to its advantages, running computer based systems has caused many disadvantages because of open doors of such systems. It is known that hackers are dangerous actors of the digital world with their unstoppable abilities and knowledge to manipulate any computer system and even people to reach their objectives or just having fun by causing harmful or dangerous situations at the end [20, 21]. With more developments in technologies and transformation to the society of informatics, roles of such people have become more important to be prevented from for achieving safety of especially information – data. Because the future of the world will be structured over mostly intelligent, autonomous systems, hacking such systems will be too critical for many safety reasons. Hacking such intelligent systems may cause not only losing data but also allowing evil-minded people or systems to re-program any system to behave according to different laws – directives. In a more general perspective, hacking a machine connected to a wide network of machines (Internet of Things: IoT) [22-24] may cause hackers to reach critical environments by using even simple machines, which are intelligent for specific tasks but having lots of open doors. That scenario can be improved by thinking about self-driving military vehicles, drones, war machines, or satellites, which are key elements for security of a nation – country.

3.5. *Producing Intelligent Systems*

In the future, it will be a common task to produce intelligent machines. Even today, many different companies like Tesla have already taken many steps in this manner. Industry 4.0 is a concept to define that revolutionary age and it will even probably transform production approaches into newer forms. But because there is the fact of machines, which can work – behave intelligently and autonomously, the production done here may be defined again by answering some questions rising. As related especially ethical subjects, one key answer that should be answered is about who will be responsible for any after-learned dangerous or harmful behaviors of such machines. That question also causes some additional questions to rise: ‘Should intelligent, autonomous machines have their lawful status like a human?’, or ‘How can an intelligent machine be responsible for its actions, decisions…etc.?’

3.6. Copyright Issues

Artificial Intelligence based systems are also used for creating different artistic products like pictures, poems, novels…etc. But at this point, there is a remarkable issue on copyright. One common question that should be asked here is: ‘Who is the copyright holder of an artistic product; the intelligent system or its developer?’ This issue is similar to the issue on producing intelligent systems but the main point considered here is about the products by intelligent systems.

3.7. Machines Created by Machines

It is also another ethical issue if it should be allowed one intelligent machine to create – develop other ones autonomously. That’s a rising issue because it is still unclear how after-learned, intelligently done behaviors can result to differences in new type of machines developed as benefiting from experiences – after-learned data of previous, ancestor machines.

4. Suggestions on Solution Ways

In addition to the mentioned ones in the previous section, there are many other ethical and safety oriented issues caused by Artificial Intelligence. Newer developments and improvements appeared in the context of Artificial Intelligence day-by-day cause us to derive and think about new issues that should be answered. At this point, there is always hope to think about also alternative solutions and at least such efforts are necessary to improve the associated literature covering such solution oriented works. Here, the authors also have some suggestions for eliminating or controlling the mentioned issues:

- For especially Machine Ethics, moral dilemma is a milestone to see if it is possible to achieve a desired moral – ethic behaviors in a machine – computer system based on Artificial Intelligence. In this sense, newer Artificial Intelligence techniques focused on only solving ethical problems with new methods or well-known problem solution methods (i.e. classification, clustering, pattern recognition, and optimization) should be designed and developed.
- It is important to design some laws for controlling working mechanism of intelligent, autonomous machines. At this point, it is suggested to design and develop complex algorithmic structures limiting abilities – behaviors of machines in different situations and also trying to eliminate moral dilemmas at the same time.
- In the literature of Artificial Intelligence Safety, researchers are working currently on designing safe agents, which are small but important parts of bigger problems solved carefully in detail. At this point, more research works to achieve the following types of agents should be done more [25-27]:
    - Interruptible Agents,

  - Ignorant Agents,
  - Inconsistent Agents,
  - Bounded Agents.

- As an important learning approach within the Machine Learning techniques of Artificial Intelligence, Reinforcement Learning [28] is widely discussed in the scientific community because of some dystopian scenarios on dangerous machines learned in a wrong way or via wrong feedback. In order to provide safe Reinforcement Learning, there is a technique called as Inverse Reinforcement Learning and this technique is known as an effective Artificial Intelligence Safety technique [29, 30]. Moving from that, it could be an effective way to run alternative safety oriented works over this technique and also improve it. Furthermore, it is of course possible to develop alternative techniques to deal with disadvantages of learning approaches like Reinforcement Learning.
- For a better, controllable Artificial Intelligence, it is necessary to form some professions and jobs. Such professions and jobs may include the ones like Artificial Intelligence engineer, Machine Ethics engineer, Artificial Intelligence safety expert, Artificial Intelligence training rule expert…etc.
- It is thought by the authors that it may be a good way to design a global, hierarchical structure defining priorities of humans, other living organisms and also intelligent, autonomous machines to prevent from any possible Existential Risks by Artificial Intelligence and also undesired situations of society transformations caused by i.e. Technological Singularity or any other form of changes in the future. Such structure can employ some laws – rules to be followed by all its members and in this way both ethical and safe living over the world is finally achieved.

5. Conclusions and Future Work

This paper has provided a general discussion and overview of ethical and safety oriented issues on Artificial Intelligence. When we consider the current situation, some issues are just imaginations and scenarios, which have not realized yet but they are all theoretically possible of course. Even some of currently observed technological developments are some critical signs of a need for controllable Artificial Intelligence and because of that there is an important number of researchers, who think that intelligent systems of the future are potential threats for our future. The discussed subject in this paper are just some remarkable examples regarding Artificial Intelligence in the future and it is clear that we are still desiring a future, which has lots of unclear, misty problems waiting to be discovered and solved.

Discussions provided in this paper are some typical findings and also ideas by the authors and all these research efforts done so far have encouraged them for some future works. Future works in this manner include development of some ethics and safety related algorithmic systems as alternative solution scenarios and also strong theories in order to support the mentioned research interests and issues, which include some gaps because of their age in the scientific arena.

References

[1] S. Russell, and P. Norvig, Artificial Intelligence: A Modern Approach. 1995, Prentice-Hall.

[2] M. Negnevitsky, Artificial intelligence: a guide to intelligent systems. 2005, Pearson Education.

[3] D. T. Pham, and P. T. N. Pham, “Artificial intelligence in engineering”. International Journal of Machine Tools and Manufacture, 39(6), 1999, 937-949.

[4] S. Russell, Ethics of artificial intelligence. Nature; London. 521.7553, 2015, 415-418.

[5] N. Bostrom, and E. Yudkowsky, The ethics of artificial intelligence. The Cambridge Handbook of Artificial Intelligence, 2014, 316-334.

[6] M. Anderson, S. L. Anderson, and C. Armen, Towards machine ethics. In Proceedings of the AOTP’04-The AAAI-04 Workshop on Agent Organizations: Theory and Practice, 2004.

[7] L. Muehlhauser, and L. Helm, The singularity and machine ethics. In Singularity Hypotheses (pp. 101-126). 2012, Springer Berlin Heidelberg.

[8] B. Hibbard, Ethical Artificial Intelligence. 2014, arXiv preprint arXiv:1411.1373.

[9] M. Anderson, and S. L. Anderson, “Machine ethics: Creating an ethical intelligent agent”. AI Magazine, 28(4), 2007, 15.

[10] R. V. Yampolskiy, Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In Philosophy and Theory of Artificial Intelligence (pp. 389-396). 2013, Springer Berlin Heidelberg.

[11] A. Pavaloiu, and U. Kose, “Ethical artificial intelligence-An open question. Journal of Multidisciplinary Developments, 2(2), 2017, 15-27.

[12] R. Kurzweil, The Singularity is Near: When Humans Transcend Biology. 2005, Penguin.

[13] B. Goertzel, “Human-level artificial general intelligence and the possibility of a technological singularity: A reaction to Ray Kurzweil's The

Singularity Is Near, and McDermott's critique of Kurzweil". Artificial Intelligence, 171(18), 2007, 1161-1173.

[14] N. Bostrom, Superintelligence: Paths, dangers, strategies. 2014, OUP Oxford.

[15] N. Bostrom, "Existential risks". Journal of Evolution and Technology, 9(1), 2002, 1-31.

[16] MIT Technology Review. Do We Need Asimov's Laws?. 2014, TechnologyReview.com. Retrieved December 3, 2017, from https://www.technologyreview.com/s/527336/do-we-need-asimovs-laws/

[17] The Associated Press, For Driverless Cars, a Moral Dilemma: Who Lives and Who Dies?, 2017, NBC News Web Site. Retrieved December 3, 2017, from http://www.nbcnews.com/tech/innovation/driverless-cars-moral-dilemma-who-lives-who-dies-n708276

[18] M. R. Montaigne, The Complete Essays of Montaigne. 1958, philpapers.org.

[19] S. Bernezzani, 10 Jobs Artificial Intelligence Will Replace (and 10 That Are Safe). 2017, HubSpot.com. Retrieved December 2, 2017, from https://blog.hubspot.com/marketing/jobs-artificial-intelligence-will-replace

[20] P. A. Taylor, Hackers: Crime in the digital sublime. Psychology Press, 1999.

[21] T. Jordan, and P. Taylor, "A sociology of hackers". The Sociological Review, 46(4), 1998, 757-780.

[22] F. Xia, L. T. Yang, L. Wang, and A. Vinel, "Internet of things". International Journal of Communication Systems, 25(9), 2012, 1101.

[23] H. Kopetz, Internet of Things. In Real-time systems (pp. 307-323), 2011, Springer US.

[24] F. Wortmann, and K. Flüchter, "Internet of things". Business & Information Systems Engineering, 57(3), 2015, 221-224.

[25] O. Evans, and N. D. Goodman, Learning the preferences of bounded agents. In NIPS 2015 Workshop on Bounded Optimality, 2015.

[26] O. Evans, A. Stuhlmüller, and N. D. Goodman, Learning the preferences of ignorant, inconsistent agents. 2015, arXiv preprint arXiv:1512.05832.

[27] L. Orseau, and S. Armstrong, Safely interruptible agents. In Uncertainty in Artificial Intelligence: 32nd Conference (UAI 2016), edited by Alexander Ihler and Dominik Janzing, 2016, (pp. 557-566).

[28] R. S. Sutton, and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge: MIT Press, 1998.

[29] P. Abbeel, and A. Y. Ng, Inverse reinforcement learning. In Encyclopedia of machine learning (pp. 554-558). 2011, Springer US.

[30] A. Y. Ng, and S. J. Russell, Algorithms for inverse reinforcement learning. In Icml, 2000, (pp. 663-670).