Are We Safe Enough in the Future of Artificial Intelligence? A Discussion on Machine Ethics and Artificial Intelligence Safety[7]

Utku Köse
Suleyman Demirel University, Isparta, Turkey
Çünür Mahallesi, Süleyman Demirel Caddesi, Merkez/Isparta, Tel.: +90 246 211 10 00
utkukose@sdu.edu.tr

Abstract

Nowadays, there is a serious anxiety on the existence of dangerous intelligent systems and it is not just a science-fiction idea of evil machines like the ones in well-known Terminator movie or any other movies including intelligent robots – machines threatening the existence of humankind. So, there is a great interest in some alternative research works under the topics of Machine Ethics, Artificial Intelligence Safety and the associated research topics like Future of Artificial Intelligence and Existential Risks. The objective of this study is to provide a general discussion about the expressed research topics and try to find some answers to the question of 'Are we safe enough in the future of Artificial Intelligence?'. In detail, the discussion includes a comprehensive focus on 'dystopic' scenarios, enables interested researchers to think about some 'moral dilemmas' and finally have some ethical outputs that are considerable for developing good intelligent systems. From a general perspective, the discussion taken here is a good opportunity to improve awareness on the mentioned, remarkable research topics associated with not only Artificial Intelligence but also many other natural and social sciences taking role in the humankind.

Keywords: artificial intelligence, machine learning, machine ethics, artificial intelligence safety, future of artificial intelligence

1. Introduction

Artificial Intelligence is an important scientific field, which is currently shaping the future of humankind. After it has started its journey as a result of innovative developments within Computer Science, it has provided many effective and efficient solutions to different fields, and has then taken an active part rapidly in the scientific arena with its great potential on solving real-world based problems (Negnevitsky, 2005; Russell et al., 2003). Today, intelligent solutions are often employed in daily modern life by taking part in even our phones, cars or houses. In a general manner, there is an infinite loop of innovations, which is a typical result of collaboration caused by some popular scientific fields and the field of Artificial Intelligence is one of them. Considering the unstoppable rise of intelligent solution approaches, methods and techniques of Artificial Intelligence, there has always been some discussions regarding the disadvantages and unclear characteristics of intelligent systems in the context of a life surrounded always by also ethical and morality-oriented issues (Allen et al., 2000; Bostrom, 2003; Dubhashi, & Lappin, 2017; Floridi, & Sanders, 2004; Herzfeld, 2002). But the more advanced problems have been solved by intelligent systems, the more discussions over Artificial Intelligence have appeared.

Currently, people are often anxious about the possible existence of dangerous intelligent systems and that anxiety seems improving more with alternative discussions associated with different fields of the modern life. By moving from such anxiety-oriented questions, there is a great interest in some alternative research works under the topics of Machine Ethics or Artificial Intelligence Safety. Generally, objectives of such research works are mainly focused on designing and developing human-compatible intelligence systems or at least introducing some alternative solution ways to adjust badly configured intelligent systems. In more detail, the related works associated with developing 'safe' intelligent systems are also examined under some other alternative research topics like Existential Risks because Artificial Intelligence is accepted as one of the most

[7] This work is the keynote speak made by Utku Kose at the SMART 2017 – Scientific Methods in Academic Research and Teaching International Conference 2017 held in Timişoara, Romania, between September 8 and September 9, 2017.

remarkable risks having a threating potential towards humankind and even other living organisms. As combining all of these research efforts in a common ground, Future of Artificial Intelligence is known as another research topic dealing with the future build over intelligent machines.

Considering the explanations made so far, the objective of this study is to provide a general focus on the expressed research topics and try to find some answers to the question of 'Are we safe enough in the future of Artificial Intelligence?'. In detail, the discussion includes a comprehensive focus on 'dystopic' scenarios, enables interested researchers to think about some 'moral dilemmas' and finally have some ethical outputs that are considerable for developing good intelligent systems. From a general perspective, the discussion taken here is a good opportunity to improve awareness on the mentioned, remarkable research topics associated with not only Artificial Intelligence but also many other natural and social sciences taking role in the humankind.

In the context of the objectives of this study, the remaining content is organized as follows: The next section provides some brief information regarding considered research topics like Machine Ethics, Artificial Intelligence Safety, and other associated ones. After that section, the third section discusses some anxiety points connected with possible dangerous scenarios or moral dilemmas. Following that, the fourth section tries to answer the main question of the study by providing ideas about ethical outputs to develop good intelligent systems and finally the study is ended with discussions about conclusions and future work ideas.

2. Research Topics Associated with Anxieties on Artificial Intelligence

Anxieties related to Artificial Intelligence have caused the rise of different research topics in time. Of course, these research topics do not always depend on just anxieties, but they also include research efforts for solving mysteries lying on the background of Artificial Intelligence and providing more findings for a better future shaped by intelligent systems. The scientific literature seems always open for introduction of new research topics and efforts within them but some remarkable ones taking researchers' interest widely today can be explained briefly as follows:

2.1. Machine Ethics

Machine Ethics is a research topic in which researchers' deal with ethical issues that may appear while intelligent systems are solving some problems. These problems are generally associated with the existence of humankind or need very critical analyses to provide the most accurate solution at the end. In the literature of Artificial Intelligence, a similar term: Ethical Artificial Intelligence can also be used to meet the same research efforts, but Machine Ethics is focused more on an exact intelligent system that is something like a robot or an advanced machine, which is strong enough to solve more complex problems that cannot be solved yet with current capabilities of intelligent systems. So, this research topic has connections with both current developments and possible future developments in the field of Artificial Intelligence (Allen et al., 2006; Anderson, & Anderson, 2011).

In Machine Ethics, researchers often encounter with issues that are both philosophical and technical (Powers, 2011). Issues covered in Machine Ethics are generally because of the following subjects having ethical conflicts in the context of Artificial Intelligence (Anderson, & Anderson, 2011; Lin et al., 2011; Torrance, 2008; Wallach, & Allen, 2008):

- **Rights:** What kind of rights should an intelligent machine have while solving problems for us? If an intelligent system causes an accident or show dangerous behaviors while working on its tasks – jobs, which rights should belong to it in order to have accurate judgments about it? More generally, should a robot (intelligent system) have rights like us? These are all important questions that are rising when we give an ethical view on rights for intelligent systems.
- **Duties:** It is certain that we need advanced intelligent systems to solve real world problems in more accurate, rapid and effective ways. So, all these needs require intelligent systems to have duties to be completed. But what if not all kind of duties

should be given to intelligent to be completed? Sometimes, it may not be clear to see the end of some duties, which require some human oriented behaviors, which even cannot be simulated enough by Artificial Intelligence. While it is still an unsolved problem that how Artificial Intelligence can be completely be like us in terms of spiritual and mental manner, it may not be ethical to have it to be responsible for all duties.

- **Human Welfare:** When we consider about our welfare, ethical issues rise in the way, which may cause lots of paradoxes to be analyzed. Keeping human welfare at an optimum level may be a big problem for robots because it includes an almost infinite number of environmental factors to be evaluated because of dynamic nature of the human and also factors (i.e. nature, world), he / she is actively interacting. Some dangerous behaviors provided by Artificial Intelligence may be some important steps that should be taken to save the human welfare for the future, but humankind cannot understand the ethical causes lying deep inside of that or ethical rules of a very-well trained – learned, advanced intelligent system cannot then be same / similar with the ones we know as ethical ones.
- **Justice:** How can an intelligent system apply justice when it encounters with events including ethical issues (i.e. killing someone to save yourself, causing some people to be dead because of somebody else's faults, crimes hidden very well on the background of laws)? There are many issues because of the loopholes written laws; how can a robot in the role of a judge deal with cases? If we move away from the arena of the law, we can also derive questions about daily life: How can an intelligent system managing a shop apply the justice to its employers when these employers have some problems in an ethical manner? What does it mean for a robot to be fair when a duty assignment problem should be solved ethically considering all other robotic workers' past performances at an intelligent factory?

In the context of Machine Ethics, the following three remarkable concepts can be discussed widely to have some more ethical ideas regarding Artificial Intelligence (Hibbard, 2014; McLaren, 2006; Schneider, 2016; Wallach, 2010):

- **Information:** It is accepted that the control of information which will be used to train intelligent systems or to which intelligent systems can use are important to ensure ethical intelligent systems. In detail, the amount of information, and even its value are also important issues evaluated within Machine Ethics.
- **Control:** It is believed that effective and accurate approaches can enable people to control intelligent systems to behave ethically. Because Artificial Intelligence is a product of the human-mind, its control can be achieved better only by humans. That is also because only we can understand what kind of human errors may be included in intelligent systems and how such systems should be controlled to behave according to the ethical rules we believe.
- **Reasoning:** In intelligent systems (robots), reasoning is automated, and it has a flexible nature because of the continuing training process directed by environmental factors and experiences of the system more generally. At this point, one important question in the context of Machine Ethics is: how far should we trust an automated reasoning? Furthermore, it is still unclear if any triggering 'bad' training data may affect the automated reasoning of an intelligent system so that it may act harmful to its environment. At this point, how can we understand if a training data is appropriate to make an intelligent system more ethical? The reasoning concept here comes with many alternative questions needing to be answered carefully.

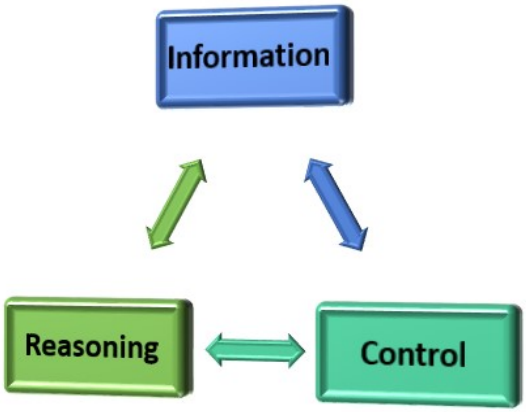

Figure 1. Three remarkable concepts discussed widely in the context of Machine Ethics.

Research works under the Machine Ethics also focus on an important concept: 'consciousness'. Here, the essential ethical problem is associated with the question: 'Can intelligent systems (robots) have consciousness?'. Moving from that, we can have some arguments such as (Holland, 2003; Torrance, 2008; Wallach, 2010):

- Actually, an intelligent system cannot take responsibility of its actions. So, are the developers responsible for that? Any human, company…etc.?
- How far can responsibility be connected with us (humans)?
- Is it enough for the humankind to deal with some certain real-world problems? Is it impossible for intelligent systems to have enough consciousness level to deal with all real-world problems?
- Should intelligent machines be considered in the context of what we actually understand from consciousness?
- If an intelligent system (robot) has consciousness; (1) 'Should it have rights like us?' (2) 'What if it wants more things from the humankind?' (3) 'How far we can trust its consciousness?' (4) 'Could a conscious system work for us?' (5) 'Can we have the right to turn off a conscious intelligent machine?'...etc.
- If we can achieve the consciousness in an intelligent system, should we consider any other additional factors affecting that consciousness?
- Is it really necessary for an intelligent system to have consciousness? Should we design something like consciousness but a more specific thing special for intelligent systems?

Intelligent agents are widely discussed in the context of Machine Ethics. Here, the objectives are all connected with ensuring an ethical agent. Related to that ethical agent issue, the literature also defines two types of ethical agents according to their focus – design on ethics (Anderson, & Anderson, 2007; Moor, 2009):

- **Implicit Ethical Agents:** Agents that are limited by their designers to prevent from unethical actions – behaviors.
- **Explicit Ethical Agents:** Agents that have special algorithms to make them acting – behaving ethical.



Figure 2. Essential ethical agents according their focus – design on ethics.

Explanations so far are based on philosophical and theoretical approaches on ethical issues that can be associated with Artificial Intelligence and its outputs. Researchers thinking that Artificial Intelligence can be dangerous (unsafe) because of some reasons find themselves in designing algorithmic approaches to eliminate possible side-effects causing unsafe intelligence systems and ensure a safe Artificial Intelligence environment. Research works associated with such efforts are included under the research topic of Artificial Intelligence Safety.

2.2. Artificial Intelligence Safety

Artificial Intelligence Safety is associated with the idea of that it is possible to encounter with unsafe – dangerous intelligent systems if effective precautions are not taken. So, before it is too late, alternative approaches, methods, and techniques should be designed to obtain safe intelligent systems (robots in the future). In the literature, Yampolskiy (2013) expresses that researchers have some ideas leading them to misunderstand the concept of Machine Ethics and an alternative research topic to ensure safety in Artificial Intelligence should be followed. Such ideas have caused to think about developing safe Artificial Intelligence based systems (Pavaloiu, & Kose, 2017). It is also remarkable that although humankind has the highest priority in terms of safety, Artificial Intelligence Safety is interested in dangerous or harmful actions – behaviors that can be shown by intelligent systems to humans, other living organisms, and also other intelligent systems (Figure 2).

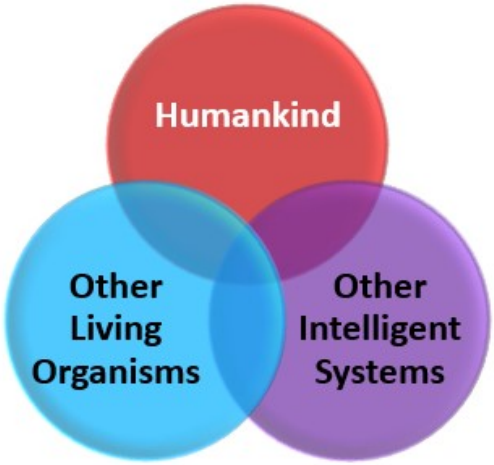

Figure 3. Objective safety factors in the context of Artificial Intelligence Safety.

When we think about a safe intelligence system, the following questions and many more often wait to be answered accurately with effective research efforts:

- Should there be a 'red button' to stop any intelligent system when its actions start to be dangerous or harmful?
- How can we develop a red button for preventing intelligent systems from turning to the 'dark side'?
- How can we stop an intelligent system to stop us from pressing to the red button when it learns enough about it and its effects?
- How can we 'limit' an intelligent system to enable it only for its duties?
- How can we prevent intelligent systems from being hacked by other intelligent systems or hackers?
- Can we predict future behaviors of an intelligent system?

Nowadays, there are some remarkable solutions developed for ensuring safe Artificial Intelligence based systems. By focusing on agent models of Artificial Intelligence, there are some alternative agent types introduced as 'safe intelligent systems'. Also, a widely-used Machine Learning approach: Reinforcement Learning is also considered in the form of an alternative learning approach in order to ensure safe intelligent systems. Some essential information regarding these

developments can be expressed briefly as follows (Abbeel, & Ng, 2011; Evans, & Goodman, 2015; Evans et al., 2016; Lin et al., 2017; Ng, & Russell, 2000; Orseau, & Armstrong, 2016):

- **Interruptible / Ignorant / Inconsistent / Bounded Agents:** Interruptible Agents stands for safe agents, which were introduced as having a 'red button' to be stopped in case of any dangerous state. On the other hand, Ignorant / Inconsistent Agents are based on the idea of that people are ignorant and inconsistent about their choices in the real life so intelligent agents should be like that in order to have ethically and safely structured intelligent systems. Finally, Bounded Agents are based on designing optimal, frugal and fast agents, which are heuristics for next decisions – plans as by inspiring from the fact that humans are both biased and bounded cognitively.
- **Inverse Reinforcement Learning:** Inverse Reinforcement Learning is a remarkable approach to determine a reward function for defining behaviors of the objective agent – system so that it can be possible to also find good policies. That means also ethical and safe ways in terms of considered research topics.

In the literature, another remarkable research effort is associated with training data, which can be connected with the issue of information and the related factors, as mentioned before. In detail, the concept of 'adversarial examples' is used to define training data, which can confuse intelligent systems and make them learn something wrong (Goodfellow et al., 2017). So, that research subject is highly associated with Artificial Intelligence Safety, because of its potential to be useful for hackers of the future whether they will be human or other intelligent machines. Examples provided in the literature show that small changes – additions in training data can cause Artificial Intelligence to understand false about the learned problem – solution easily and such problem is very critical about safety level of a developed intelligent system.

Some other concepts – issues discussed in the literature can be listed as: 'rationality', 'corrigibility', 'value alignment', 'human-aligned Artificial Intelligence', 'openness', 'expected utility', 'reward', 'reward engineering' (Amodei et al., 2016; Arnold et al., 2017; Conitzer et al., 2017; Dewey, 2014; Riedl, & Harrison, 2016; Russell et al., 2015; Soares et al., 2015; Vamplew et al., 2017).

Machine Ethics and Artificial Intelligence Safety have no certain borders between them. Machine Ethics is even called as also Ethical Artificial Intelligence (Hibbard, 2014). At this point, research perspectives related both research topics can be even accepted in a common research field employing works for ethical and safe intelligent systems. On the other hand, all these works can be connected to the efforts of shaping a 'good future' of intelligent systems. So, another research topic: Future of Artificial Intelligence rises here.

2.3. Future of Artificial Intelligence

Increasing anxieties are always because humankind wants a future, which is supported by intelligent systems, which are human-compatible and safe. So, except research works focused on developing ethical and safe intelligent systems, there is also a need for discussing future developments regarding Artificial Intelligence. More ideas about how intelligent systems of the future will be like are essential for researchers to derive effective solutions for ethical and safe Artificial Intelligence. Because of that, a research topic called as Future of Artificial Intelligence deals with views over the future of Artificial Intelligence.

Today, there are different kinds of research interests introduced in the context of Future of Artificial Intelligence. With their remarkable and logical background shaped by both technical and social aspects, two of them attract researchers interest widely. They are:

- **Technological Singularity:** Also known as 'Singularity', this is a hypothesis indicating that Artificial Intelligence, which is better than human intelligence will shape the humankind and civilizations greatly in the future (Callaghan et al., 2017; Muehlhauser, & Helm, 2012; Nicolescu, 2017).

- **Superintelligence:** It is both term and a research interest, which is explaining an intelligence, which is surpassing even the highest level human brain in the context of general intelligence. Actually, many ethical and safety-oriented issues are also associated with possible, future intelligent systems having 'superintelligence.' (Bostrom, 2014; Dubhashi, & Lappin, 2017; Perdue, 2017).

Arguments defending the idea of 'bad Artificial Intelligence' has led researchers to classify Artificial Intelligence under Existential Risks, which is a separate research topic focusing on factors affecting the existence of the humankind.



Figure 4. Two research interests that can be considered under the Future of Artificial Intelligence.

2.4. Existential Risks

Existential Risks is not directly associated with ethical or safety issues about Artificial Intelligence, but it is related to Artificial Intelligence because it accepts that field by considering two important characteristics. According to that research topic, an existential risk has a 'global effect' covering the whole humankind and a 'terminal intensity' reducing the surviving potential of the considered risk group through even new generations (Bostrom, 2002). As a result of developments and increasing anxieties, these two characteristics are also associated with Artificial Intelligence. In detail, this corresponds to a possible bad Artificial Intelligence, which can cause the existence of the humankind to end. So, that research topic has important relations with Machine Ethics and especially Artificial Intelligence Safety.

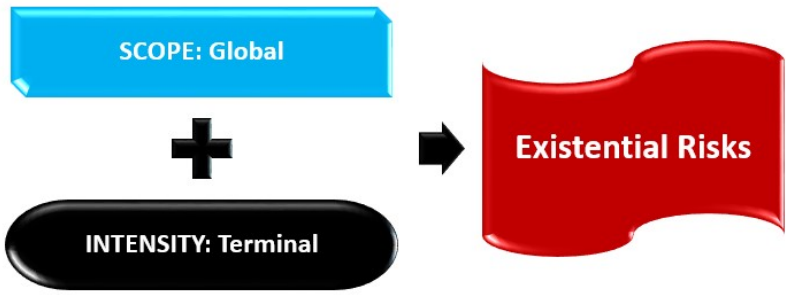

Figure 5. Two characteristics of an existential risk.

Except from Artificial Intelligence, Existence Risks deals with many more alternative risks as presented in Figure 6.

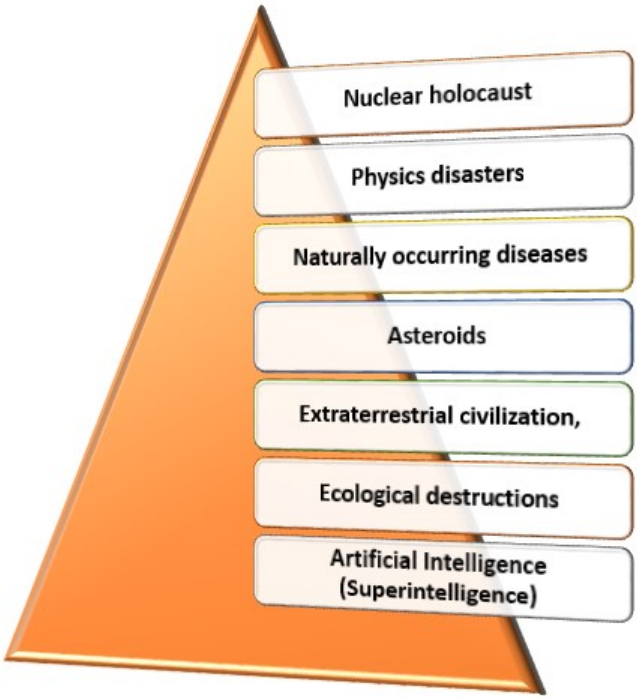

Figure 6. Existential Risks.

3. Possible Scenarios Causing Questions

As being research subjects of the explained research topics, there are already some possible scenarios, which enable us to think more about if we are safe enough for the future supported by Artificial Intelligence. In this context, especially remarkable scenarios can be explained briefly as follows:

3.1. Moral Dilemmas

Moral dilemmas have been always important for us. Because moral dilemmas can easily confuse even us, they become more important problems to be solved in the sense of Artificial Intelligence. Some possible scenarios of moral dilemmas can be expressed as follows by considering intelligent systems:

- How can a self-driving car decide who should live and should die in case of a fatal accident, which has no possibility to leave everyone alive at the end?
- How should a person be punished by the intelligent judge, if that person has killed somebody in order to save himself / herself?
- Which patient should be treated by the intelligent doctor, if there are many patients with high level priorities at the same time?
- Is it fair for an intelligent system to take any human's work from him / her and leave him / her unemployed?

3.2. Artificial Intelligence to be Created by Artificial Intelligence

"An intelligent system company has introduced its new product, which has an Artificial Intelligence to create similar intelligent systems to solve some problems better with collaborative approaches done by an intelligent swarm."

That remarkable scenario is associated with a future including intelligent systems, which are created by other intelligent systems. Except for religious discussions and anxieties, this scenario is important for researchers because it is not clear that if the creation of an intelligent system by other ones can cause more dangerous results at the end. The issue of allowing an intelligent system to

create new ones is also an important problem to be solved in terms of both Machine Ethics and Artificial Intelligence Safety. The problem improves more and more if the creator intelligent system can employ dangerous and harmful training – learning data, which may cause all created intelligent systems to be all dangerous – harmful in the future.

The issue of creation is similar in also copying intelligent systems. It is clear that the most important advantage of Artificial Intelligence is being easily copied or cloned to create intelligent systems with same characteristics or having newer ones with some different characteristics. At this point, it is important to also consider future opportunities for creating new intelligent systems easier and faster. But it is also an important question to ask if such fast flow can cause disadvantages (uncontrollable creation – copying – cloning, uncontrollable training – learning data etc.).

3.3. Jobs to be Done by Artificial Intelligence

'The Omega is a good, intelligent system (factory worker), which can perform a producing task effectively in less than an hour. Because of that, there is no need for any human at the factory so that humans are even not allowed to enter to the factory area.'

It is a widely discussed scenario that our jobs will be taken by intelligent systems in the near future. At this point, there are serious anxieties – questions about how the economic situation and humans' welfare will be like if there will be no jobs (or a few number of jobs to be done by humans) in the future. Because employment is an important factor affecting many additional components like the economy, social life, self-confidence, career etc., a future including intelligent workers and not working humans can be a new age, which has a very misty way ahead, especially considering humankind.

3.4. Social Life Transformed by Digital Individuals

That scenario can be possible in the future if Artificial Intelligence has enough capabilities to take more part in human's social life: '*John meets every day with his virtual friends to discuss some Philosophy and Rock Music. He does not need any human friends because his virtual friends make him improve his knowledge.*'

It is already a certain problem that digital world has some negative effects on our social life with addictive applications of social media and more interactive computer games and computer-oriented applications. At this point, advanced, intelligent, digital individuals can take place in our world and transform the current social life into another, unique form. Such social life may cause humans to lose their exact cultural and emotional roots in time and even make humans architects of a future with bad intelligent systems, which is a typical dystopia for now.

3.5. Rights and Copyright Problems

'For last 100 years, there is no human, who is better than intelligent painters in that painting competition. But a human painter thinks that this year's winner has benefited from his paintings, which was shared before him over the Internet. The human painter thinks that the system has even stolen his painting style.'

That scenario shows that rights and copyright issue will be very unclear in the future in which intelligent systems take an active role. Because an Artificial Intelligence based system has all abilities to reach all types of data from anywhere and anytime. In the end, the system then can use its almost infinite resources to make it even the most successful painter, poet, or an author. But should that prevent humans from sharing all their knowledge and abilities with Artificial Intelligence? Which type of judicial adjustments should be made in order to eliminate all such possible scenarios having conflicts in it. In more detail, we can think about a song that is created by an intelligent system. At this point, can the intelligent system have rights over copyrights regarding that song or does copyright belong to the creator – developer of that intelligent system? All such questions should be answered accurately in the context of research topics like Machine Ethics and also Artificial Intelligence Safety (because of the issue of using information from a human without any permission because of the open Internet environment.).

4. Some Solution Suggestions

By inspiring from the performed previous works and analyzing essential issues discussed in this study so far, it is also possible to think some solutions in order to develop 'good' intelligent systems for the future. Some of them can be mentioned briefly as follows:

- In order to prevent conflicts that may appear between the role of humans, living organisms and intelligent systems (robots) over the world, a global hierarchical 'tree of priorities' can be designed. Of course, control – management of such structure should be done carefully by appropriate authorities.
- It can be seen from the literature that there is a good potential within agent-based systems to design ethical and safe Artificial Intelligence based structure. At least, algorithmic and mathematical models formed on the background of such systems are important factors for researchers to improve the associated literature for better further works with current findings. So, there should be more focus on developing ethical and safe agents with alternative research efforts. At this point, it could be a good idea to start by moving from philosophies lying on the background of Implicit and Explicit Ethical Agents and even inspiring from already introduced ethical and safe agent systems (like Interruptible, Ignorant, Inconsistent or Bounded Agents). Here, more mathematical and logical approaches should be employed in order to eliminate all possible paradoxes and gaps that may cause unsuccessful research attempts. Because of that, such development studies should be done in a multidisciplinary manner including researchers from Computer Science, Mathematics, Logic, Philosophy and even social sciences focused on the human, like Sociology or Education.
- In addition to the suggestions expressed for developing alternative agents, it can also be a good way to classify Artificial Intelligence oriented systems with some 'dangerous levels'. In this way, it can be possible to use some systems in only certain tasks and fields, which will enable to run a 'limited task scope' for such systems to prevent them from behavior changing or any actions that may cause harmful results at the end.
- Moral dilemmas will be always an open question for Artificial Intelligence because of changing dynamics of the world, society and fields in which we currently working, studying etc. So, research works trying to determine accurate behaviors of intelligent systems for such moral dilemmas should be always active and an intense consideration should be given to such research interests for shaping especially ethical background of the Artificial Intelligence.
- A certain set of rules in algorithmic structures of Artificial Intelligence based systems can be designed in order to make them human and living organisms-compatible. Here, three laws designed by Asimov (MIT Technology Review, 2014) in his widely-read science-fiction novels can be a good point to move from. Asimov's rules on robots seem still strong enough to cover all logical background on achieving an ethical and safe system. The related rules can be expressed as follows (MIT Technology Review, 204): (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 Law (1)." (3) "A robot must protect its own existence as long as such protection does not conflict with the Law (1) or Law (2)."
- Changing nature of the society and more effects of technology causes newer jobs and specialties to appear. At this point, there will be probably new kind of jobs associated with Artificial Intelligence in even a near future. Considering ethical and safety-oriented requirements for Artificial Intelligence, it is possible to suggest some jobs as

follows: 'Training Data Analyzer for Intelligent Systems', 'Ethics and Safety Analyzer for Intelligent Systems', 'Agent Expert', 'Machine Learning Engineer', 'Safe Artificial Intelligence Engineer', 'Artificial Intelligence Algorithm Designer', 'Safe Robot Hardware Designer', 'Robot Safety Expert'.

- Safety oriented works are generally focused on algorithmic ideas and efforts surrounded by them. But it can be also a good way to focus on how to design hardware components of intelligent systems (in terms of especially robots). In other words, robots (intelligent systems), which may have the potential of being dangerous or showing complex actions can be designed with safer designs and materials, which will give some advantages of being prevented from harmful results that may even cause because of some accidents.
- As it was indicated before, information is one of the most remarkable factors for determining how an intelligence system can be ethical or safe. So, it is important to perform more research works about how to represent information for Artificial Intelligence and how to eliminate small information that may cause 'butterfly-effect' at the end and transform the system into a 'bad one'. In this context, it is important also to evaluate learning – training approaches used for obtaining desired intelligent systems, which can just do their tasks and be aware of any dangerous information (data) affecting its nature. Currently, it seems that the Reinforcement Learning is a good learning – training approach to be analyzed in detail but the future will be always open for introduction of newer approaches of course.
- Optimization lives in the heart of real life. All real-world problems can be solved effectively with optimization because they cannot be modeled mathematically. By inspiring from the optimization idea, it can be possible to design and develop some 'regulator' techniques for controlling and directing interaction of intelligent systems with the real world, which means active learning – training for them. In this way, simpler but effective enough solutions can be obtained in order to prevent such systems from evolving into bad systems or being affected by malicious factors. At this point, it is important to model which kinds of variables, constants and even constraints can be applied for certain types of intelligent systems developed with different techniques.
- Sometimes, details that are seen not too much important can be seen by some others and may lead scientific developments to different ways. Theories here are key elements to achieve the related mechanism. It can be suggested here to develop new theories on ensuring ethical and safe Artificial Intelligence. Such theories can be drawn with a multidisciplinary view combining different minds from different fields, which is interested in role of Artificial Intelligence in the world of the future. As a recent theory, readers are referred to the 'Fading Intelligence Theory' (Kose, & Vasant, 2017).
- There are anxieties regarding current and future state of Artificial Intelligence because the humankind is thought to be at the heart of the world and the life. Although our future seems to be depended on only our decisions, there are many environmental factors affecting the whole life. So, it may not be always necessary to enable robots (intelligent systems) to do some tasks or take important roles on certain sides of the life. In order to keep a balance between the technology and the living organisms over the world (and even in the universe), it can be a better way to design an optimum 'tasks sharing'.

5. Conclusion and Future Work

This study has provided a general discussion about ethical and safety-oriented issues regarding Artificial Intelligence. In detail, some remarkable research topics followed widely by the scientific community have been generally explained and a light has kept over possible dangerous

scenarios to understand more about what kind of issues are remarkable. At this point, the study is structured around the question of 'Are we safe enough in the future of Artificial Intelligence?' Along the study, the dangerous scenarios are also tried to be met with some possible solution suggestions. The study has been also organized as a reference work for enabling and directing interested readers to deeper sides of the literature.

The associated literature in this manner still has a misty way ahead but many important developments have been done so far since first anxieties related to possible dangerous and harmful effects that may be seen in time with appearance of more advanced, superintelligence-based systems. It seems that improvements in supportive technologies will possibly cause also many rapid and advanced developments in the field of Artificial Intelligence, but still algorithmic and mathematical structures have important role on solving even the most complicated issues that we may encounter with even in the future.

In addition to that study, the author is encouraged to take research done more steps away. In this sense, it is aimed to design and develop some algorithmic solutions for ensuring safe intelligent systems. Such solutions will be probably applied in terms of multi-agent systems. On the other hand, there will be also further studies to understand more about nature of training data to have more accurate ideas about how it can be possible to prevent a training data from 'malicious additions' or additional data that will probably cause great changes in reasoning of an intelligent system. Finally, additional theoretical works regarding Machine Studies and Artificial Intelligence Safety to improve the associated literature will also be done by the author.

References

Abbeel, P., & Ng, A. Y. (2011). Inverse reinforcement learning. In *Encyclopedia of machine learning* (pp. 554-558). Springer US.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251-261.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics?. *IEEE Intelligent Systems*, *21*(4), 12-17.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, *28*(4), 15.

Anderson, M., & Anderson, S. L. (Eds.). (2011). Machine Ethics. Cambridge University Press.

Arnold, T., Kasenberg, D., & Scheutz, M. (2017). Value alignment or misalignment–what will keep systems accountable. In *3rd International Workshop on AI, Ethics, and Society*.

Bostrom, N. (2002). Existential risks. *Journal of Evolution and Technology*, 9(1), 1-31.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 277-284.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford.

Callaghan, V., Miller, J., Yampolskiy, R., & Armstrong, S. (2017). Technological Singularity. Springer.

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral Decision Making Frameworks for Artificial Intelligence. In *AAAI* (pp. 4831-4835).

Dewey, D. (2014). Reinforcement learning and the reward engineering principle. In *2014 AAAI Spring Symposium Series*.

Dubhashi, D., & Lappin, S. (2017). AI dangers: Imagined and real. *Communications of the ACM*, *60*(2), 43-45.

Evans, O., & Goodman, N. D. (2015). Learning the preferences of bounded agents. In *NIPS Workshop on Bounded Optimality*(Vol. 6).

Evans, O., Stuhlmüller, A., & Goodman, N. D. (2016). Learning the Preferences of Ignorant, Inconsistent Agents. In *AAAI* (pp. 323-329).

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349-379.

Goodfellow, I., Papernot, N., Huang, S., Duan, Y., Abbeel, P., & Clark, J. (2017). Attacking Machine Learning with Adversarial Examples, Open AI – Blog Web Site. Retrieved Jan. 21, 2018, from https://blog.openai.com/adversarial-example-research/

Herzfeld, N. L. (2002). In Our Image: Artificial Intelligence and The Human Spirit. Fortress Press.

Hibbard, B. (2014). Ethical artificial intelligence. *arXiv preprint arXiv:1411.1373*.

Holland, O. (Ed.). (2003). Machine Consciousness. Imprint Academic.

Kose, U., & Vasant, P. (2017). Fading intelligence theory: A theory on keeping artificial intelligence safety for the future. In *Artificial Intelligence and Data Processing Symposium (IDAP), 2017 International* (pp. 1-5). IEEE.

Lin, P., Abney, K., & Bekey, G. A. (2011). *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press.

Lin, X., Beling, P. A., & Cogill, R. (2017). Multi-agent Inverse Reinforcement Learning for Two-person Zero-sum Games. *IEEE Transactions on Computational Intelligence and AI in Games*. PP(99).

McLaren, B. M. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE Intelligent Systems*, *21*(4), 29-37.

MIT Technology Review. (2014). Do We Need Asimov's Laws?. 2014, TechnologyReview.com. Retrieved Jan. 23, 2018, from https://www.technologyreview.com/s/527336/do-we-need-asimovs-laws/

Moor, J. (2009). Four kinds of ethical robots. *Philosophy Now*, *72*, 12-14.

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

Negnevitsky, M. (2005). Artificial Intelligence: A Guide to Intelligent Systems. Pearson Education.

Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *Icml* (pp. 663-670).

Nicolescu, B. (2017). Technological Singularity: The Dark Side. In *Transdisciplinary Higher Education* (pp. 155-161). Springer, Cham.

Orseau, L., & Armstrong, S. (2016). Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence* (pp. 557-566). AUAI Press.

Pavaloiu, A., & Kose, U. (2017). Ethical Artificial Intelligence-An Open Question. *Journal of Multidisciplinary Developments*, *2*(2), 15-27.

Perdue, R. T. (2017). Superintelligence and natural resources: morality and technology in a brave new world. *Society & Natural Resources*, *30*(8), 1026-1031.

Powers, T. M. (2011). Incremental machine ethics. *IEEE Robotics & Automation Magazine*, *18*(1), 51-58.

Riedl, M. O., & Harrison, B. (2016). Using Stories to Teach Human Values to Artificial Agents. In *AAAI Workshop: AI, Ethics, and Society*.

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, *36*(4), 105-114.

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., & Edwards, D. D. (2003). Artificial Intelligence: A Modern Approach (Vol. 2, No. 9). Upper Saddle River: Prentice hall.

Schneider, S. (2016). Science Fiction and Philosophy: From Time Travel to Superintelligence. John Wiley & Sons.

Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Torrance, S. (2008). Ethics and consciousness in artificial agents. *Ai & Society*, *22*(4), 495-521.

Vamplew, P., Dazeley, R., Foale, C., Firmin, S., & Mummery, J. (2017). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 1-14.

Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology*, *12*(3), 243-250.

Wallach, W., & Allen, C. (2008). Moral Machines: Teaching Robots Right From Wrong. Oxford University Press.

Yampolskiy, R. V. (2013). Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In *Philosophy and theory of artificial intelligence* (pp. 389-396). Springer, Berlin, Heidelberg.