The Moral Implications of Artificial Intelligence

Jaime Reborn

University of North Texas

2019

Abstract

This research paper analyzes the moral implications of artificial intelligence (AI). Existing AI technologies and future advances will uplift people's social welfare by making it easier to manage and treat chronic illnesses. They will also improve the efficiency with which human beings undertake complex tasks. As the thinking, reasoning, and problem solving capabilities of AI machines continues to advance and their integration into people's lives widens, the technologies will generate serious moral consequences. These consequences will include the harm that arises from the machine's inability to make ethical decisions, the extinction of the human species, the dystopian life characterized by suffering and totalitarianism, the paternalistic existence in which the state will manipulate citizens' everyday living decisions, and the social disintegration that will develop from the provision of personalized solutions.

Moral Implications of Artificial Intelligence

By improving the efficiency levels of existing technologies, current and future artificial intelligence (AI) applications will be instrumental in uplifting people's social welfare in diverse professions and sectors. The affected industries and fields will include environmental protection, personal health management, public welfare, precision medicine, education, public safety, service robotics, entertainment, employment, communication, transportation, and economic development. In medicine, AI tools with machine learning competencies will assist physicians and other medical practitioners in uncovering complex association in the human brain (Maruthappu, 2018). Using sophisticated networks of interconnected neurons that support cognitive development in the same manner as the human brain, machine learning capabilities of the AI applications will identify and solve complex medical problems by rapidly and carefully evaluating various sources of insight to reach reasoned conclusions (Maruthappu, 2018). In London, an AI-powered smartphone App with machine learning capabilities has triaged more than 1.2 million patients to the Emergency Room in a single year (Maruthappu, 2018). This success highlights the efficiency and social welfare improvements associated with AI.

As with medicine, other sectors like communication and transport will report the same level of progress efficiency and social welfare improvements after the installation of AI applications. In communication and transport, developers will integrate AI technologies that can learn about and understand the thinking patterns of the residents of Smart Cities, simulate their behaviors, and develop personalized services that improve their wellbeing and quality of life (Allam & Dhunny, 2019). With the collected and developed knowledge in Smart Cities, the technologies improve policymakers' ability to make informed decisions (Allam & Dhunny, 2019). In computational paralinguistics, AI applications with machine learning, unsupervised

learning, and human-machine collaborative learning capabilities have improved the accuracy of technologies that analyze and predict a speaker's personality traits based on textual cues and voice acoustics. The machine learning capabilities have strengthened existing technologies' ability to analyze, learn, see-through, and develop precise accounts of the speaker's emotional state, height, weight, personality traits, and other traits that human speaker analysts cannot identify (Schuller, Zhang, & Weninger, 2018). These AI innovations have generated substantial improvements in human beings' social welfare and, in the process, contributed to a massive upturn in the value of AI systems. Indeed, conservative estimates indicate that the value of social and economic welfare opportunities created by AI systems will increase to $11 billion by 2024 (Hidemichi & Shunsuke, 2017).

In spite of the projected progress in social welfare in the aftermath of AI installation, society's continued utilization of the applications generates serious moral perils. The more the advancement of AI improves, the more society will confront the moral implications arising from the technologies. Collectively, AI-based systems utilize machine learning programs that can learn, mimic, and simulate human behavior (Dreyer & Geis, 2017). While some experts and lay readers may regard the determination of the moral implications of AI as entirely subjective, none of them can deny the fact that technologies are only valuable when they act in the service of the human being and create the foundational knowledge needed for people to undertake their day-to-day tasks. When technology has neural nodes that give it learning capabilities that are superior to the human brain, serious moral consequences arise. With this in mind, this research paper will delineate the moral implications that spring from the widespread adoption of AI technologies. To obtain an understanding of these results, the researcher will analyze the learning capabilities of AI applications and review scholars' claims on their ethical consequences.

The overall objective of the study is to analyze the moral implications associated with AI applications. This topic will be of service to the reader in three ways. First, it will give the target audience the intuition it needs to understand the actual impact of AI. At present, scholars' focus on the efficiency and social welfare implications of AI applications has prevented readers from appreciating the adverse effects that may arise from increased utilization of AI applications. The results of the present study will give readers the knowledge they need to understand that the integration of AI into people's day-to-day lives will have serious moral implications. Second, the study will educate readers by highlighting alternative accounts that may rule out or offset many of the efficiency and social welfare benefits that will arise from AI application. By emphasizing the moral implications, readers will understand that AI integration will lead to positive and negative developments. In many cases, the negative developments will be so severe that it might prompt the readers to reevaluate their enthusiasm towards the technology. Finally, the study will benefit the readers by discussing AI application in a way that does not conflate issues and avoids unproven or misleading connotations.

Limitations of Study

Based on the delineated aim, this study will focus on two things—the learning capabilities of AI technologies and the moral repercussions associated with the incorporation of AI applications into various aspects of people's lifestyles. In the course of analyzing the learning capabilities of AI, the researcher will not delve into an investigation of the developed technologies. The researcher will merely assess the knowledge capabilities of recently developed and future AI applications. Because many of the AI advancements are recent, the author will rely solely on recently published scholarly journals, newspaper articles, and textbooks. In other words, the researcher will only use sources published between 2017 and 2019.

Methodology

To collect information on the moral implications of artificial intelligence, the author used the qualitative research design. Using this research design, the author identified the content analysis technique and used it as the method of data collection. After identifying the relevant peer-reviewed journals, textbooks, and online articles, the author developed codes that he used to identify relevant phrases and arguments.

Literature Review

In many of the reviewed studies on AI, authors have claimed that AI applications are autonomous systems that can learn and make thousands of cognitive, information-driven decisions in the absence of direct supervision or guidance from human actors. In other words, the machines have the capacity to evaluate the risks and benefits associated with a given scenario and make a reasoned judgment (Etzioni & Etzioni, 2017). Certainly, cognitive architecture studies confirm that developers have created AI applications that can capture, at the formulation level, human cognition mechanisms (Lieto, Bhatt, Oltramari, & Vernon, 2018). These applications can independently and efficiently perform underlying brain functions like perception, adaptivity, memory, learning, control, and reasoning (Lieto, et al., 2018). The long-term objective of these cognitive architecture developments is to develop AI-based robots and computing systems that can achieve human intelligence levels.

Over the last decade, advancements in cognitive architecture have ushered new technologies like iCub, CLARION, ACT-R, and SOAR (Lieto, et al., 2018). These superhuman applications can perform a wide range of specific cognitive tasks, including recognition, action execution, perception, learning, reasoning, and selective attention (Lieto, et al., 2018). With the aid of neural networks, the super intelligent technologies review large volumes of conflicting

information, draw on additional insight from the surrounding environment, evaluate the merits of various causes of action, and make accurate inferences and conclusions (Etzioni & Etzioni, 2017). For instance, many of the advanced GPS systems can instantaneously identify accidents or upcoming traffic and advice drivers to reroute (Etzioni & Etzioni, 2017). Although none of the technologies can simultaneously undertake several human level cognitive tasks, developers believe that they will realize this objective in the coming years with the development of AI technologies with General Artificial Intelligence skills. These cognitive architecture advancements and the predicted development of AI applications with superhuman intelligence competencies have serious moral implications. The next section of the research paper will analyze these moral implications.

Analysis

How Intelligent Are Artificial Intelligence Technologies?

Researchers have undertaken many studies in their quest to enhance readers' appreciation of the concept of AI, AI trends, and the moral consequences of AI. In relation to the concept of AI, all of the reviewed studies acknowledge that AI is a technology that mimics, recreates, and even surpasses human cognitive development. Concerning AI trends, the reviewed studies confirm that technology firms and researchers are still developing AI applications. At present, all of the AI applications can excel at specific tasks like thinking, handwriting analysis, speech processing, and video content analysis. They cannot perform the general thinking functions of the human brain. However, future AI machines will simultaneously perform general brain tasks. Regarding the moral implications, scholars are still making speculative arguments because technology firms are yet to realize their AI development objectives. Nonetheless, there is general agreement that continued creation and adoption of machines with super intelligence will expose

human beings to serious moral consequences. Due to the speculative character of assertions on the moral implications of AI, scholars have not undertaken a comprehensive analysis of the topic. This paper will bridge that gap by offering a concrete evaluation of the moral effects of AI adoption.

Moral Implications of AI

AI technologies' inability to make ethical decisions raises serious moral concerns. Driverless cars and other AI technologies have machine learning features that give them the power to make autonomous, unsupervised decisions, but they do not have the power to ensure that their decisions are ethical or moral. The developers' failure to equip the technologies with ethical control functions can cause considerable harm to users and innocent bystanders (Etzioni & Etzioni, 2017). In May 2016, Americans witnessed the risks associated with AI-based autonomous technologies and systems when a Tesla vehicle on autopilot mode crashed and killed its passenger (Etzioni & Etzioni, 2017). In February 2019, another Tesla car on autopilot caused a crash on a highway in New Brunswick, New Jersey (Silverstein, 2019). The driver informed the police officer that the vehicle's sensors caused the accident after mistaking diagonal white lines for a different lane (Silverstein, 2019). These incidents highlight the risks associated with developers' failure to incorporate systems that render the decisions of AI technologies ethical (Silverstein, 2019). Many other autonomous machines like childcare robots, robotic surgeons, and unmanned aerial vehicles can collect data, analyze it, form conclusions, and alter their behavior without guidance or oversight from human actors (Silverstein, 2019). Such technologies can make decisions that can visit deleterious consequences on people's physical and emotional wellbeing, but they lack the capacity to tell the difference between morally appropriate and morally reprehensible decisions. If these machines can make learn and

make instantaneous data-driven decisions about when to yield, when to stop, when to slow down, developers must also equip them with ethical decision-making competencies. Failure to incorporate ethical decision-making features into their machine learning capabilities raises serious moral questions regarding the risks that these technologies pose to humans. Inclusion of ethical consideration will improve the technologies' ability to differentiate between decisions that are morally sound and those that violate the rules of morality. Such a capability will mitigate the perils that human beings face as a result of the continued use of AI by transforming the applications into moral thinkers.

In addition to the absence of ethical controls, the uninhibited adoption of AI innovations will pose an existential threat to human beings. The unfettered creation of super-intelligent AI-based technologies will eventually lead to the development of robots and human bionic systems that will replace humans in all aspects of life. As the AI-based technologies become more and more intelligent and society's big data problems become more and more challenging, human beings will be left with no option but to delegate all important decisions to the machines (Spyros, 2017). In the long-term, this state of affairs will render humans totally dependent on machines. In this utopian existence, human beings will be relegated to a second class status (Spyros, 2017). Their stature will be equivalent to that of computer pets (Spyros, 2017). In this subhuman status, robots and machines can use and misuse humans as they please (Spyros, 2017). With the relegation of humans to the periphery of existence, companies, governments, leading transnational corporations, and the ruling elite will begin to implement measures that will lead to their systematic elimination (Spyros, 2017). They will eliminate the communities that they regard as surplus and leave a small population of humans that can enjoy the abundance brought about by the new crop of superhuman technologies (Spyros, 2017). This catastrophic outcome

highlights the correlation between the AI advancements and the transformation of human beings into an endangered species.

Closely related to human beings' transformation into an endangered species are the possible dystopian outcomes associated with AI. When the AI inventions spiral out of control, they will usher humans into a totalitarian existence characterized by great suffering and injustice. Breathtaking advances in the field of AI have led to the development of technologies that can think, learn, and continuously develop their neural nodes. Recently, Google Inc. underscored this fact in its announcement that its *DeepMind* algorithm had to itself the modalities of winning 49 versions of a complex strategy game. The news came at a time when studies showed that AI technologies can now read and understand handwritten language and outwit human beings in complex tasks like describing video and photo contents, evaluating financial transactions for evidence of fraud, and reviewing citizens' tax returns for evidence of tax fraud. Within the next 10 to 20 years, these radical advancements will pose a threat to half of today's jobs and plunge the world into an unprecedented unemployment crisis. By 2060, supercomputers will outdo human beings in all areas and contribute far-reaching changes in the ways in which countries have organized their economies and governance structures. China, Britain, and the US have already developed super-intelligent AI technologies that have an influence on economic and social policies. Britain and China's AI technologies can do an internet search, collect information about their respective citizens' online activities, and give them a "Citizen Score" that will have an influence on the types of personalized services they will receive. Further advancements AI will lead to the emergence of persuasive computing applications that will give governments the power to steer citizens towards desired outcomes like voting for a particular presidential candidate or supporting controversial legislation. The government's increased ability to shove

citizens towards certain courses will signal the death of democracy, the emergence spread of totalitarianism and human suffering. These developments suggest that AI will push the world towards a morally reprehensible programmed society in which governments will program their citizens and have complete control over their mobility, decisions, and social interactions.

Furthermore, the growth of paternalism is among the moral consequences of the increased integration of AI innovations into people's day-to-day lives. Using big data retrieved from social networking sites, healthcare records, and other sources of citizens' personal information, governments can now use AI machines to nudge citizens towards behaviors that it regards as friendly, healthy, or sustainable. The concept of "big nudge" is now part of the routine parlance in policymaking circles. The concept, a manifestation of the phenomenon of paternalism, refers to the combination of big data, AI, and nudging. Under the big nudge systems, governments will extend their reach beyond the knowledge of what citizens are doing to the control of what citizens are doing. Governments will identify the "appropriate" behaviors and use sophisticated mechanisms to persuade citizens to adopt them. To the lay citizens, this behavior will be evidence that legislators, policymakers, and federal government officials care about their physical and psychological wellbeing. However, experts in governance will interpret this as a digital scepter in which members of the executive and legislative arms of government govern the citizens effectively and efficiently complying with rules requiring citizens' participation in the governance process. In the long-term, the paternalistic conduct could pave the way to an optimized capitalist world in which citizens serve the needs of the government and the ruling elite. In this world, the masses will be under the control of a data-empowered benevolent king who will use a digital magic wand to influence them to meet predetermined social and economic targets.

A further moral consequence will be the internal erosion of AI machines due to the dearth of transparency and democratic oversight over the functionality of AI machines. Even with the highlighted advancements and the moral implications, governments have failed to establish governance institutions and implement oversight policies that would prevent the internal erosion of the functionality of AI machines. Recent hacking incidents and corporate scandals have shown that AI technologies are not immune from dubious influences. In the past decade, hackers have managed to hack all institutions and corporations with complex IT infrastructures, including the NSA, the Pentagon, and the White House. Extremists, terrorists, criminals, and other categories of activists will attempt and succeed in taking control of the digital magic wand that gives governments and technology firms control over consumers and citizens. Similarly, government contractors and other unscrupulous individuals can access and influence the activities of the AI machine without the knowledge of the government or the technology firms that are managing them. After accessing these technologies, hackers or insiders will reprogram the technologies in a way that leads them to influence citizens' behavior to generate outcomes that are favorable to their interests. The hackers may infiltrate the system and program it to influence citizens to engage in riots by undertaking arbitrary alterations in the cost of electricity, fuel, taxes, and other essential services. During the ensuing riots, the hackers might seek to destabilize the government further by asking citizens to call for the president's resignation. The recent revelations that Russia hacked into the DNC servers and used Facebook and Twitter to influence Americans to vote for Donald Trump provide a snapshot of what might happen when hackers take advantage of the absence of democratic control and transparency over AI technologies to infiltrate the AI machines and erode their functionality from the inside. External and internal actors' ability to

access and gain control over citizens' emotions, behaviors, and choices is a moral implication that the society will have to grapple with after the incorporation of AI into people's lives.

Apart from the risk of internal erosion due to the absence of democratic oversight mechanisms and many of the delineated moral implications, human-AI interaction agency may lead to a scenario in which AI machines and their human creators are acting rationally but in ways that are contrary to each other's interests. The development of AIs that can think, reason, and engage in rational behavior may foster an environment in which AI machines are acting in ways that are not aligned to the interests of human beings. This dangerous turn may manifest in three ways. First, it may manifest when the AI machines engage in paperclip maximization behaviors in which they exhibit values that are not aligned to the values of their human creators. This may arise when the AI machines turn into serial killers by deciding that killing children suffering from chronic infections is the right thing to do. Second, the contradiction may develop when the AI applications misinterpret human values. This may develop when a command urging them to maximize citizens' happiness leads to the opening of all borders. Third, the conflict will develop when AI machines decide to align their interests with those of an unscrupulous individual or corporation. In this regard, the corporation or individual will weaponize the technology and collaborate with it in efforts targeted at destroying people's physical and psychological wellbeing. This human-AI agency interaction conflict is an issue that AI developers have continued to ignore. Governments incorporating the technology into their systems have also expressed indifference towards the possibility that rational AI technologies may decide to behave in ways that are inconsistent with the interests of their creators. Nonetheless, the possibility that the technologies might opt to act in ways that are contrary to the wishes of their creators and cause catastrophic damage to the public is something that the public

that supports the technology cannot afford to ignore. The masses must consider whether it is morally right for such technologies to think independently and make decisions that are inconsistent with the thinking of their creators.

Along with the human-AI interaction agency conflict, the manipulation of services to degrade the levels of social cohesion in society is a moral consequence the public cannot afford to ignore. Governments and other organizations will use the service customization features in AI machines to polarize citizens and prevent them from developing shared experiences and engaging in collective action. To conceal the manipulation of the masses through AI machines, governments will customize suggestions and solutions to each citizen's unique needs. In this way, state operatives will create the echo chamber or filter bubble effect by buttressing local trends through citizens' repeated exposure to a highly personalized solution. As a result, citizens will get their individual opinions and views relayed back to them whenever they complain about the ineffectiveness of government service. Social polarization will be the long-term consequence of this state of affairs. As more and more citizens become dependent on personalized solutions, they will begin to develop separate social groups that have no shared experiences or common interests. In the absence of unifying values, aspirations, and experiences, the conflict will be the common feature in social groups. People will find themselves increasingly find themselves at odds with the membership of their respective social groups (Helbing, et al., 2018). In this way, the personalization of solutions will inadvertently destroy the levels of social cohesion in the society. Evidence from the current levels of social polarization in the US can provide an insight into the role that the increased use of AI will use in undermining citizens' unity (Helbing, et al., 2018). The use of AI in social media and search engine applications has led to the generation of information that appeals to each citizens' political interests (Helbing, et al., 2018). When dealing

with online users who self-identify as Democrats, the search engines and social media networks generate search results with political and social information from progressive media sources like CNN, MSNBC, and NBC. When the internet user identifies as Republican, the AI-based information search applications will search results that appeal to his political philosophy. These personalized results have entrenched social divisions and widened political rifts within families and other social circles. As levels of social polarization widen, political compromises are becoming impossible to achieve. In the long run, the divisions will fragment the American society and occasion its disintegration. This situation foretells the negative consequences that will spring from increased dependence on AI technologies.

By the same token, the continued existence of superhuman AI applications on the internet generates serious moral implications surrounding their possible impacts on human beings. The development of AI with superhuman capabilities, when combined with the online world, will render AI machines vulnerable to narrow AI attacks that will expose human beings to catastrophic consequences. The narrow AI viruses will have the ability to use human language to undertake phishing attacks. When successful, the viruses will trick the AI-based robots and influence them to adopt certain behaviors (Turchin & Denkenberger, 2018). For instance, the AI virus may influence the auto-pilot controlled planes to crash into nuclear power plants. Considering that thousands of commercial airlines with computerized autopilots are in the air at any given moment, the possibility of a coordinated plane attack in the aftermath of a narrow AI infiltration increases tremendously (Turchin & Denkenberger, 2018). Admittedly, Elon Musk has acknowledged that that AI's presence on the internet poses a serious risk to humans (Clifford, 2018). According to Musk, viruses could manipulate these AI technologies into publishing fake news, blackmailing consumers, and fostering suicidal thoughts in citizens (Clifford, 2018).

Russia's *Blue Whale* game's successful suicide ideation offers an insight into the mechanisms that a narrow AI virus would use to manipulate human behaviors and attitudes. *Blue Whale* used intuitive and engaging tasks to convince 130 teenagers to commit suicide (Adeane, 2019). The game would send regular tasks to subscribers. In the initial stages, the game sends users fairly innocuous tasks like "Watch a horror film" or "Wake up at 3 am" (Clifford, 2018). However, the tasks become complex as users progress to higher levels. The users receive tasks telling them to "Stand on the edge of a cliff" or "Etch the image of a whale on your right arm" (Clifford, 2018). The final task in the game was a request to commit suicide (Clifford, 2018). At least 130 teens adhered to the request and died. Equally, narrow AI viruses will infiltrate supercomputers, corrupt their knowledge, and force them to perform tasks that will harm human beings (Clifford, 2018). This negative effect is something the human population will have to contend with if they continue to prioritize convenience over their physical and emotional wellbeing by maintaining their reliance on AI machines.

Conclusion

The more the advancement of artificial intelligence improves, the more it will expose human beings to adverse moral implications. AI technologies' inability to make ethical decisions and the harmful effects that may arise from such a flaw is among the moral implications that may arise from human beings' sustained embrace of the superhuman machines. The existential threat that the super intelligent machines will pose to humans is another moral consequence of human beings' must grapple with. As society maintains its grip on AI applications and integrates them into people's day-to-day technologies, they will reap important benefits like convenience and improved quality of life. However, these benefits may come at the expense of the human species.

The dystopian consequences associated with the technology is also a moral repercussion that will arise from human beings' unrelenting dalliance with AI technologies.

References

Adeane, A. (2019). *Blue whale: What is the truth behind an online 'suicide challenge'?* Retrieved from BBC: https://www.bbc.com/news/blogs-trending-46505722

Allam, Z., & Dhunny, Z. (2019). On big data, artificial intelligence and smart cities. *Cities, 89*, 80-91.

Clifford, C. (2018). *Elon Musk: "Mark my words- A.I. is far more dangerous than nukes.* Retrieved from CNBC: https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-dangerous-than-nuclear-weapons.html

Dreyer, K., & Geis, R. (2017). When machines think: Radiology's next fontier. *Radiology, 285*(3).

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *Journal of Ethics*, 1-16.

Helbing, D., Frey, B., Gigerenzer, G., Hagner, M., Hofstetter, Y., Hoven, J., . . . Zwitter, A. (2018). Will democracy survive big data and artificial intelligence. *Towards Digital Enlightenment*, 73-89.

Hidemichi, F., & Shunsuke, M. (2017). Trends and priority shifts in artificial intelligence technology invention: A global patent analysis. *RIETI Discussion Paper Series*, 1-38.

Lieto, A., Bhatt, M., Oltramari, A., & Vernon, D. (2018). The role of cognitive architectures in general artificial intelligence. *Cognitive Systems Research, 48*, 1-3.

Maruthappu, M. (2018). Artificial intelligence in medicine: Current trends and future possibilities. *British Journal of General Practice*, 143-144.

Schuller, B., Zhang, Y., & Weninger, F. (2018). Three recent trends in paralinguistics on the way to omnicient machine intelligence. *Journal on Multimodal User Interfaces, 12*(4), 273-283.

Silverstein, J. (2019). *Driver says Tesla car gets "confused" and crashes on highway.* Retrieved from CBS News: https://www.cbsnews.com/news/tesla-autopilot-car-gets-confused-and-crashes-on-highway/

Spyros, M. (2017). The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms. *Hephaestus Repositoryy*, 1-27.

Turchin, A., & Denkenberger, D. (2018). Classification of global catastrophic risks connected with artificial intelligence. *AI & Society*, 1-17.