Tyler Toussaint
Prof. Ignani & Prof McLaughlin
Honors Capstone
4/18/2019

AI and Ethics

Technology is a factor of society which extends our abilities, solves world problems or eases some of our daily routines. Naturally, as new problems arise, we seek after new technology to fix them, and because of societies dependency on technology, it is rapidly progressing at speeds that most people are unaware of. With the growth of technology came the creation of Artificial Intelligence, a system that could think similar to that of a human being. Initial ideas of artificial intelligence lead to the thought of robots or human like inventions that would resolve societies problems. Technology of this caliber does benefit human life, but it also brings fear and consequence into question. Movies such as Irobot fantasize a future where an Artificial Intelligence system takes over and destroys human civilization. Inventing a product that can eventually outthink humans is a realistic concern, however, there is a more pressing risk, and that is if the future of an Artificial Intelligence will adhere to an ethical code. Without following ethics, the pursuit of advancing AI would be morally wrong and could bring harsh consequences.

Understanding what AI is and where it is applied is the key first step in forecasting the future progression of the technology. Contrary to movie depictions, artificial intelligence is not as simple as the making of human acting robots, in fact it is embedded in various widely used applications that can often go unnoticed. Artificial intelligence is broken up into three basic concepts which are machine learning, deep learning, and neutral networks. Knowledge of these allows one to understand where AI is implemented. Machine learning is a branch of AI that, without pre-existing code, gives the machine the ability to learn using large amounts of trial

examples for a task. By examining these trials, the machine is able to adapt and choose a strategy to complete the task given. The machine, after endless permutations of data, acquires the ability to recognize patterns and features. Areas useful for machine learning are prevalent in image recognition, speech recognition, medical diagnosis, classification and even prediction systems (Sciglar 2018). An article from Sigmoidal explains the use of machine learning AI in the healthcare in which it was used to treat war veterans with PTSD and detect brain bleeds. Tiatros Post Traumatic Growth for Veterans program partnered with IBM Watson, an AI that used analytics to ensure that the patients would complete psychotherapy and increased the completion rate from under ten percent to seventy-three percent. IBM Watson was again paired with an Israel health group using clinical insight patient data and machine vision to automatically flag patients with brain bleeds. Both examples show that through given data, these artificial intelligence systems can learn patterns and behaviors and access situations more effectively than humans (Sigmodial 2017).

The next concept in AI, deep learning, is similar to machine learning however it deals with the machine learning more than one task. For the AI to have the ability to perform more multiple tasks, computer scientists formulate general-purpose learning algorithms that help machines learn more than just one task. Now, the AI can analyze something such as a picture and use the information in another type of data set (Sciglar 2018). Implemented in Google's AlphaGo project, an AI played against human opponents in a strategic game called Go. The basic premise of the game is for one player to surround more territory than the opponent. Seemingly impossible due to the games complexity and focus on practice and human intuition, the AI was able to beat the three-time European champion and would later go on a sixty-game win streak. Initially the artificial intelligence learned how to play the game using thousands of

human amateur and professional games. Later advancement allowed it to play itself starting from completely random play allowing it to arguably become the strongest player. Through analyzing previous games and moves the AI was able to produce highly inventive moves shocking to players (AlphaGo 2016).

The last concept of AI are neutral networks which makes the concept of deep learning possible. Inspired by human biology, neutral networks act as neurons or brain cells through the use of math and computer science principles. By taking in information to a certain threshold, passing it to the next layer, and then comparing that with the other information processed the same way, AI gains the ability to learn (Sciglar 2018). Having the knowledge of the use of artificial intelligence, one may now question how a machine that is beneficial and common in daily routine could pose a threat if not ethically bound.

First, AI systems are known for their ability to make tasks automated and by doing so decreases many of the jobs in the markets they are implemented in. The concern that over automation can lead to the depletion of jobs and high unemployment is a potential ethical risk as companies invent more new ways to automate jobs (Bossmann 2016). In "AI and Jobs: the role of demand", James Bessen states that there is widespread concern today that artificial intelligence technologies will create mass unemployment during the next 10 or 20 years. Although the decline in the automotive industry gives more reason for concern, Bessen focuses on the role that demand plays and believes it is key to understanding whether major new technologies will decrease or increase employment in affected industries. He believes that if the demand for AI technology remains elastic and does not completely replace humans, then it could create jobs rather than deplete them. As of now, most AI is focused on automating sub tasks of jobs rather than complete automation therefore leaving room for human workers (Bessen &

James 2018). Bessen's argument eases the fear of high unemployment under the condition that the AI implemented does not fully replace the human workers. An ethical boundary is drawn here, and if progression allows for AI to become self-sufficient, then jobs will be depleted without renewal. Two alternative scenarios, a growth perspective and a crisis perspective, were created due to the rising concern of future AI and unemployment. Both scenarios are calculated predications based on how the technology could progress. The growth perspective shows that automation reduces cost and frees up labor which allows for new growth and jobs in unexpected areas. The main argument is that although artificial intelligence can automate knowledge, it lacks higher order functionalities that are unique to humans. Unlike the growth perspective, the crisis perspective believes that the latest automation will be so close to human abilities that jobs creation will not happen. Again, the same ethical boundary is present in that if given higher order functions, AI can deplete the job market that it is used in. Glimpses of AI removing jobs can be seen at a factory in China named Foxconn. Here the company automated ninety percent of their jobs and are looking to automate two thirds of the remaining. The company is an example that AI has the potential to fully run businesses without the addition of new jobs. As each day goes on artificial intelligence becomes more capable, to the point where jobs which required human versatility are now being completed. IBM Watson is able to perform research, a skill unique to humans, that employs twenty percent of lawyers' billable hours. Another concerning aspect is that an AI, designed to imitate all human functions, is not needed to devastate the job market rather an AI, designed to complete specific tasks, could do the same under the right criteria. If jobs do get created the crisis perspective believes that due to machine learning, AI will be able to learn the jobs faster and eventually take over (Halal 2016). Current AI suggests that automation is beneficial and creates growth however without the ethical question of unemployment kept in

mind, future AI could surpass an intellectual level where jobs become completely automated and the crisis perspective becomes a real crisis.

A more subtle ethical issue pertaining to artificial intelligence deals with the aspect of inequality (Bossmann 2016). Economic inequality has been a concern as more companies explore the uses of artificial intelligence. AI has proven its ability to replace human workers, and as the number of workers decrease, there is more money that goes to fewer people. A gap is then created as individuals who own AI driven companies make all the money. Steven Rattner did a comparison between the revenue and amount of workers of the top three automakers in 1990 and the top three companies in Silicon Valley in 2014. What he found was that both companies made 250 billion dollars in revenue, however, the top three automakers employed 1.2 million workers where Silicon Valley only employed 137 thousand (Parke 2017). If not dealt with caution, company owners may be able to exploit future AI to make large amounts of money because automation would reduce the need for high a volume of workers. Future AI has the ability to cause inequality in a sense that as AI becomes more involved in society, people will be forced to become more technical. Most jobs will deal with maintaining or fixing AI systems and those who cannot gain the technological education needed to qualify for these jobs will have a harder time finding other work. Basically, those who cannot reach a high level of technological intelligence will be forced into lower level jobs with often times less pay. If artificial intelligence progresses without ethics, a gap in society will be created between those who are technologically inclined and those who are not.

Next, artificial intelligence has influence on human behavior and interaction (Bossmann 2016). In 2014, a computer program posed as a 13-year-old boy named Eugene. The computer program was able to convince 33 percent of the people that it chatted with that it was actually

human. Eugene passed what was called the Turing test which is a 5-minute conversation between a human judge and either another human or a computer program. If the judge can't tell within 5 minutes whether they are talking to a computer or not, then the test suggest that it is capable of thinking because it consistently able to make us believe it is doing so (Ackerman 2014). Without having full human functionality, a present AI can fool scholars that it too is a human. Not only can AI have human-like conversation, the technology can also be used to figure out one's personality. Similar to market strategies, artificial intelligence is able to track recent websites, purchases, common searches and other data to filter ads and certain popups that the users see. These ads and popups tend be of interest to the user based off the data the AI has studied. Sentient is a company that uses AI to optimize the conversion of users browsing to buying from a website. In a case study optimizing the web interface that connects users to online education program, the conversion rate was found to be 5.61 percent. Sentient's program, Ascend generates, tests, and evaluates designs then dismisses inefficient designs. The best design in the time allotted is output as part of the learning process and tested through live online testing. Through Ascend, the conversation rate of the web interface rose from 5.61 percent to 8.22 percent proving that AI is effective at influencing people towards certain transactions (Miikkulainen 2018). With current AI technology computer programs are able to make people believe that they are human and influence us into interacting more with certain web interfaces. As new AI technology is discovered these abilities only get stronger. If full human function is reached in AI, AI could possibly manipulate people emotionally or spirituality while the user thinks the program is human. Ethics is needed to restrict future AI from being too invasive reviewing people's data and using it against them.

When dealing with current artificial intelligence, one must remember that it is a computer and can be fooled in ways that humans cannot (Bossman 2016). In 2015, a study was released on how a deep neural could recognize images with high certainty that humans could not perceive. The human eye sees that pictures as static however the AI would say with 99 percent certainty that it was a bus or an animal. Due to the algorithm behind the program, the AI picks up patterns that is imperceptible to the human eye (Nguyen 2015). The danger here is that if the algorithm picks up these unrecognizable pictures they can easily be mislabeled as something else. In multiple cases a false reading from an AI could be detrimental such as in the health field, if a person is misdiagnosed then the treatment would be unnecessary. Social factors can be threatened due to AI misinterpretation. For example an individual who identifies as a woman gets classified as a man because an AI analyzing their features found similarities closer to that of a male. The result could offend the individual because their correct identity is not being used causing the AI to be unethical.

Not only can an AI be fooled, but it has been proven that some are bias (Bossmann 2016). In 1996, Batya Friedman and the Helen Nissenbaum reported AI bias when observing SABRE flight book scheduling system. The system provided flight listings and routing information for airline flights in the United States however the algorithm and information sorting created a systemic bias towards its sponsor. American Airlines was always shown to agents on the first page even if there were other options that were cheaper or a more direct flight. The National Resident Match program further proved AI bias as its algorithm rules favored the hospitals preferences over the resident preferences. Examples of bias in AI systems referenced previously may seem trivial and the effects minimal, however, certain AI programs have shown to be biased on sociological reasons such as race. Northpointe's Correctional Offender

Management Profiling for Alternative Sanctions is a software used for sentencing and parole but also represented the probability they would return to jail. Black convicts were rated higher than non- black convicts in returning to prison and were twice as likely as Whites defendants to be misrepresented as higher risk of violent recidivism (Osoba 2017). Racial bias brings a huge threat to human ethics because all people are to be seen as equal, so to have technology that includes bias would be ethically wrong to those the program segregates. Artificial intelligence, being a program, makes it unaware of the value of human dignity. Due to its ignorance of human morality, AI can easily make decisions to exploit what it deems to be lesser without regard to who the machine has affected. If the future progression of artificial intelligence doesn't hold its algorithms accountable for bias basic rights to equality would be infringed on.

Safety and security of artificial intelligence systems must be managed properly to ensure that the data received is secure and the task it was given is completed correctly (Bossman 2016). A challenge that computer scientists see today is the advancement of cybersecurity because every security system eventually fails. That being said an AI system is bound to fail and depending on the task could cause serious harm. In 1983, nuclear attack early warning system falsely claimed that an attack was happening, and in 2010 an AI stock trading software caused a trillion-dollar crash in seconds. Not only did an AI kill an economy, but also killed a man when an AI robot designed to grab auto parts grabbed a person instead (Yampolskiy 2017). Faults or bugs in the code of AI algorithms are what causes the system to give incorrect outputs and with maintenance and cybersecurity developers try to minimize the errors. Given that no AI is completely secure, code must be ethically developed and maintained with the primary goal of safety and security of users. Straying from morals when creating algorithms diminishes the benefits because the failure of such AI systems proves to have detrimental consequences. As

technology improves the problems faced becomes more complex and without a guideline our control of the problems could breach our level of intelligence.

As scientists focus on studying technology to further the performance and efficiency, some begin to see the issue of controlling such advanced systems (Bossman 2016). Blinded by what could be created and drive to solve problems, the control problem has been able to evade large discussion. Nevertheless, control over AI is a pressing concern that needs to be established prior to future progression of the technology. Nick Bostrom states that "any sufficiently intelligent artificial mind could be capable of having devastating effects on the world, so approaches to controlling such a creation should be carefully analyzed beforehand". His reasoning for the statement was that "if machine brains surpassed human brains in general intelligence, then this new superintelligence could become extremely powerful—possibly beyond our control. As the fate of the gorillas now depends more on humans than on the species itself, so would the fate of humankind depend on the actions of the machine superintelligence" (Chong 2017). By allowing AI technology to gain functions equal or greater to that of humans, humanity relinquishes superiority over the world because AI systems would more efficiently and effectively run society. As more problems arise, AI complexity is upgraded to fix them, and the control humanity has over them continually decreases.

Lastly, the design of artificial intelligence is to imitate the functions of humans; so, when achieved where does the program fit into the rights of humans? Initially the question seems comical, however, if a system becomes self-aware with human functions it would be aware of its right (Bossmann 2016). By denying rights to the AI could be considered a form of slavery because of its higher functions comparable to humans. An article published by Harvard claims that the brain is a complex algorithm, however questions whether consciousness is what

separates us from an AI. If the mind only consist of algorithms and consciousness is the effect of that, then there is little choice but to give the same moral status to machines (Risse 2018). How could humanity give unequal value to that of an AI that could reciprocate the same or greater human faculties? The solution is simple, and that is to stop the advancement of artificial intelligence before it reaches the level of rationality that humans have. If scientists are creating systems to benefit society, then ethics should halt the advancement of AI past human's intellect unless humanity is ready to relinquish its control over them.

Technology proves its benefits throughout history as people continuously invent and advance it. Artificial intelligence being a creation in the progression of technology serves as a tool for automation and machine learning which is useful for various applications. Although there are many positives of AI systems, the future of the technology foreshadows fears and consequences that could arise. By pointing out the relation artificial intelligence has to ethics, moral boundaries are recognized. These boundaries serve to limit the problems that future AI may cause and without ethics pursuing the advancement of AI would be morally wrong. The choice is up to humanity to adhere to ethics but breaking of these restrictions makes movies like Irobot seem even more possible.

Works Cited

Ackerman, Evans. *A Better Test than Turing [News] - IEEE Journals & Magazine*, ieeexplore.ieee.org/abstract/document/6905475.

Miikkulainen, www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17332/16374.

"AlphaGo." DeepMind, 2017, deepmind.com/research/alphago/.

"Artificial Intelligence and Machine Learning for Healthcare." *Sigmoidal*, 16 Mar. 2018, sigmoidal.io/artificial-intelligence-and-machine-learning-for-healthcare/.

Forecasts of AI and Future Jobs in 2030: Muddling Through Likely, with Two Alternative Scenarios, William Halal , Dec. 2016, jfsdigital.org/wp-content/uploads/2017/01/JFS212Final（已拖移）-6.pdf.

Bessen, and James. "AI and Jobs: The Role of Demand." *NBER*, University of Chicago Press, 10 Jan. 2018, www.nber.org/chapters/c14029.

Bossmann, Julia. "Top 9 Ethical Issues in Artificial Intelligence." *World Economic Forum*, www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/.

Chong. *The Control Problem*. 2017, www.ieeecss.org/sites/ieeecss.org/files/documents/Presidents_Message_April_2017.pdf.

Osoba, Osonde. "An Intelligence in Our Image." Eds, 2017, eds-b-ebscohost-com.sacredheart.idm.oclc.org/eds/detail/detail?vid=4&sid=802b519f-9556-4366-aab2-a30f000cd155@sessionmgr120&bdata=JnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ.

Parke, Matt. “Silicon Valley's Top Dogs Make Billions with Fewer Workers. Why This Is Bad News for the Auto Industry.” *WorkingNation*, 23 June 2017, workingnation.com/silicon-valleys-top-dogs-make-billions-fewer-workers-bad-news-auto-industry/.

Risse, Mathias. *Human Rights and Artificial Intelligence*. May 2018, carrcenter.hks.harvard.edu/files/cchr/files/humanrightsai_designed.pdf.

Sciglar, Paul. “What Is Artificial Intelligence? Understanding 3 Basic AI Concepts.” *Robotics Business Review*, Robotics Business Review, 19 Apr. 2018, www.roboticsbusinessreview.com/ai/3-basic-ai-concepts-explain-artificial-intelligence/.

Nyugen, A. “University of Wyoming.” 2015 *Evolving AI Lab*, www.evolvingai.org/fooling.

Yampolskiy. *Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images*. 2017, arxiv.org/ftp/arxiv/papers/1610/1610.07997.pdf.