Implementation of Morality in Artificial Intelligence

Amanuel Matias

Abstract:

The spark of artificial intelligence is often attributed to Alan Turing, who was one of the first people to explore this, then foreign, concept of AI. However, his studies were restricted by the unavailability of higher powered computing technology. With the continuous development of powerful technology in recent years, significant advancements have been made in the field of artificial intelligence. These advancements have reached the point where people's lives are put at the hands of AI in certain situations. This brings in many complications regarding the morality of artificial intelligence technology. One route that scientists are taking involves using a database of humans' ethical decisions to provide a foundation for AI decision making. My objective is to increase the depth of this database by traveling up the coast of California and collecting data on different people's responses to various ethical dilemmas. This interview process will occur between the dates of May 15, 2019 and June 3, 2019. After I gather this information, I plan to look for trends in responses. I will use these trends to write an article about common perspectives in ethics and how they can be applied to artificial intelligence in society.

Introduction:

As the relevance of artificial intelligence in human society increases daily, there are many specific matters that need to be addressed before the successful integration of AI can occur. The issue that I am focusing on centers around the implementation of morality into technology possessing artificial intelligence. Without a set definition of morality, scientists and engineers arrive at the following question: How can machine ethics be programmed into artificial intelligence technology that plays a significant role in our society? This technology includes but is not limited to, autonomous vehicles, facial recognition, and aerial drones. As time progresses, an increase in the prevalence of artificial intelligence is certain, so it is important to address issues like this one before major advances are made regarding AI involvement in everyday life.

Background:

During the mid-twentieth century, the notion of artificial intelligence cemented itself into the minds of scientists and engineers. Alan Turing, an English mathematician, was one of the first to dive into the arithmetic behind artificial intelligence. One of his base questions was why are machines not able to make decisions and solve issues using attained information, if humans are able to do so? Turing saw this as a viable task, and in 1950 he wrote a paper about his testing of machine intelligence (Anyoha). However, Turing's hypotheses were short-lived. Due to the extremely high costs and lack of capability of computers in the 1950s, Turing was not able to carry out all the experiments that he had hoped to (Stone et. al).

Shortly after Turing's endeavors, three scientists published a program, *Logic Theorist*, which had the intention to imitate the problem-solving skills of a human being. This program is seen as the first to genuinely practice artificial intelligence. Between the introduction of *Logic*

Theorist in 1957 and the year 1974, AI prospered (Anyoha). A large part of this was due to the fact that computers made significant advancements in availability, efficiency, and performance. The repeated success found by scientists during this era persuaded government agencies to agree to fund institutional AI research.

Nowadays, the capability of computers continues to increase at a steady pace. For this reason, scientists and engineers are not limited in the same sense that Alan Turing was, back in the 1950s. In fact, computers now process vast amounts of information that people could by no means do themselves. The increase of reliance on computers has spread to areas where artificial intelligence is now more common than not. For example, a large number of phone calls one makes to companies will initially be received by an automated telephone operator. In the near future, many of the cars seen on a given road will not be operated by a human, but by a computer software.

With AI playing a more significant role in society, sometimes there is uncertainty when it comes down to the ethical values this technology should hold, and how programmers will go about implementing these values. According to the report, *Engineering Moral Agents- from Human Morality to Artificial* Morality, mathematical approaches can lead to the formation of moral theories which can then be relayed to machines of artificial intelligence (Fisher). Before doing so, a large collection of moral-dilemma situations is put together, and this is to be used as a base for the technology to look back onto for decision-making reference (Katte). This approach is similar to that which I will take in my own research. In order to work towards answering the question, “How can machine ethics be programmed into artificial intelligence technology?”, I will collect information from a variety of people, and in turn, expand AI’s ethical database.

Methods:

As a nineteen-year old college student with only a brief computer science background, there is not much that I can contribute in the programming realm of artificial intelligence. However, I am very much capable of collecting data regarding moral decision making of the general population. My plan is to come up with 15 ethical dilemmas that both humans and AI technology would come across in everyday life. Once I have these dilemmas laid out, I will travel by car to eight different California cities, and in each one, I will ask these 15 ethical dilemmas to ten different people. The eight cities I plan to go to are the following: San Diego, Los Angeles, Santa Barbara, San Luis Obispo, San Jose, San Francisco, Berkeley, and Oakland. After starting in San Diego, I am going to drive up the coast of California. I will stay for two days in each city, until I reach my final destination. On each one of the days in a given city, I will interview five strangers who will be chosen at random, to ensure an unbiased data set. After the fourth city I visit, San Luis Obispo, I will take one day of rest before continuing onto San Jose. Once I finish up my interviews in Oakland, my last stop, I will conclude my data-gathering, and fly back down to San Diego. One may ask, "Why can't this data collection be done through an online service?" If I were to proceed with my surveying through the web, there are a few disadvantages in comparison to face-to-face interviews. First, I will lose the benefit of having the general population as my interviewee pool. If online surveys were my approach, people who are more often on the internet will be providing the bulk of the responses. Also, out of the people who are often on the internet, mostly people with ample free time will be inclined to spend some of their day filling out surveys. Second, it is exponentially more difficult to find someone who will agree to fill out an online survey that presents itself while this person is browsing the web. Instead, if I were to be walking around a Santa Barbara beach and I asked someone to answer a

couple questions about morality, I believe the positive response rate would be much greater. For these reasons, I see the importance in incorporating travel into my research process.

One important step I must take before my data collection commences is receiving IRB approval. This has its own process, and it starts by completing a mandatory online certification. After doing so, I must fill out an IRB Research Project application and prepare informed consent documents. Once these steps have been taken and the IRB has approved my process, I will be able to carry on with my research.

Expected Results:

After the end of my data collection process, I will analyze the results and look for patterns and similarities in the responses given by the California-residing individuals. Using these patterns, I will put together an article which briefs the general Californian viewpoints on different aspects of morality and ethics. Since there is no objective definition of morality, the best that engineers can do is find generalized opinions on human ethics. The larger the ethical database is, the more likely it will be approved by a greater number of people. This means that if artificial intelligence technology applies this database to its own moral decision making, it will also likely receive approval of the general public.

Conclusion:

At the end of my research, my goal is to have benefitted scientists' process of improving the human resemblance of artificial intelligence's morality. By increasing the science world's grasp on the general consensus of ethics, artificial intelligence can be more seamlessly integrated into our society. The end goal of human and AI cooperation has many advantages that can

expedite humanity's constant pursuit of efficiency. The start of integration has already begun with concepts like automated vehicles, and a spike in integration is not too far off into the future. By conducting this research, I will be playing a role in this scene.

Timeline:

May 15-16, 2019- Start research in San Diego.

May 17-18, 2019- Drive up and conduct interviews in Los Angeles.

May 19-20, 2019- Drive up and conduct interviews in Los Angeles.

May 21-22, 2019- Drive up and conduct interviews in Santa Barbara.

May 23-24, 2019- Drive up and conduct interviews in San Luis Obispo.

May 25, 2019- Rest in San Luis Obispo.

May 26, 2019- Drive up and conduct interviews in San Jose.

May 28, 2019- Drive up and conduct interviews in San Francisco.

May 30, 2019- Drive up and conduct interviews in Berkeley.

June 1, 2019- Drive up and conduct interviews in Oakland.

June 3, 2019- Conclude research and drive back down to San Diego.

Budget:

Travel – Gas for ~1000 miles: $150

Food- 3 meals per day for 20 days: $480

Lodging- 19 nights of hotels: $1900

Total: $2530 + $70 for miscellaneous expenses = $2600

Works Cited:

Anyoha, Rockwell. "The History of Artificial Intelligence." *Science in the News*, 28 Aug. 2017, sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.

Fisher, Michael. "Engineering Moral Agents – from Human Morality to Artificial Morality." *Dagstuhl Seminar 16222*, 19 May 2016, http://drops.dagstuhl.de/opus/volltexte/2016/6723/pdf/dagrep_v006_i005_p114_s16222.pdf

Katte, Abhijeet. "Can Morality Be Engineered In Artificial General Intelligence Systems?" *Analytics India Magazine*, 10 Oct. 2018, www.analyticsindiamag.com/can-morality-be-engineered-in-artificial-general-intelligence-systems/.

Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, and Astro Teller. "Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, Sep. 2016, http://ai100.stanford.edu/2016-report.