**Module 1: Option #1**

Lance I. Evans

Colorado State University Global

CSC510-1: Foundations of Artificial Intelligence

Dr. Jonathan Vanover

26 December 2022

**Safety in AI**

The advancements of technology as a whole and more specifically artificial intelligence (AI) have continued to progress far beyond what we thought capable 60 years ago. With these advancements, there are many concerns regarding AI and our future. With the immediate concern being the replacement of humans for warehouse work by machines, which will later replace the entire factory with self sufficient AI robots. This concern is well founded, but the bigger picture is being missed, when AI is able to run factories without human intervention, humans will not have to work anymore, they will then be able to fill their time with pass times and hobbies. Though to get to that point there are several things that need to be addressed and fixed. The first being the biases in AI and the present concern that we are faced with. The next being the reality of the concern of AI safety. Finally, how are we to give AII morals and ethics?

**AI Bias**

The current issues AI biases in today’s advancements revolve around how humanity perceives itself. While observing the societal norms of humanity today, there are many conscious and unconscious biases that are prevalent. Whether that is from the misaligned viewpoints of racial inequities or from the ability to be honest about oneself. The simple solution to correcting the biases in AI today would be to either create a dataset without “undesirable biases” (Cirillo et al., 2020) or answer questions regarding human values (Irving & Askell, 2019). An undesirable bias is defined as “that which exhibits unintended discrimination” (Cirillo et al., 2020). Sources of undesirable bias span from a wide range of potentials, from the lack of diversity (meaning lack of appropriate representation of appropriate socio-economic groups within each racial background) of the sample size, to the algorithm itself having biases (Cirillo et al., 2020). Sufficiently vetting the sample size with an appropriate tool to ensure the sample size meets the requirements of the project and meets appropriate diversity standards could help. In most projects involving AI, understanding human values is essential in creating a successful AI project. In an attempt to understand the values of human, there are issues that are prevalent. The most important being the inability of humans to answer questions accurately about values, which could be caused by the misunderstanding of questions being asked (Irving & Askell, 2019). Other reasons that lead to inaccurate responses are cognitive and ethical biases, lack of domain knowledge, limited cognitive capacity, and correctness may be local (Irving & Askell, 2019). As we progress towards a more AI-centric reality, the fact that it is difficult to remove undesirable biases from basic datasets, the more advanced concepts of AI will also reap undesirable biases and may lead to unwanted outcomes. Although, there are biases that are intended to be there, like within the medical field. Some biases need to be there to ensure the best solution is presented to ensure proper diagnosis and treatment. These biases include gender, sex, genetics, social status, education, among many more (Irving & Askell, 2019). With the aforementioned biases included, there can be a more accurate determination of mental health well-being. However, using the biases improperly could result in detrimental determinations.

**Reality of AI Safety**

With the understanding that most people have seen the science fiction movies involving AI uprisings, the greatest fear amongst these is that AI learns that either to save this world, humanity must be extinguished, or to save self, humanity must be extinguished. This fear is built on the assumption that through analyses of everything that humanity has done (e.g., war, global warming, food shortages, etc.), humanity is the root of all the problems that exist in the suboptimal world that we live in, and to mitigate the risk of further degradation of the planet or progress, the best way to do that is through ridding the world of the disease “humanity.” In an attempt to produce an AI robot with human-like movement, companies like Boston Dynamic have subjected their robot to torture-like instances to develop real-world movement mechanisms (Goode, 2018). These sorts of displays of animosity have received both awe and fear expressions from the general public, the awe from the way the robot is able to maintain its’ balance, and the fear from the memory of these robots experiencing the torture and what could happen in the future if these robots were to develop a semblance of sentience. Though, the idea of these conditions was to determine or test whether the robots are able to handle unexpected conditions and maintain their balance. These tests may seem inhumane, but we must note that these robots are not yet human, nor do they have consciousness. If and/or when they become conscious, then we have another question to be visited. When Elon Musk stated “there should be some regulatory oversight, maybe at the national and international level, just to make sure we don’t do something very foolish” he is right. We, as humans, make many mistakes before correcting to fix a problem, in engineering it is no different. “As humans, we are inevitably constrained in our capacity to foresee and prevent the kinds of risks that only superintelligent machines could pose” (Goode, 2018). With appropriate regulations in place to mitigate the potential disastrous outcomes, Musks’ worries may be quelled. The threat of AI is seen as a serious one by influential intellects, including Musk and Harking. The famous macro-historian and futurist Yuval Noah Harari speaks of AI and climate change in the same breath, as both being existential crisis of a similar magnitude (Goode, 2018). Though the future is unwritten, there is an undeniable agreement that the progression of AI can be a scary one, especially without rules in place to protect humanity.

**Hardcoded Morals and Ethics?**

Can ethics and morals be hardcoded into the programming of AI? The argument would have to be yes, due to the ability to actually do so. However, it is not merely that simple, there may be a way to hardcode morals and ethics, but in that way the compassion or empathy we are seeking will not be there. As a basic guideline, there can be specific values or rules that can be hardcoded as a staple of their being. Just as with humans, ethics and morals are not hardcoded into our DNA, we learn these as we progress through life. To ensure AI receives the ethics and morals that we want them to have, they must be trained through supervised learning. Although, after they have learned through a training dataset, they will continue to learn through real-life interactions, which could allow them to learn from undesirable data, which could corrupt the entire morals and ethics that were previously “instilled” in them. “By contributing to create a more complex chain of decision-making and action, AI may make more difficult for individual agents to make sense of the reasons why a certain decision was taken, what their role exactly was in the operation, and, in general, whose reasons and what reasoning were governing the system they are part of” (Santoni de Sio & Mecacci, 2021). Were morals to be hardcoded into the AI program, the creator would then be responsible for whatever actions are taken by the AI, were they to be learned, the creator would still be partially responsible because of the datasets used to train it initially, but who is to take responsibility for an AI that is trained and then learns differently down the road? Should it still be the creator? But that wouldn’t be right, because as children grow and become adults, the parents aren’t held to the responsibility of an adult’s actions, so wouldn’t it be right for the creator to have the same treatment as the parent in this circumstance?

**Conclusion**

To conclude, the developments of AI have been substantial over the years, but there are serious concerns about where we are going and what precautions are being taken moving forward. We initially discussed the biases in current models of AI and how they have already shown that they have caused issues with current projects. Next, we discussed the perception of AI and how projects have reinforced the fear that people have with the overall safety of AI. Finally, we discussed how morals and ethics could be hardcoded or learned, but who should take responsibility after the AI learns from other people?

**References**

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and Healthcare. *Npj Digital Medicine*, *3*(1). https://doi.org/10.1038/s41746-020-0288-5

Goode, L. (2018). Life, but not as we know it: A.I. and the popular imagination. *Culture Unbound*, *10*(2), 185–207. https://doi.org/10.3384/cu.2000.1525.2018102185

Irving, G., & Askell, A. (2019). AI safety needs social scientists. *Distill*, *4*(2). https://doi.org/10.23915/distill.00014

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with Artificial Intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x