I certify that I have read the requirements set out on this form, and that I am aware of these requirements. I certify that all the work I will submit for this course will comply with these requirements and with additional requirements stated in the course outline.

Course Number: ENCS 393
Name: GEORGE MAVROEIDIS
Signature:

Instructor: Dr. Brandiff Caron
I.D. # 40065356
Date: 25-11-2021

Can AI be designed with morality?

There is no doubt that technology has dictated humanity's evolution in recent centuries. Throughout different eras, moral standards have been reinvented based on the co-constructivist relation between technological innovation and humanity's social progress for a better life. Current ideologies challenge the idea of artificial intelligence cooperating with humans morally with optimal outcomes. There is a definitive approach to design an artificial machine to maximize fairness and make great moral decisions. This is possible in the conditions that moral values can be explicitly defined and recognized by a machine as well as doing so in a transparent and regulated fashion so the appropriate design choices are objectively enforced.

Humans process dilemmas through reasoning, an aspect that a machine can only possibly simulate through complex metrics. It is argued that we are still not able to define clear standardized ethical norms soundly, but we are able to make moral decisions based on what we naturally feel, regardless if they are morally right or wrong [1]. Making ethical values computable, including the natural flaws that are part of the human factor, is the most optimal introduction to moral artificial intelligence. This means that artificial life will no longer be entirely value neutral and will carry a sense of responsibility. Therefore, personality traits must be characterized into them to align with the human factor for decision-making and to maximize fairness across multiple moral compasses. For example, self-driving cars must not only make the swiftest decision, but also the morally correct one that humans would if it was possible for them to execute it on time. "Society has to trust self-driving cars in order for them to ultimately realize their potential for saving lives, and they won't do that if the cars behave in ways that conflict with their moral values" [2]. Although a machine's purpose is to make better and faster decisions, it won't reach that stage until it starts to understand the human notion of decision-making first.

Despite wanting machines to facilitate moral complexities and discovering better moral compasses, it cannot be accomplished under the control of a singular creator. Regulation, openness and transparency of the appropriate data and architecture choices of a machine, will determine the least dangerous and destructive path of its evolution. Crowdsourcing human morality can offer the most diverse solutions for the proper training model. For instance, MIT's Moral Machine project shows how crowdsourced data can be used to effectively train machines to make better moral decisions in the context of self-driving cars [1]. In addition to data regulation, the process of converting ethical norms to programmable metrics must be transparent in order to determine whether AI systems are responsible for their decisions. Ethical accountability can be determined for self-driving cars if detailed logs of all automated decisions are kept at all times [1]. Thus, the system's behavior can be determined and experts will make the proper adjustments for creating a smarter machine.

In conclusion, artificial intelligence has reached a point where it is expected to replicate or even surpass the level of moral decisions made by humans. This has led to criticism regarding whether current AI systems are capable of such actions independently. In reality, it is possible for artificial machines to be designed with moral capabilities only if moral values are designated clearly for a system to understand and ensure transparency and regulation throughout its life cycle. Machines won't understand morality unless humans understand it first. A daunting task like this will be challenging with many pitfalls along the way. But its success will encourage us to improve on our human flaws.

References:

[1] Dr. S. Polonski, "*Can we teach morality to machines? Three perspectives on ethics for artificial intelligence*", 19-Dec-2017. [Online]. Available: https://medium.com/@drpolonski/can-we-teach-morality-to-machines-three-perspectives-on-ethics-for-artificial-intelligence-64fe479e25d3 [Accessed: 26-Nov-2021]

[2] *C. Q. Choi, "The Moral Dilemmas of Self-Driving Cars",* 25-Oct-2018. [Online]. Available: https://www.insidescience.org/news/moral-dilemmas-self-driving-cars [Accessed: 26-Nov-2021]