LUMSA UNIVERSITÀ

Integral Human Development

Annual Seminar of the International PhD Program & Research Network Contemporary Humanism in cooperation with the Postdoc-Fellowship Program in Integral Human Development

at CADOS – Católica Doctoral School

5-9 September 2022

PhD Candidate: Marco Tassella (Contemporary Humanism - LUMSA, UCLy)

"Enhancing free will? New perspectives on Moral Artificial Intelligence"

Keywords: "Free Will", "Moral Luck", "Moral Responsibility", "Artificial Moral Agents", "AI"

This project's objective is to analyze the relationship and the influence that the upcoming Artificial Intelligence technologies ("AI" for short) may have on our current knowledge of the human mind, of its transcendental structures, and of its ability to implement free will and intentional actions.

Specifically, this research tries to tackle a double challenge: on the one hand, we will examine how the development of the AI could also help us to better understand the basic moral structure of human reasoning and evaluation; on the other hand, we will try to understand how these discoveries could be used to develop technological tools to serve our needs, and to create better "Artificial Intelligence for social good" (Floridi, 2022). Before we begin, let us start with a couple of quick, but essential premises:

1) Our current knowledge about the conceptual status of free will is, unfortunately, quite lacking. From a philosophical standpoint, the many theories that exist on the topic are often contradictory. Furthermore, the real problem is that none of those theories are grounded on scientific knowledge, nor based on any sort of resistant reasoning to ultimately *justify* the existence of free will.

2) The field of AI is an ever-evolving research; yet, what today seems challenging, may be true as of tomorrow. It is admitted that, nowadays, these "intelligent" machines are still functionally very different from the human species and its abilities; however, this might slowly change in the future.

The claim that I wish to study is if, *by working on new AI models who emulate some of the specifically human skills, we could get a better understanding of our mind, of the way it works, and how it interacts with our actions*. By "specifically human skills", essentially, I refer to the specific set of abilities that are inherently tied to our decision-making skills and to our free will (e.g., our ability for reasoning, judgement, risk-management, moral evaluation, self-forming actions, torn decisions, etc.).

A new sub-field of AI, called "Artificial Moral Agents", is currently trying to develop and verify the above-mentioned claim. Let us start with a basic definition (Cervantes, 2016): an Artificial Moral Agent (or "AMA" for short) is *"a complex piece of software (or "virtual machine"), capable of engaging in decisions that involve moral implications, in a direct or indirect way*.". This implies two fundamental facts: one is that Artificial Moral Agents do not require to be implemented into robots or physical machines to work, and the other is that most of their actions are taken to have moral consequences or implications, meaning that these agents are already considered to be "morally apt" by definition. Cervantes goes on by specifying that "*this moral behavior may be based on ethical theories such as teleological ethics, deontology, and virtue ethics, but not necessarily"*. This implies that we could surely develop these AMA using one of the many ethical theories that we already know and understand, but also that said list of ethical theories is not necessarily exhaustive: we might also be able to create a new kind of ethical framework that could be valid for both humans and machines. This claim remains controversial, but it might still be interesting to entertain the thought.

In any case, these Artificial Moral Agents are already being used in many fields (like healthcare or the justice system), to help humans in some of their "moral" endeavors: they partake in moral decisions by calculating odds, by confronting data, by recalling past occurrences of a specific situation, by presenting similar cases, etc. The actual human, which is ultimately responsible for the decision (see Kane, 1996), should be able to interact with the machine and make a more "informed" choice (*as if* he or she were taking an "advice" from someone): by receiving more information, he or she may be able to achieve a better, more complete understanding of the situation, by enlarging his or her field of considerations ("wisdom") and being able to decide and act in a more free and "up-to-date" way.

In the future, these machines may become even more useful than they currently are. And yet, as of today, there is still a fundamental *caveat*: we currently cannot understand how to teach them some of

those specific tasks that we, as humans, are able to perform "naturally". These AMAs are missing many essential skills, skills that nowadays are considered too difficult to implement in a virtual system: AMAs cannot understand the meaning of things, nor act in a rational (albeit not necessarily "free") manner.

There are essential philosophical problems involved in this situation. Many of these issues are related to the link between "knowledge" and "know-how", others relate to our freedom, others refer directly to our insights on the human mind: how can we give machines a common sense, so they can "understand" the task they have been assigned? (Searle, 1980); what is the difference between a human "mistake" and a machine "error"? (Brown, 1994); what does it mean to act "freely"? (Kane 1996, Mele 2006); what would it take, for a machine, to be conscious? (Dennett 1991, Kurtzweil 2006); does the knowledge of good necessarily imply the goodness of actions? (Davidson 1980, Mele 1986).

By developing these AI, we might be able to achieve a more complete picture of our mind: we could have a better understanding of what it means to be *human*, what it is like to be free, what freedom entails in terms of moral responsibility, and so on. All of this would be done hoping that, by better understanding our mind, we could then proceed in building better, more useful AI. The final objective will be, therefore, the enhancing of our lives and perhaps even of our ability to take better, more informed decisions. In this regard, there would be much more to be said: during the upcoming years, my objective will be to try and approach some of these very controversial questions.

Finally, let us briefly consider the methodology for my research: I will begin by working on the main literature on Artificial Moral Agents (Cervantes 2016, Costantinescu 2022), on philosophy of consciousness (Chalmers 1996, Dennett 1991), and on machine ethics (Anderson 2011, Floridi 2022). The general context for this study will be a non-reductivist, non-dualist conceptual framework (Macarthur, 2010), which will make it possible to work on non-material entities (e.g., "beliefs", "objectives", "intentions" …), without lapsing into a cartesian-like dualism.
In addition to this, during the next year, I will collaborate with the "*New Humanism & AI*" (NHAI2022) and "*Anthropology in the era of trans-humanism and AI*" (CArisMA) projects at the *Université Catholique de Lyon*, in France, where I will be strengthening my general and technical knowledge about artificial intelligence, its social and philosophical implications, and its ongoing practical aspects. Moreover, the topic of AI is in constant development, and the literature on Artificial Moral Intelligence is growing exponentially: many new scientific articles are being published on a daily basis, in an ever-expanding field of fascinating new ideas.

Bibliography (presentation)

- Anderson, M., Anderson, S.L., *"Machine Ethics"*, Cambridge University Press, 2011.
- Brown, H.D., *"Principles of Language Learning and Teaching"*, Prentice Hall, 1994
- Cervantes, J.A., Rodriguez, L.F. (et al.), *"Autonomous agents and ethical decision-making"*, Springer, 2016.
- Chalmers, D., *"The Conscious Mind: In Search of a Fundamental Theory"*, Oxford University Press, 1996.
- Costantinescu, M., Vica, C., Uszkai, R., *"Blame it on the AI?"*, Springer, 2022.
- Davidson, D., *"How is Weakness of the Will Possible?"*, Oxford University Press, 1980.
- De Caro, M., Macarthur, D. (ed.), *"Naturalism and Normativity"*, Columbia University Press, 2010.
- Dennett, D., *"Consciousness Explained"*, Allen Lane, 1991.
- Floridi, L., *"The Ethics of Artificial Intelligence"*, Springer, Cortina, 2022.
- Kane, R., *"The Significance of Free Will"*, Oxford University Press, 1996.
- Kurzweil, R., *"The singularity is near: when humans transcend biology"*, Penguin, 2006.
- Mele, A., *"Is Akratic Action Unfree?"*, Phenomenological Research 46, 1986.
- Mele, A., *"Free Will and Luck"*, Oxford: Oxford University Press, 2006.
- Searle, J., "*Minds, Brains and Programs*", The Behavioral and Brain Sciences, 1980.