Title: A Defense of the Principle of Distributed Responsibility in Artificial Intelligence

Abstract: The world is rapidly moving towards a new era: the age of artificial intelligence. We are now surrounded by manmade systems and spend most of our days dealing with the artificial – using our computers, phones, ipads, televisions – and dealing less and less with that which is natural. The relationship between the artificial and the human has traditionally been characterized in normative terms.

One of the biggest challenges posed by Artificial Intelligence is the status of the normative concept of responsibility. With the emergence of increasingly developed technologies and artificial agents impacting the day by day, several problems arise but, at this stage, we don't seem to have the conceptual tools to engage with its solutions.

The main caveat for that is a specific responsibility thesis that is being used in the traditional scenarios of human-machine interactions, that we will henceforth call the "conventional thesis" (CT), which can be logically stated as follows:

(CT): x is a responsible agent iff x possess y;

A negative version is also applied: no Artificial Agent (AA) is responsible for its actions or motivations. On the basis of (CT) we can find a fundamental ontological assumption: the thesis that no (AA) posses, ontologically, the y property that is relevant to give moral status to an entity. Note that y is not a property of an unlimited open set but, rather, must be a morally relevant property for the normativity of responsible actions. Thus, y is conventionally equivalent to a set of properties whose legal codes and common sense commonly denominate "consciousness" or "self-awareness" that we can summarize as mentalistic or intentionalist properties. That is: for an agent x to be held responsible, x must necessarily have the property of being a conscious or/and self-conscious agent

This is, in a nutshell, what (CT) is about: a very plausible way of making moral judgments about actions caused by human beings and it actually works pretty well for a pre-artificial intelligence realm. The problem is that, again, we are reaching the age of artificial intelligence. The world is changing but our concepts of normativity regarding moral agency and moral responsibility remain the same as before (cf. Verbeek, 2011). We are trying to create fire with stones whilst ignoring that stones are no longer the best way to create fire: we are using the same ethical assumptions to think a world of actions where human beings are no more the only player in town.

In this talk, we will argue that (CT) creates what it is known as the "responsibility gap" (RG). We will show that (CT) is implausible to deal with the rapid development of current technologies and, at the shoulders of Luciano Floridi (2013, 2016), we will propose the concept of Distributed Responsibility (DR) to close the current normative gap (cf. Gouveia, 2018, 2019).

Keywords: Artificial Intelligence, Artificial Agents, Responsibility Gap, Distributed Responsibility

References:

Floridi, L. (2013) “Distributed Morality in an Information Society” *Science and Engineering Ethics*, 19 (3): 727-743.

Floridi, L. (2016) “Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions” *Phil. Trans. R. Soc*, A 374: 20160112.

Gouveia, S. S. (2018) *Philosophical Reflections: Art, Mind and Justice* (Portuguese Edition), Famalicão: Editora Húmus.

Gouveia, S. S. (2019) “A Defense of the Principle of Distributed Responsibility in Artificial Intelligence” *Revue Roumaine de Philosophie*, 63 (2): 235–248.

Verbeek, P. (2011) *Moralizing technology: Understanding and designing the morality of things*, Chicago: University of Chicago Press.

Steven S. Gouveia is a doctoral researcher at the University of Minho (funded by FCT: SFRH / BD / 128863/2017), supervised by the philosopher Manuel Curado and the neuroscientist Georg Northoff. He is a visiting researcher at Minds, Brain Imaging and Neuroethics Unit at the Royal Institute of Mental Health at the University of Ottawa. He is a researcher at Mind, Language and Action Group, Institute of Philosophy at the University of Porto, and at the Lisbon Mind & Reasoning Group, IFILNOVA - Universidade Nova de Lisboa. He has published 8 books on various topics such as Philosophy of Mind, Bioethics, Democracy or Artificial Intelligence, with prefaces by Noam Chomsky and Peter Singer. He will soon publish his new book “Homo Ignarus: Rational Ethics for an Irrational World” and he is the host of the international documentary “The Age of Artificial Intelligence”. More information: www.stevensgouveia.weebly.com.