Massimiliano L. Cappuccio (UAE University)

Autonomous Artificial Agents, Ethical Decision, and the Frame Problem

According to various influential theorists, including most notably philosophers Daniel Dennett, Hubert Dreyfus, Michael Wheeler, and roboticist Rodney Brooks, artificial systems that rely exclusively on stored heuristics and internal representations of the world are irremediably destined to suffer from the so called "Frame Problem" (or, to be more precise, the metaphysical version of it): that is, the structural impossibility of deciding what information is relevant to a certain task. While many solutions to particular versions of this problem have been proposed, no general solution has been successfully implemented yet. Today, the philosophical discussion about the Frame Problem is at the center of the foundational debate in social robotics: in particular, the Frame Problem seems to be one of the most resilient difficulties in programming "ethical machines", i.e. machines that are capable to take justified decisions in morally and socio-culturally sensitive circumstances. Three case studies analyzed by the literature are particularly significant for the discussion of this problem, and my paper aims to briefly highlight the similarities and their communal normative structure: 1) the first is the case of self-driving cars facing ethical dilemmas comparable to the famous "trolley problem"; 2) the second is the case of autonomous combat drones devised for the use of lethal weaponry; 3) the third is the case of social robots equipped with mindreading modules meant to allow them to infer the intentions and reasons that motivate the instructions of a human operator. This capability would allow the robot to decline or protest the instructions that appear to contradict the real intentions of the operator.

Given the holistic and pervasive nature of the Frame Problem, even the implementation of a schematic code of conduct seems to pose insurmountable complications for these kinds of human-robot interactions. I will compare the two theoretical approaches proposed by Dennett and Dreyfus, pointing out that – while the former understands the Frame Problem in terms of cognitive overload, the second interprets it as an intrinsic insensitivity to context of the cognitive architectures based on representations. My paper suggests that Alan Turing, in describing his famous "Imitation Game", was already aware of the general sense of this problem. Like Dreyfus, Turing had already anticipated that a solution, if any exists, involves the creation of truly embodied and environmentally situated self-learning machines: these are the only systems that could ever be capable to interpret the real value and the ethical implications of their decisions against a concrete background of contextually situated and massively interconnected reasons. I will argue that this approach is the most promising one, but I will also discuss the a priori reasons that, given the current state of AI research, make its implementation in the previously examined case studies extremely challenging.

Briggs G. (2012), Machine ethics, the frame problem, and theory of mind, *Proceedings of the AISB/IACAP World Congress*.

Cappuccio M., Wheeler M. (2012), Ground-level intelligence: action-oriented representation and the dynamics of the background, in Z. Radman, *Knowing without Thinking*, Palgrave-MacMillan, 13-36.

Dennett D. (1984), 'Cognitive wheels: The frame problem of AI, in *Mind, Machines, and Evolution*, Cambridge University Press.

Dreyfus H. (1992), *What Computers* Still *Can't Do*, The MIT Press.

Klincewicz M. (2015). Autonomous Weapons Systems, the Frame Problem and Computer Security. *Journal of Military Ethics* 14 (2):162-176.

Turing M. (195), Computing machinery and intelligence, *Mind*, 59(236), 433–460.

Dr. Massimiliano (Max) Cappuccio is Assistant Professor in Philosophy of Mind and Cognitive Science at the Department of Philosophy of UAE University, where he directs the Interdisciplinary Program in Cognitive Science. He is also a member of UAEU Laboratory of Psycholinguistics, run in collaboration with New York University Abu Dhabi, and a founding member of the UAE branch of the IEEE Robotics and Automation Society. He is a correspondent member of the Neurophilosophy Lab of the State University of Milan, Italy. He is currently working on a UAE-NRF-funded interdisciplinary experimental project on the neuronal bases of skill acquisition and disruption. He is currently editing the *MIT Press Handbook of Embodied Cognition and Sport Psychology*.