Review of: "More Human Than All Too Human: Challenges in Machine Ethics for Humanity Becoming a Spacefaring Civilization"

Giuseppe D'Acquisto

Potential competing interests: No potential competing interests to declare.

The paper deals with the theoretical, yet very interesting, situation for ethical considerations of an AI system assisting humans on long distance travelling towards a space-faring civilization. The ethical implications are evident considering the “non return” effects and the need not to expose humans to irreversible consequences after a decision has been taken by a machine. The paper revolves around the question as to whether a machine can be, or become, a fully ethical agent, and be designed to pursue ethical goals, without which it may pose an existential threat to humanity. After a review of the existing literature, the author introduces the notions of “implicit ethical agent” and an “explicit ethical agent”. Machines that are implicit ethical agents are ones that have been programmed by a human to follow ethical behaviour or avoid unethical behaviour, whereas machines that are explicit ethical agents can use ethical principles in calculating the best course of action when confronted with an ethical dilemma. The latter case is the most radical and difficult to tackle, because it involves the notions of consciousness and responsibility. The opinion of the reviewer is that these notions cannot have an algorithmic equivalence, because, saying it with Pylkkö (1998), “*What we say about the thing which is in present-day neuroscience called the brain, may, in light of the future research, turn out to be based on a ridiculous conception... What today we happen to be denoting by the word brain may, according to future conceptions, simply turn out to be nonexistent*" and also “*experience includes both perceptual and non-perceptual elements, i.e. it covers more than sense perception*”. Being responsible means being accountable also for possibly wrong decisions or apparently irrational, or without perfect knowledge of the environment, undertaken with a view to pursuing a broader beneficial goal (or a selfish one), and being capable of justifying the criterion of the choice. Machines are designed to cope with a deterministic environment, or they can handle uncertainties (through bayesian or causal - à la Pearl - decisions) but cannot cope with indeterminate situations, which require creativity and blind risk acceptance. These (again with Pylkkö) are "aconceptual concepts": humans just survive issuing general predicates that, according to their senses, describe the functioning of the world, until they are falsified and replaced by other postulates equally waiting to be falsified (as Popper put it). A reductionist mechanistic, or automatic, approach to ethics, pretending that machines will take over humans on fully autonomous decisions, may be at least problematic, or difficult to justify on the sole premise of data and mathematical models. The reviewer believes that rather than focusing on the threats that machines, especially super-intelligent machines, as fully autonomous ethical agent will bring to humanity when they substitute humans in decisions, we should better develop a theory of co-existence of humans and intelligent machines, based on the premise that humans will not relieve their role of dominance vis-a-vis artificial intelligent systems. In this respect, it is very worth considering the work conducted in the area of cooperative inverse reinforcement learning (see Russell, “Human Compatible: AI and the

Problem of Control", 2020) as a practical and non conflicting way to solve ethical dilemma when machines (even super-intelligent ones) are involved in the decision.

Additionally, the reviewer would like to point out some repetitions (the sentence "some argue that it can also be evaluated from the perspective of ethical norms" is repeated three times) and inconsistencies in the references (for instance, Alan Turing (1989)).