Software with Ethical Discernment

Luís Moniz Pereira and Ari Saptawijaya

NOVA LABORATORY FOR COMPUTER SCIENCE AND INFORMATICS (NOVA LINCS)

UNIVERSIDADE NOVA DE LISBOA, PORTUGAL

ABSTRACT

Machine ethics is a burgeoning field of enquiry that emerges from the need of imbuing autonomous agents with the capacity for moral decision-making. It has attracted interest from the artificial intelligence community, and brought together perspectives from various fields: philosophy, cognitive science, neuroscience and evolutionary biology. The overall result of this interdisciplinary research is not just important for equipping agents with some capacity for making moral judgments, but also for helping better understand morality, via the creation and testing of computational models of ethical theories.

Keywords: Machine Ethics, Computational Logic, Evolutionary Game Theory, Evolution of Morality.

Our research has focused on logic programming techniques for modeling of morality aspects, namely moral permissibility and the dual-process of moral judgments. The main characteristics are captured by its available ingredients, which include abduction, integrity constraints, preferences, argumentation, counterfactuals and updates, framed together in an agent's life cycle architecture. This agent life cycle architecture concerns itself only with the realm of the individual, where computation is a vehicle for modeling the dynamics of knowledge and moral cognition of an agent.

In the collective realm, norms and moral emergence have been studied computationally, using the techniques of Evolutionary Game Theory. Our research shows that the introduction of cognitive capabilities, such as intention recognition, commitment, and apology, separately and jointly, reinforce the emergence of cooperation in the population, comparative to absence of such cognitive abilities. Modeling such individuals within a population helps understand the emergent behavior of ethical agents in groups, in order to implement them not just in a simulation, but also in the real world of future robots and their swarms.

Our work thus contemplates two distinct realms of machine ethics, the individual and collective, and identifies bridges of connection. In studies of human morality, these distinct interconnected realms are evinced too: one stressing above all individual cognition, deliberation, and behavior; the other stressing collective morals, and how they emerged with evolution.

Machine ethics is becoming a pressing concern, as machines become ever more sophisticated, autonomous, and act in groups, amidst populations of other machines and of humans. But ethics, jurisprudence, and legislation, are lagging much behind in adumbrating the new ethical issues.

A fundamental question arises concerning the study of individual cognition in groups of often morally interacting multi-agents (that can choose to defect or cooperate with

others): whether from such study we can obtain results equally applicable to the evolution of populations of such agents. And vice-versa: whether the results obtained in the study of populations carry over to groups of frequently interacting multi-agents, and under what conditions.

Specifically with respect to human morality, the answer would appear to be a resounding 'Yes'. For one, morality concerns both groups and populations, requires cognition, and will have had to evolve in a nature/nurture or gene/culture intertwining and reinforcement. For another, evolutionary anthropology, psychology, and neurology have been producing consilient views on the evolution of human morality. Their scientific theories and results must per force be kept in mind, and serve as inspiration, when thinking about machine ethics. All the more so because machines will need to be ethical amongst us human beings, not just among themselves.

The very study of ethics and evolution of human morality too, can now avail themselves of the experimental, computation theoretic, and robotic means to enact and simulate individual or group moral reasoning, in a plethora of circumstances. Likewise regarding the emergence of moral rules and behaviors in evolving populations. In bridging these realms, cognition affords improved emerged morals in populations of situated agents.

At the end of the day, we will certainly wish ethical machines to be convivial with us.