Simulation as a Method of Ethics: simulating agents by programming, to elucidate autonomy, agency and ethical virtues

Fernando da Costa Cardoso [1], **Luís Moniz Pereira** [2]

[1]NOVA LINCS, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
CNPQ, Brazil. Email: promenadex@gmail.com

[2]NOVA LINCS, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal. Email:lmp@fct.unl.pt

Abstract

We set forth a case study of the integration of philosophy and computer science using artificial agents ruled by abductive logic and emergent behavior. Our first step is to highlight different models that we developed of such agents (a set of them related with evolutionary game theory and another a model of a narrative storyteller robot). As we indicate, each model exemplifies different aspects of the ethical dimension that can be investigated with resource to simulation in silicon.

Simulation can guide ethical investigation

Our aim is to suggest that simulation can guide ethical investigation. We propose to see, in the emergence of artificial agency, the multi-pole development of different characteristics such as guilt, tolerance and reciprocity and how the development of such capacities of artificial agents accordingly seeds grey areas of the ethical inquiry. Although this suggestion is not exactly new (Danielson, 1992), ours is a proposal for a more global reevaluation of what consists investigation in ethics on the basis that we have good reasons to recognize suitably developed artificial systems as constituting something beyond simple tools for this inquiry, as suggested by Grim (Grim, 2004). This suggestion relates with these systems' crescent autonomy and independence toward its creators, something that helps to provide not only a toolbox but also a relatively closed universe (a laboratory) for the investigation of the aforementioned characteristics that constitute the ethical dimension.

Based on this distinction between a tool and a laboratory, our strategy is to open the space between programming efforts and philosophical investigation presenting, in a sense, a philosophical interpretation of a programming effort and to use such effort, without fears of circularity, as a method of ethics. This will allow us to reassesses important ethical concepts in a way that may prove useful. The distinctiveness of our proposal is to present different programming efforts and to try to determine their role for ethical investigation. This leads to the elaboration of a general methodology, complementary to experimental approaches, to ethical investigation.

Two frameworks

Instead of providing an *a priori* argument to defend our conceptions, we analyze different programming models that were developed to better understand the grey area between simulation and emulation of ethics. Different approaches to this simulation/emulation have been suggested, like the ones that can be found in (Allen, Varner, and Zinser, 2000), but here we are going to describe models that rely on logical programming and Evolutionary Game Theory (EGT). Although limited, and low on the autonomy scale, we propose that these programming models highlight aspects that are informative of our own embodied moral conditions, due to these very limitations. Thusly, in addition, we wish to emphasize both the influence that this work might have in tutoring philosophical intuitions and, recursively, on the influence of this tutelage on the further development of these very same computational agents. The gains, relative to the complementarity that could exist in the joint development and co-evolution of the twine of computational models and philosophical theories that can be achieved through a better comprehension of us, outweigh those risks, such gains being themselves demonstrations of moral autonomy.

Although this modeling research has been undertaken for several years now, we are going to pay attention only to some of its latest developments. In particular, our aim forthwith is to analyze two specific models developed at our NOVA-LINCS center and its partner institutions:

● Agents developed in the context of evolutionary game theory simulations, in non-repeated and in reiterated two-person and public good games, where a diversity of successful simulations and analytic demonstrations have been made to better understand the joint role of recognizing in-

tentions, of commitment and of apology, for the promotion of emergence, in a population of agents, of combinations of stable morally cooperative behaviors, by agents who are at times able to recognize intentions, establish commitments, and accept apologies (Pereira, Santos, and Han, 2014) (Pereira, 2012) (Han and Pereira, 2013) (Pereira et al., 2014).

● The narrative storyteller about a robot that, as it attempts to save a princess, needs to successively deal with moral updating dilemmas, using the ACORDA logic programming system (Lopes and Pereira, 2006) (Pereira and Lopes, 2007) (Lopes and Pereira, 2010) (Pereira and Saptawijaya, 2011) (Pereira and Saptawijaya, 2015).

Evaluation

The aim of these models has been to establish, through logical formalization, frameworks where single agents and multiplicities of agents are able to employ flexible behavior in answer to the demands of virtual environments. We suggest going beyond this first immediate aim and reassessing these models, after describing them. As we tried to show before (Cardoso and Pereira, 2015), in those models we touch on something else that is valued from the point of view of Ethics, though in a manner deeply different from the kind of rationalist effort that contains itself only within thought experiments.

In the first simulation or, more precisely, in the first sets of simulations, the agents therein, which are proffered as possessing some degree of autonomy, are nevertheless simple in their evaluations reflecting the closed up system of the prisoner's dilemma matrix of losses and gains. However, this should not be taken as a limitation since the aim is to analyze the role of the interactions among multitudes of agents having different interests and strategies, in a framework that allows for distinct aims, in order to envisage how these different strategies evolve over an extended number of generations. The essential point is this: those strategies that emerge and become stable correlate with emergent norms. The second experiment we present provides a chance to evaluate autonomy during circumscribed social interactions or under social constraints, where autonomy plays a wider role even when dealing with simple agents.

An interpretation of these two computational efforts can tell to one interested in Ethics that the phenomena he tries to understand could be captured at a simpler level with fruitful results for that inquiry. A level certainly not yet surrounded by the great values that he so promptly tries to identify with Ethics, but whilst losing in the process of that very same identification a perspective that could have permitted a multitude of agents, one having different degrees (and attending constraints) of autonomy, and therefore providing a richer account of this dimension.

These unexpected and contra intuitive results allied with the prospect of an easier public evaluation of the models supports our belief that in them we find not only a simple tool but a global method of enquiry that could redirect mainstream philosophical ethics.

G. E. Moore, at the moment of the foundation of contemporary metaethics (Moore, 1903), sustained that a simple effort of will, from his cabinet in Cambridge, would clear the field of its confusions. We agree with Anscombe (Anscombe, 1958) that far from it, metaethics as the leading form of investigation produced not a translucent truth but a stalemate where different positions, some apparently convincing, some as exotic as Moore's own form of non-naturalism, comes and goes without fulfilling its task as a "Prolegomena to any future Ethics that can possibly pretend to be scientific" (Moore, 1903, p. iv).

This is related to the limitations and partiality of thought experiments and of any pure analytical effort enlarged by its difficulty with sharing these results in a way that permits public evaluation. We propose to adapt in the investigation of ethics an already established forum, here exemplified by logical programming and EGT, where these problems can be fixed by recognizing that, at one same moment, this forum reaches a point where we, indeed, face the task of building agents but where our evaluation of these agents demands something akin to the choice that Putnam (Putnam, 1972) thought would be increasingly a necessity: that we treat those agents not as tools but as ends.

Acknowledgments

FCC thanks Conselho Nacional de Desenvolvimento cientifico e tecnológico (CNPQ/Brazil). LMP thanks FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013.

References

Allen, C., Varner, G., and Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), pp. 251-261.

Anscombe, G. E. M. (1958). Modern Moral Philosophy. *Philosophy*, *33*(124), 1-19.

Cardoso, F., Pereira, L. M. (2015) On artificial autonomy emergence -- the foothills of a challenging climb, in: J. White, R. Searl (eds.), *Rethinking Machine Ethics in the Age of Ubiquitous Technology*, Hersheys, IGI Global, pp.51-72.

Danielson, P. (1992). *Artificial morality: virtuous robots for virtual games*. Routledge.

Han, T. A. and Pereira, L. M. (2013). Intention-based Decision Making via Intention Recognition and its Applications. in: Guesgen, H. and Marsland, S. (eds.), *Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security,* pp. 174-211, Hershey: IGI Global.

Han, T. A.; Pereira; L. M.; Santos, F. C.; and Lenaerts, T. (2013).

Good Agreements Make Good Friends. *Sci. Rep.*, 3, doi: 10.1038/srep02695.

Lopes, G. and Pereira, L. M. (2010). Prospective storytelling agents. In Carro, M. and Peña, R. (Eds.), *Proceedings of the Twelfth International Symposium on Practical Aspects of Declarative Languages*. LNCS series, Vol. 5937, pp. 294-296. Berlin: Springer-Verlag.

Moore, G. E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.

Pereira, L. M. and Lopes, G. (2007). Prospective Logic Agents, in: Neves, J. M.; Santos, M. F.; and Machado, J. M. (eds.), *Progress in Artificial Intelligence*, Procs. 13th Portuguese Intl. Conf. on Artificial Intelligence (EPIA'07), pp.73-86. Berlin: Springer.

Pereira, L. M. (2012). Evolutionary Tolerance. In Magnani, L. and Ping, L. (eds.), *Philosophy and Cognitive Science – Western & Eastern Studies*. SAPERE series, Vol. 2, pp. 263-287. Berlin: Springer-Verlag.

Pereira, L. M. and Saptawijaya, A. (2007). Moral Decision Making with ACORDA. In *Local Proceedings of the Fourteenth International Conference on Logic for Programming Artificial Intelligence and Reasoning (LPAR'07)*, Yerevan, Armenia.

Pereira, L. M. and Saptawijaya, A. (2011). Modelling Morality with Prospective Logic. In Anderson, M. and Anderson, S. L. (eds.), *Machine Ethics* (pp. 398-421). New York, NY: Cambridge University Press.

Pereira, L. M.; Han, T. A.; and Santos, F. C. (2014). Complex Systems of Mindful Entities – on Intention Recognition and Commitment. In: Magnani, L. (ed.), *Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues*, pp. 499-525. Berlin, Springer-Verlag.

Putnam, H. (1975). *Mind, language and reality: philosophical papers 2*. Cambridge, Cambridge University Press.

Saptawijaya, A. and Pereira, L. M. (2015). The Potential of Logic Programming as a Computational Tool to Model Morality. In Trappl, R. (ed.), *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*. Lecture Notes in Cognitive Technologies series. Berlin: Springer-Verlag.