The emergence of artificial autonomy: A view from the foothills of a challenging climb

Fernando da Costa Cardoso

Conselho Nacional de Pesquisa (CNPQ/Brasil)

NOVA Laboratory for Computer Science and Informatics

Universidade Nova de Lisboa, Portugal

promenadex@gmail.com

Luís Moniz Pereira

NOVA Laboratory for Computer Science and Informatics

Universidade Nova de Lisboa, Portugal

lmp@fct.unl.pt

Abstract

In this chapter we set forth a case study of the integration of philosophy and computer science using artificial agents, beings ruled by abductive logic and emergent behavior. Our first step in this chapter is to highlight different models that we developed of such agents (a set of them related with evolutionary game theory and one model of a narrative storyteller robot). As we indicate, each model exemplifies different aspects of the bottom of the hill of autonomy as an emergent property of artificial systems specified through three aspects ("Self control", “Adaptivity to the environment" and "Response to environment"). In summary, our conception is that autonomy, when presented as an emergent characteristic, could fill the important place given it by elaborations in philosophical ethics and one that leads us to a clearer comprehension of where to direct our efforts in the field of artificial agents. We conclude this chapter with the notion that this reevaluation of autonomy is necessary for the enhanced comprehension of human morality.

Keywords: Artificial morality, Moral machines, Autonomy, Ethics.

Introduction

Autonomy plays a central role in modern ethics, as on its basis we distinguish the class of ethical agents from the other beings - "only things" or at most Kant's "heteronomous" beings, but not fully ethical agents. Forged from medieval Christian conceptions of the divine origin of morality and tempered with the rediscovery of classical Greek philosophers, with their natural order establishing standards for fitness, excellence and virtue, traditional conceptions of autonomy continue to inform notions that often appear more central to the field. Freedom, personhood, responsibility and intentionality, aspects of autonomous agency by way of which some beings differentiate themselves from the mass of objects, tools, artifacts and things littering this physical world, at root derive their value in the aid they render in answering this fundamental question: autonomous, or not?

This binary is complicated by apparent grey-area cases. The existence of human children, incapacitated humans (mentally ill, intoxicated) and intelligent animals imply that "ethics" is not the simple domain of pure rational agents. Sensitivity to such cases is evident in the "moral patient" of Regan (1986) and Singer (1993). But, this sensitiveness remains exceptional. Autonomy often remains presented as an especial characteristic distinct to human beings differentiating them from other beings.

Our aim in this chapter is to suggest that artificial agency emerges with the multi-pole development of different characteristics, and that the development of capacities of artificial agents accordingly seeds the grey areas of agency. Ours is a proposal for the reevaluation of autonomy on the basis that we good reasons to recognize as autonomous suitably developed artificial systems. Our strategy is to open the space between programming efforts and philosophical investigation presenting, in a sense, a philosophical interpretation of a programming effort. This will allow us to redefine autonomy in a way that may prove useful. The first step is the recognition of the legitimacy in the conception of autonomy as is develops in response to challenges coming from different areas of historical investigation such as sociology and psychology. The second step, the one that distinguishes our proposal, is to present different programming efforts and to try to determine any sense of autonomy in these systems without diminishing what stands out from the preceding review of the modern tradition. This leads to our third step, the elaboration of the aforementioned frame, wherein different beings may be inserted as in a map revealing the different dimensions of autonomy as an emergent property. Lastly, in the final part of the chapter, we present some far-reaching consequences that we believe need to be addressed, and toward which we plan to aim in future investigations.

1- Setting the stage

Philosophers, particularly those in the rationalist tradition, have pictured autonomy as affording an important role not just in the practice of morality but also in the distinction of agents that are moral beings from others which are not. This tradition of so establishing autonomy or self-governance at the center of Ethics has a history that can be characterized globally and episodically. Globally, following Schneewind (1998), we may take it as a reaction against "the conception of morality as obedience" to some divine order. It is an internalization. Episodically, we may highlight the establishment of the Kantian formulation crowning autonomy as the central concept essential for morality, and thus Ethics, in the first place.

This crowning follows a natural development inside the modern moral tradition binding Ethics with human psychology. In this vein, when Kant established autonomy as the foundation of human dignity, he established an association between these two fields, something that could be verified in the following supportive statement: "Autonomy is thus the ground of the dignity of the human and of every rational nature" (Kant, 2002, AK 4:436). Autonomy reveals the particular constitution of our souls. In this sense modern Ethics, and with it the question of how we should live, merges a conception of what we should do with a conception of how our minds operate optimally.

The route that lead to this position certainly evolved in correlation with our views about our own psychology and accordingly can be captured in the progression that links the formulations of Descartes and Kant. Descartes affirms, when establishing his dualist view of the mind-body problem in the *First Meditation* of the *Meditations* (1641), that Ethics manifests itself as Freedom. By that is meant that the search for the proper way to act should be related with the best part of us, a part identified by Descartes with what in us is most perfect: our rationality. This leads Descartes in the sequence to the prescription of a certain way of life related with a certain way of thinking:

> Indeed, the more strongly I incline in one direction the more free my choice is (...) Freedom is never lessened – indeed it is increased and strengthened – by natural knowledge and divine grace. When no reason inclines me in one direction rather than another, I have a feeling of indifference – that is, of its not mattering which way I go – and that is the poorest kind of freedom. What it manifests is freedom considered not as a perfection but rather as a lack of knowledge – a kind of negation. If I always saw clearly what was true and good, I should never have to spend time thinking about what to believe or do; and then I would be wholly free although I was never in a state of indifference (Descartes, *Fourth Meditation*).

In this sense, the detachment between our free soul and our bodily passions establishes a thin line where real autonomy, identified as freedom, situates. Leave aside how the references to supernatural entities and their influence on us due to Descartes' historical context. Focus, instead on the role of this "natural knowledge" - this thin line is occupied exclusively by human beings that are then established as autonomous because they are able to see those motives of choice, the reasons, and are able to evaluate them as the "right ones" to serve as leads for actions. In summary, a capacity in our souls links our ability to proceed in a certain sense, and this ability is both the motive for our particularity in the animal kingdom and, again, a motive for the identification of this particularity with morality itself.

Certainly these connections could be further clarified and that is exactly what Kant did. The choices derived by those autonomous agents, human beings to be sure, constitutes the domain of the practical pure reason as Kant addresses in his *Groundwork of the metaphysics of morals*. There, in the third formula of moral law, also dubbed by Kant the formula of autonomy, he affirms that, when we act, we should act "not to choose otherwise than so that the maxims of one's choice are at the same time comprehended with it in the same volition as universal law". Here, Kant is cementing autonomy at the center of morality, because to self-govern according to pure rationality is a requirement for any moral action. Freedom is not choice, something that Descartes had already denied when equating choice with mere appearance dependent on our bodily functions. Since in this center we find a self-legislator, that is a purely rational being detached both from heteronomous motives (such as those that relate to passions and interests), and also from contexts, the dignity of this self-legislator becomes the dignity of autonomy.

This equivalence between freedom and autonomy certainly blinded the philosophers to certain challenges. This blindness could be justified with the shining of this self-legislator in positioning itself against all the odds, something that Kant so beautifully expressed with the will of this self-legislator. This is the good will, a will that “shines like a jewel for itself, as something that has its full worth in itself”. And the shine of this jewel is brightest as the good will distinguishes itself through action:

> though under certain subjective limitations and hindrances which, however, far from concealing it and making it unrecognizable, rather elevate it by contrast and let it shine forth all the more brightly (Kant, 2002, G, 4:397).

Consider that this Kantian "shine" is directed through the prism that is his notion of duty, and we have thus a rough sketch of the mechanism motivating Kantian moral philosophy, with duty providing the type of explicit action guidance central to the regulation through reason of practical life. The aim of this sketch is to highlight the centrality of human beings both on the ethical and evolutionary stages. It is because of a particular characteristic of human souls, rationality, that human beings are unique members of the "Ethical Association", a membership validated in its exercise, vis-à-vis autonomy.

Maybe is time to challenge this formulation. We suggest that, as stated, this formulation occludes any proper comprehension of both our ethics, and ourselves.

Certainly, there are historical reasons for this. For example, in her evaluation of modern moral philosophy, G. E. M. Anscombe argued that, behind the Kantian formulation - and in fact behind all formulations of Ethics since the beginning of modernity – there are the skeletons of that element so salient in the Descartes quote, above: our morality is something granted by God to human beings, alone. Anscombe attempted to show that we cannot sustain our modern conceptions of autonomy and privileged place in the universe alongside the view of a supernaturally derived moral duty. According to her, the idea of duty developed by Kant is just an impressive philosophical concealment of what is at root a divine gift to human beings.

From a sociological point of view, the same issue was raised in the notion of a progressive disenchantment of the world, first developed by Max Weber. Fast forward to today, and recent Psychology leads us also to question the aforementioned grounding, and to search for new ways to establish Ethics especially on natural grounds, a search facilitated by empirical comprehension of how our minds have evolved (c.f. Doris, 2010; Haidt, 2007; Greene, Nystrom, Engell, Darley, & Cohen, 2006). This work conceives of an innate moral grammar in the gray area between moral philosophy and psychology, a grammar patterning both our minds and our moral rules, and has been pursued as a challenger to the picture of ethical life inherited.

We are thus left with two competing visions. A disjunct, rationalist Ethics that segregates human beings from nature in such a way that we end up believing in clouds, and an empiricist Ethics confirmed and clarified through diverse scientific discoveries. Of these two, only one seems to exhibit the very dignity that rationalist conceptions of autonomy aimed to exalt.

Given this result, we are going to pursue here the suggestion of a necessary disentanglement between freedom and autonomy. This disentanglement may aid in abandoning and overcoming inherited conceptions of ourselves and of our Ethics. In order to motivate this transition, we will offer a demonstration of the evolution of artificial autonomous agents, suggesting that parameters necessary for their emergence provide us with reasons to transform conceptions of moral agency accordingly. Our attempt is based on the idea that developments in programming can be used as one set of tools, suggesting how to establish a better view on ethical concepts, and thus on ourselves. Thus, the twofold aim of this chapter: to analyze (1) how autonomy emerges in the context of artificial autonomous agents; and (2) how this development influences our own comprehension of what it means to be autonomous. As we will demonstrate, this dual aim of analysis becomes a matter of necessity in face of the increasing blur of the lines between, on the one side, the simulation of ethical traces – with programming tools – as a way to study aspects of ethics and, on the other, the emulation of those traces in agents that are becoming increasingly autonomous.

2- Emergence of autonomy in the context of artificial autonomous agents

In order to achieve this aim we start by pointing to the elements that convinced us that the emergence of artificial agents requires the development of a notion of autonomy that is useful and accepted by programmer and philosopher, alike. These fields share increasingly the same arena, and this sharing occurs as artificial agents become increasingly free from direct human intervention, an independence that is necessarily associated with the recognition of aspects of self-governance, and this ultimately leads us to question the idea of autonomy as established in the rationalist tradition just presented.

But how to do that? Our final answer will be to highlight emergent aspects of autonomy in opposition to the view of autonomy as a characteristic mark unique to human beings. Autonomy seems to be a much more fluid notion, one that implies degrees associated with different beings that populate an expanded – more inclusive – ethical space. In the end, we will establish a conception of autonomy that is a naturalist and evolutionary, in opposition to the aprioristic grounding of the notion staked out by the received rationalist tradition.

Accordingly, 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, as the ones that can be found in Allen (2010), but here we are going to describe models that rely on logical programming. Although limited, and low on the autonomy scale, we propose that these programming models highlight aspects of autonomy that are informative of our own embodied moral conditions, due to these very limitations. Thusly, in addition, we wish to emphasize both for 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. Besides 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, the present effort attempts to dismiss the apparent risks of a comparative decrease in our own autonomy when contrasted with the autonomy of those lesser autonomous agents and, therefore, uphold the stance that those apparent risks are weak. The gains that can be achieved through a better comprehension of ourselves outweigh those risks, these 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 intentions, 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 & Han, 2014; Pereira, 2012, Pereira & Saptawijaya 2011; Han & 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 & Pereira, 2006; Pereira & Lopes, 2007; Lopes & Pereira, 2010).

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 are going to go beyond this first immediate aim and reassess these models, after describing them. As we try to show, autonomy here touches something else that is valued from the point of view of Ethics, though in a manner deeply different from the kind of rationalist effort we described in the previous section.

In the first experiment or, more precisely, in the first sets of experiments, the agents therein 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 wish to present provides a chance to evaluate autonomy during social interactions or under social constraints, where autonomy plays a wider role even when dealing with simple agents.

The ground level of autonomy?

Our first programming effort models the reiterated prisoner dilemmas with various aspects of uncertainty taken into account, including when there is no full information about actions, in order to investigate the emergence of strategies of cooperation. The mechanisms driving the emergence and evolution of cooperation – in populations of abstract individuals with diverse behavioral strategies in co-presence – have been an object of mathematical study via

Evolutionary Game Theory (EGT), informed in part by Evolutionary Psychology. Programming efforts in this area have been ongoing at least since Axelrod (1984) offered the classical original formulation, with other models having been presented by authors like Danielson (1992). Their work depends on implementation and simulation techniques on parallel-processing computers, thus enabling the large-scale, fine-grained and yet relatively rapid study of aforesaid mechanisms under varieties of conditions, parameters, and alternative virtual games. The theoretical and experimental results coming from this field, thus, have continually been surprising, rewarding and – especially important for us – potentially informative for an empirically founded Ethics fully sensitive to the requirements of moral autonomy.

Recently, in our own work we have simulated groups of individuals with innate cognitive abilities represented by established AI models, namely those pertaining to Intention Recognition (Han, Pereira & Santos, 2011; Han, Pereira & Santos, 2012; Pereira, 2012, Han & Pereira 2013; Pereira, Han & Santos, 2014). This framework facilitates the modeling of agent tolerance or intolerance to errors in other agents – deliberate or not – and tolerance/intolerance to possible communication noise. As a result, our work has shown that both the emergence and stability of cooperation are reinforced in the presence of such cognitive abilities, of tolerance to error.

But we are getting ahead of ourselves. How is Intention Recognition inserted into computational agents? In the EGT approach, the most successful strategies become more frequent in the population. Kinship, neighborhood relationships, and individual differences, may or may not be considered. In indirect reciprocity (Nowak & Sigmund, 2005), players interact at most once, but they have knowledge of their partners' past behavior. This introduces the concern with reputation, and with moral judgment (Pacheco, Santos & Chalub, 2006; Pereira & Saptawijaya, 2011; Han, Saptawijaya & Pereira, 2012).

Our other recent work (Han, Pereira, Santos & Lenaerts, 2013; Han, Pereira & Lenaerts, 2015) has shown that, after the evaluation of interactions between third parties, strategies are adopted which allow for the emergence of kinds of cooperation that are immune to exploitation, because these interactions are channeled just to those who cooperate. This – to return to the figure of the shining Kantian goodwill passing through a prism of duty – replaces the dislocated prism of rationalist constructions with an empirical equivalent. Likewise, questions of justice and trust, with their negative (punishment) and positive (help) incentives, are fundamental to games with large and diversified groups of individuals gifted with intention recognition capabilities.

In our work, intention recognition is implemented using Bayesian Networks (BN) taking into account the information of current signals for intent, as well as trust and tolerance amassed during previous iterations. We experimented with populations with different proportions of diverse strategies in order to calculate, in particular, what is the minimum fraction of individuals with Intention Recognition for cooperation to emerge, invade, prevail, and persist in agents that self-regulate their operation. One hope in understanding these capabilities is that they may be transformed into mechanisms for the spontaneous organization and control of swarms of autonomous robotic agents. With this objective, we have studied how players' strategies adapt in populations during cooperation games. We have used the techniques of EGT and have considered games such as the "Prisoner's Dilemma" and "Stag Hunt" successively repeated, and have showed how actors participating in repeated iterations within these games can benefit from having the ability to recognize the intentions of other actors, leading to an increase in cooperation.

Declarative rules as a basis for moral reasoning

Without expecting any kind of magic common to fairy tales, the narrative developed in Lopes & Pereira (2010) and Saptawijaya & Pereira (2014) provides a parallel account of the emergence of autonomy over the course of normal life. The games created in order to develop this approach are similar in form to stories that we use in order to educate, or at least to entertain, children. For example, in our work, we "conjure" (program) a robotic knight and a princess in distress, kept in a tower by a powerful wizard, and make it his duty to figure out how to save her. This wizard, knowing that the access to the castle is only possible through two bridges, positions at each a defensive guard that can be, in different simulations, either a giant spider or a human warrior with different but measurable strengths. Different capabilities to act are taken in consideration by the robot-knight. And, differently from most fairy tales, the princess and hero in our stories can be sensitive to moral demands. For example, the model can be set up in such a way that the princess is sensitive to murder, and she can reject her "savior" if he chooses a course of action against her values, i.e. killing the human warrior to get to her, for example.

The different cases that are simulated in our work show, even if in a simple format, the interplay between (1) a set of preferences that are generated a priori, and that we can associate with that part of our moral reasoning described by deontological theories, and (2) a set of choices and preferences (to deal either with a spider or with a human guard, with a weaker guard or a stronger one). Further, these choices can be interpreted as relating to the role of imagination in us, how we can visualize consequences and determine actions according to the expectations created by that imagination, in coordination with the perceived situation.

In our work, this interpretation is facilitated by the set of abductive rules that drive the simulated agents. The key to their effective reasoning – as limited as it is by the few considerations available to our virtual agents – lies in what should be taken into account. This is exactly what we are going to cover next, further detailing the model of an interactive princess-saving storytelling, and showing how knowledge updates are employed for moral updating.

Apart from dealing with incomplete information, knowledge updates (as realized by EVOLP/R in Saptawijaya & Pereira, 2014) are essential to account for moral updating and evolution. They concern the adoption of new (possibly overriding) moral rules in additional to those that an agent currently follows. Such adoption is often necessary when the moral rules that one is currently following have to be revised in the light of situations faced by the agent, e.g., when an authority contextually imposes other moral rules. This is not only relevant in a real world setting, but also in imaginary ones. For example, in our work, the robot in the story must save the princess in distress while pursuing two possibly conflicting aims, enacting the princess's moral rules while ensuring its own survival.

We represent this capacity for revision through Prospective Logic Programming (PLP) (as refined in (Pereira & Saptawijaya, 2011). This work employs declarative non-monotonic reasoning, demonstrating that it is possible to build an integrated architecture for embedding these reasoning techniques in the simulation of embodied agents in virtual three-dimensional worlds. Further, a concrete graphics supported application prototype was engineered in order to enact the story of the princess saved by the robot and imbued with moral reasoning. Our work with PLP supports the view that autonomous agents are those capable of anticipating and reasoning about hypothetical future scenarios. This capability for prediction is essential for proactive agents working with partial information in dynamically changing environments.

In order to illustrate the basic PLP framework constituting the basis of our ACORDA framework (Lopes & Pereira, 2006 and Pereira & Lopes, 2007), consider that the robot is asked to save the princess in distress, and then he is confronted with an ordeal. A gap, a river, crossable by two bridges, blocks the path to the castle. Standing guard at each of the bridges are minions of the evil wizard. In order to rescue the princess, he will have to defeat one of the minions to proceed, and overcome the river's gap.

As the PLP robot reasons, a balloon displays its thoughts, in real time[1]. Prospective reasoning involves the combination of hypothetical scenario generation – into the future – followed by preference assignments taking into account the imagined consequences of each proffered scenario. By reasoning backwards from the goal to save the princess, the agent (i.e., the robot) generates three possible routes for action. Either it does not cross the river at all, thus negating satisfaction of the rescue goal, or it crosses a bridge. In order to derive the consequences for each scenario, the agent has to reason forward from each available hypothesis. The goal of the robot, i.e., save(princess, after(gap)), is further tempered by an integrity constraint (ic0) according to which the robot will prefer the scenario with likelihood of survival that does not fall below a preset specified ic0 threshold. The robot's self-regulatory mechanism evaluates the field in order to determine which is the best way to achieve its aim (save the princess) whilst suffering the smallest loss to it (something one could relate to a minimization of pain).

Certainly the moral issues presented here are admittedly limited, but in some sense do illustrate a rudimentary sort of reflective equilibrium over possible ends (in the sense that the robot works back and forth between a series of options in order to establish a judgment). As soon as these consequences are known, meta-reasoning techniques are applied to weigh the partial scenarios. For example, if the goal of the robot is expressed as save(princess, after(gap)), which can be satisfied either by hypothetically abducing kill(ninja) or kill(spider), we can conjure a robot that exhibits something like a consequentialist procedure of moral reasoning. That is, the decision of the robot for choosing which minion to defeat (i.e., to kill), in order to save the princess, is purely driven by maximizing its own utility, predicated first of all on its survival.

Consider the following permutation on that case. The princess becomes angry because the robot

[1] It can be seen in this video: http://centria.di.fct.unl.pt/%7Elmp/publications/slides/padl10/quick_moral_robot.avi

decides to kill a man (the ninja) in order to save her since the robot had identified this as maximizing utility (its own survival) and without any concern about the princess' moral rules. In other words, the spider is easier to kill than is the human guard. She then asks the robot to adopt moral rules, namely that no man should be harmed in saving her, a constraint that we tag "ghandi_moral".

The robot learns about gandhi_moral interactively, by being told. We represent this communication by the literal update, "knows_about(gandhi_moral)". This demand changes the permissibility of the path previously chosen (the one that maximizes personal utility). Consequently, now the killing of the human guard is no longer triggered even if it is the choice that would be considered best from a purely consequentialist point of view.

The gandhi_moral update allows consistent ends to be abduced, as a rule to follow even in the face of conflicts between personal survival and the princess's sensitivities. Since the robot's knowledge contains the new overriding impediment, killing the overwhelming spider becomes the only path to be followed if save(princess, after(gap)) is to be achieved. The goal to save(princess, after(gap)) implies that the princess has to be saved whatever it takes. However, there are limitations imposed on this goal. For example, the knight wants, above all, to survive, but now is forbidden from killing the guard, thus jeopardizing his own survival by facing the undefeatable spider, and gives up saving the princess. How are such conflicts to be resolved between efficient means and ends that threaten survival?

A moral update could take place. The literal knight_moral represents a further constraint on means, demanding moral conduct in forcefully achieving the goal to save the princess. If knight_moral is not yet imposed, then the survival integrity constraint does not necessitate save(princess, after(gap)) to be an active goal that must be achieved. An agent without knight_moral considers killing the spider an "unreasonable rescue" – kill(spider) satisfies unreasonable_rescue(princess, after(gap). The consequence in this case is that the robot decides not to kill any of its enemies – the ninja because the means do not justify the end of saving the princess, and the spider because of his sense of self-preservation. It just simply aborts its mission to save the princess.

But we can consider a case too where this happens differently. So, in our third plot, the princess justifiably becomes angry again, this time because she is not saved. She then imposes the 'knight_moral' conduct rule which leads to an update in the form of moral rules that override previous moral rules conflicting with the new rules.

By the integrity constraint, solving conflict is once more triggered as a goal, making both abducibles – kill(ninja) and kill(spider) – available as solutions to the conflict, again. Next, having been told about knight_moral expressed by its update rules, the knight adopts the knight_moral and this results in abducing follow(knight_moral). Recall that gandhi_moral had already been adopted, and this leaves kill(spider) as the only abducible compatible with the new knight_moral. That is, the a posteriori preference chooses the scenario with both gandhi_moral and knight_moral to be followed. Thus, the robot has only one way to save the princess: kill(spider). This option respects both gandhi_moral and knight_moral, adopted before and still in force. As a result, the robot fails to save the princess. Indeed, the robot's survival requirement is now lower than the survival threshold with respect to the spider, and thus the spider kills it. This highlights an important issue in moral philosophy about our dependence on luck (Williams, 1981) and the eternal prospect of failing that surrounds any good will, irrespective of its brilliance.

These simple scenarios already illustrate the interplay between different logic programming techniques and demonstrates the advantages gained by combining their distinct strengths. Namely, the integration of top-down, bottom-up, hypothetical, moral updating and utility-based reasoning procedures result in a flexible framework for dynamic agent specification. The open nature of the framework embraces the possibility of expanding its use to yet other useful models of cognition such as counterfactual reasoning and theories of mind. Certainly, we can debate the value and the extension of the autonomy in this case. This debate can be more or less fruitful. The rejection of autonomy in this case, by an a priori formulation based on the fact that the very reactions of these agents do not denote freedom, falls in this last category and, standing on the argument expounded in the previous section, about the innumerous reasons to reject this procedure, we aim now to set aside this formulation in order for a more fruitful approach to arise in its place.

3- How do these developments influence our own understanding of what it means to be autonomous, and how analysis can help with the task of establishing an acceptable use of autonomy applicable to artificial agents, but nevertheless within a continuum of the general use of the concept in Ethics

In our challenge to the dominant tradition on autonomy, we need to deal exactly with its strong association with freedom and of free will. We will delimit autonomy as a form of self-legislature detached from our senses, and by implication non-contextual. It can be otherwise developed, and we are going to uphold that it should, by having other priorities in mind. Our models and the different autonomous artificial agents that appeared in the past few years provide some of the reasons for this disengagement but we should point out, too, that there exist a diversity of philosophical reasons to adopt other stances.

A first one is due to the Cartesian and Kantian tradition as reviewed previously, because it is in itself the cause of a common reaction well spread among scientists and naturalist philosophers, and summarized by Margaret Boden when she affirms that “autonomy is a problematic concept partly because it can seem to be close to magic, or anyway to paradox” (Boden, 2008). This "magic", brought by the relation between autonomy and freedom, is mainly derived from the attempt of this tradition to place human beings in a somewhat special position – i.e. as those beings that have free will and the duty of self-legislation because of their rationality and, consequently, the only members of the natural kingdom to be included in the kingdom of morality.

And, even in the face of skepticism, there certainly are as many good reasons to follow this path, if no other than for the aforementioned dignity that it apparently guarantees. But, then again, on the other hand there are reasons, which authors as diverse as Anscombe and Dennett deliver, to denounce this idea as not only opposed by our best theories of ourselves as natural beings that evolved in a certain context, but also grounded in a conception of morality awarded to us by some supernatural entity.

This first reason should on its own provide enough motivation to follow other paths, and there are different ways of diverging from this dominant tradition. For example, that the imprecisions around the cluster ‘autonomy-freedom’ itself do not help it to fulfill autonomy's centrality in Ethics, and that, indeed, may undermine the “will” to adopt this concept as having such an important role in Ethics. Nomy Arpaly, for example, develops her theory of moral worth by basing it on the more accessible notions of praise and blameworthiness, suggesting leaving autonomy out of the picture altogether. This is because, she suggests, there are “at least eight distinct things” being called “autonomy” and, although she assesses those extensively (Arpaly, 2004, p.118-126), she thinks that the mere existence of such complex and polyphonic use should warn us against attempting to define it in a single useful version, acceptable to all of those involved in the discussion.

Accepting these warnings, we are not going to try to clear the ambiguity around the notion of autonomy. Instead, we are going to agree with Boden's diagnosis that the problem lies in the denaturalization of the concept caused by the connection between autonomy and freedom. We take the admixture of magic and paradox that surrounds the concept as incentive towards a different formulation, one that, at the end of the day, is similar both to Boden's and to Arpaly's formulations.

We share Boden’s criticism of autonomy, and in particular reject autonomy:

(1) as an absolute value only susceptible of two states (either present or absent);

(2) through a simplistic top-down (disembodied) conceptualization, which rejects what seems central to our own experience of it as something always in relation, and that comes to existence in systems, like the one in our first model, that respond to the demands felt by the interplay between senses and reasoning in some context or other (that we control and understand in different degrees and within imposed constraints); and

(3) insensitive to the emergence of new agents that have inaugurated the field of artificial morality.

Ethics and moral philosophy can be enriched if the concept of autonomy can be developed in such a way that would warrant it to be applied to a wider range of behaviors and agents. So, the question we first posed mutates into an investigation into how to separate the unnecessary baggage due to received traditional wisdom from a more inclusive concept of autonomy. We will focus on freedom, and try to dissociate freedom from autonomy, thereby freeing autonomous agents from this baggage.

Gerald Dworkin, after a brief historical account reestablishing autonomy in the sense we have adopted, i.e. as that referring to the property or ability to act according to reasons and motives taken as one's own, suggests a manner we believe diminishes the magic and paradox that Boden much criticizes. His strategy is to distinguish between something relative to specific acts – freedom – and autonomy as that global capacity that certainly is part of freedom but which is more generic. In his own words:

> Putting the various pieces together, autonomy is conceived of [as] a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or attempt to change these in light of higher-order preferences and values (Dworkin, 1988, p.20).

There are two aspects here, one more useful for us and another that, although interesting, is not well suited to our aims in the field of artificial morality. The usefulness of this characterization is that it enables us to understand autonomy as relative to series of events that seemingly constitute a flow, and are intrinsic to agents as a form of reasoning that can be described as reflective self-regulation over the aspects and transformations accessible to agents in their contexts.

Interpreting Dworkin's "second-order capacity to reflect critically upon their first-order preferences, desires, wishes and so forth" as this very reasoning procedure, we sustain that typical cases of this ability are conscious, and are accessed through the production of beliefs in the process of reasoning. This reframing reflects a more general change in the field of Ethics itself. Scanlon, for example, in his most recent book (see the first lecture in Scanlon, 2014), points this out when he effectively comments on the transition of the focus of Ethics, from the establishment of what is right and wrong to that of a broader study of normative life in a bigger picture wherein the focus of the study of Ethics tends to a psychologizing.

The second aspect of his characterization links autonomy to persons. Dworkin is interested in this link because of the goal to avail himself of a point of view that permits him to judge experiences as more or less worthy. However, his aim is to analyze something already granted within what we call the Autonomous Entities Association (AEA). Our focus remains on the first aspect, as it is through psychology that we are able to establish the prerequisites for membership in this association.

Qualifications for inclusion in the AEA must be reevaluated in light of historic developments whereby, for the first time ever, the aim of providing a natural account of our relation with the world seems finally aligned with the developments in artificial agency. In the context of this alignment, we may suggest a clarification about prerequisites for membership in the AEA. We base this effort on Boden's, in dismissing autonomy's “magical” character (Boden, 2008). Membership depends on three things:

1- “The extent to which response to the environment is direct (determined only by the present state of the external world) or indirect (mediated by inner mechanisms partly dependent on the creature's previous history)”;

2- "[T]he extent to which the controlling mechanisms were self-generated rather than externally imposed," an aspect that is certainly impacted by the developments in artificial intelligence that were presented above;

3- "The extent to which any inner directing mechanisms can be rejected upon, and/or selectively modified in the light of general interests or the particularities of the current problem in the environmental context".

In addition to this formulation, she provides this concisely:

> “In general, an individual's autonomy is the greater, the more its behavior is directed by self-generated (and idiosyncratic) inner mechanisms, nicely responsive to the specific problem-situation yet reflexively modifiable by wider concerns” (Boden, 2008).

This formulation allows a visualization of how levels of autonomy may be distinguished, as a series of different candidates for membership in the AEA. We suggest the figure of a mountain, where those objects and artifacts that are not autonomous (naturally or artificially) occupy the very bottom. Certainly some of these not-autonomous-yet things seem to challenge one or another of these three proposed axes: self-organized phenomena and even some self-replicating processes seem to blur the lines at the start of the ascent, sharing these low-elevations with some or other specifically built artifacts, or even with some forms of life. But the real climb is composed by different autonomous beings that seem to move up and down its slope, climbing it as their capacities of processing reality in their limited nervous system develop and aided by different artifacts and even – recently – artificial agents. In addition, we can identify some conversion between those as in the case of *e. elegans* and artificial models as the recent project to simulate its capacities in a 1:1 scale, so in a sense there is integration between these domains.

As artificial agents are developed farther up, along this hill, and we move further up this hill with them, we see beings that increasingly motivate a distinction of ‘quality’ in the interaction between these axes, in a fashion similar to that of the utilitarian tradition in the measure of pleasure. We can imagine such a quality distinction being inserted into our courageous robot, also.

The very top of this mountain image is certainly not occupied by human beings. Our more recent accounts of ourselves certainly support that we are not in the business of pure autonomy, as assumed by the Kantian tradition, and for at least two reasons. One, we are not built to be purely rational, and we certainly value things that contradict its demands. The reasons for that are probably related with how contextualized our reactions and our decisions are in opposition to the Cartesian formulation that detached our souls from our senses. Thus, the top of this mountain is surely inhabited only by imaginary beings or supernatural ones, but this should not diminish the importance of this top as somewhat of an ideal or regulatory frame. (Recall Aristotle's distinction, in the *Politics*, between the worst of animals and a god – at the summit, live only gods.) In this sense, the establishment of this frame provides us with a horizon within which different candidates for the AEA can be classified, but we will now leave the work in determining the line delimiting “real” autonomy for others.

Instead, other important questions emerge from this reflection and we want to discuss two of them. The first is related with the possibility of a quantitative unit for the measuring of autonomy, and the second is related with the phenomenology of autonomy as an emergent property of certain agents.

Can we develop a unit of measure of autonomy?

There are in the literature some suggestions about the quantification of the autonomy of different systems, which we want to consider, even if briefly, by focusing on the formulation of Seth (2007 & 2010). Seth's formulation of autonomy is strictly analogous to our own, sharing with us even its source in Boden’s definition. This leads to a series of delimiting borders – “An autonomous system should not be fully determined by its environment, and a random system should not have a high autonomy value” – that helps to map the base which the mountain of autonomy rises, a visualization we sustain here.

Seth (2007) proposes a measure of autonomy based on the following conception of autonomy. First, future outcomes for an agent can be better accessed by considering its own past states, as compared to predictions based on past states of external variables. This notion establishes what he calls a G-autonomy variable: "a variable is G-autonomous to the extent that (1) it is dependent on its own history, and (2) these dependencies are not accounted for by external factors" (Seth, 2007, p.475)

G-autonomy, Seth applies to different forms of artificial life, and we can apply it to our models given both their limited autonomy (since their history is still incipient) and further developments, both in self-control and in richer environments wherein agent adaptivity and response can be more widely demonstrated.

The phenomenology of autonomy as an emergent property

An important question persists. This concerns an aspect of AEA members that Dworkin dismisses with the term “person”. We have pointed out that the comprehension of autonomy, as established here, rejects the view that takes it as a definitional matter, one whose presence or absence can be easily attached to agents as a whole, or otherwise to certain actions. Autonomy emerges, and even then changes. However, what emerges as autonomy? The answer is not some “thing” but rather something like a form, or pattern, or function. The concept of emergence applies to phenomena in which relational properties dominate over constituent properties in determining aggregate features. It is to configurations and topologies, not specific properties of constituents, that we trace processes of emergence.

By analogy with computing machines, cognitive scientists have argued that the “functional” properties that define a given cognitive operation are like the logical architecture of a computer program. Philosophically, this general form of argument is known as functionalism, and it is quite relevant for viewing autonomy as an emergent property. In fact, it brings us to an important aspect of our preconception of autonomy. As seen above, we departed from the view whereby autonomy is established as that property or ability to act according to reasons and motives that are taken as one's own, but we have left untouched as yet one important part of this preconception, i.e., the part relative to the expression “as one’s own”.

In moral philosophy, this expression refers exactly to the relation between autonomy and freedom. “One’s own” delimits the native power of a being in this world to deliberate and choose freely its course of actions. "One's own" may be extended, into a broader conception like that of Rawls for example, whereby this being is able to establish its character having in view a whole-life plan.

Nevertheless, with freedom set aside, following the concerns that we shared with Boden, we need to provide a conception of those beings where autonomy emerges in a different frame, one that highlights not the aspect of free will and choice but of self-regulation of interactions with the world. In this matter, we suggest "one's own" as naming the capacity to execute the cycle of observe-think-act, a view similar to that of Kowalski (2011), of agents as logical programs that exactly perform this cycle. In his conception, “the thinking or deliberative component consists in explaining the observations, generating actions in response to observations, and planning to achieve its goals”, and the acting component is understood as “a proof-procedure which exploits integrity constraints”. Since our aim here is exactly to interpret logical programs performing as agents, this definition fits the place occupied by Dworkin's “person”, and of moral philosophy's “one’s own".

Our conception of autonomy is not something that has an a priori decidability, be it as an a priori definition, be it as a predicate attached to agents in an absolute way, or be it something context independent. The notion now being presented, and here we are in agreement with some philosophical formulations like that of Arpaly (2004), recognizes autonomy as an emergent property of certain agents, able to adapt when exposed to environments, the latter being evaluated considering the ways available to the agents to enact these adaptations (i.e. if they are self-regulative or not, and if they are able or not to evolve in answer to their mutable environments). Our concept is established as referring to the property or ability for agents to act according to reasons and motives that are taken as one's own (the agent's) and, moreover, it is a capacity always in relation with something.

This permits us to present some of the attempts at developing artificial autonomous agents and to interpret those agents in a spectrum within those (Boden's) three axes, (1) the responsiveness to the environment, (2) self-generation of the mechanisms that permit this responsiveness, and (3) openness to self-modification in this process, what we can dub, for short, evolution.

This formulation does not close the door to improvements. One that can be imagined is the insertion of another axis, representing the aptitude of agents for producing explanatory or justificatory reasons of their own, in a way that would highlight autonomy as having a social-communicative dimension. Communication and responsiveness to reasons is highlighted by thinkers from Jonathan Dancy to Dan Sperber as an important characteristic of moral reasoning, and is a characteristic mark belonging to few animals besides human beings (c.f. Cheney & Seyfarth, 2007). And, this leads us to some interesting philosophical considerations.

4- **Artificiality in the animal kingdom of morality**

In the previous sections, we invited the reader to share with us the understanding that the aim of providing an account of our ourselves as ethical agents is aligned with the developments in artificiality, exemplified by the emergence of new forms of agents, the artificial moral agents. This emergence populates the world with beings clearly distinct from the most common forms of non-reactive artifacts produced by human beings in their cultural history. Although a moral Turing test has been proposed before (Allen et al., 2000 and Arkin, 2009), the perspective adopted in this chapter can contribute to an account of the consequences of this emergence in a distinct way. Certainly, the models presented here, in particular the second one, seems to aim at the simulation of moral reasoning as a declarative process where justification plays a central role, something that is open to criticism. Sometimes, it seems that the offering of this reasoning is mostly an afterthought of the agents, something quite extrinsic to the process involved in what we consider the focus of Ethics: the progress from deliberation to actions, where explanation occupies certainly its place in the aftermath.

Surely, the heuristic process involved can be investigated and set out as have Haidt (2007), Greene et al (2004) and Mikhail (2011), among others, but a more complete comprehension of the emergence of these agents needs to deal with their responsiveness through action itself. Galen Strawson suggests the following example:

> Suppose you set off for a shop on the evening of a national holiday, intending to buy a cake with your last ten pound note. Everything is closing down. There is one cake left; it costs ten pounds. On the steps of the shop someone is shaking an Oxfam tin. You stop, and it seems completely clear to you that it is entirely up to you what you do next. That is, it seems clear to you that you are truly, radically free to choose, in such a way that you will be ultimately responsible for whatever you do choose. You can put the money in the tin, or go in and buy the cake, or just walk away. (You are not only completely free to choose. You are not free not to choose.) (Strawson ,1994, p.11)

In a context such as ours, where artificial autonomous agents are expected to become ubiquitous, this question is about our next step. Whether the necessary openness and reactivity is possible in programmed systems, or if there is a final gap between coded autonomous agents and us. As we highlighted before, in our computer modeling, the first aim is not empirical evidence and lacks the brute force of it. We are in a position to wonder: Do we continue to put money, time and effort into this "tin", or take another route?

The next step in our research efforts are commitments to the former. We aim to model the ability of our agents to deceive, imitate and emulate, in order to further enlarge the gray zone in our graphic proposal. With some functions encoded, we anticipate resolving those key elements necessary for membership in the AEA. We are confronted by a gap, with fuzzy edges that cannot be dispensed with by appeals to personhood, because personhood belongs to either side of the gap, at the summit of our climb to fully articulated artificial moral agency. Our conviction is that future work further integrating moral philosophy with programming will establish necessary logical supports to complete the task. And our hope is that this work will provide evidence that, although the processes within us are complex, their complexity is not inaccessible.

According to Newell (1992, p.25), in the face of big puzzles, it is useful to first know their dimensions. Modeling and simulation are tools for this discovery. Although the unity of measure that we analyze is still far from perfect, the point to develop with the case of automata is that a small set of rules, self-applied, may counter-intuitively generate a behavior close to those of more complex beings or at least their societies. These are dimensions that should prove useful as inquiry into the essence of moral agency continues.

With such dimensions identified, we are able to assess whether any distinction between moral behavior and emerging patterns like convection cells in boiling water, and so finally draw some formal equivalence between that "magic" of autonomy and freedom with the natural systems in which they emerge. In an evolutionary context, the point is that the brain, in its biological evolution, has bootstrapped itself into generality. From computer science we know that, as soon as one boots a computer, what the hardware does – simplifying a bit – is to go to the first instruction in memory. This instruction resides as a physical pattern, with the pattern due to "software". What does this mean? It means that the instruction configures the hardware, and obliges the hardware to execute that instruction. It means that the CPU then obtains instruction-specified data from memory, combines the data according to the software instruction, and puts the result back in a memory location specified by the instruction. Then the hardware looks up the next software instruction, and so on and so forth. That is, the software becomes the master of the hardware. Our contention is that in the brain that happens too, and is called "free will" by some people. Thus, the climb up the slope of autonomy is a process of bootstrapping (see Pereira, 2014, for detailed discussion).

Different agents may evolve through interactions with surroundings, and different agents may direct their evolution differently as autonomy increases along the three axes established earlier. Consider Floridi's (2007) a "re-ontologizing" of reality in this way. Evolution is ontologization. Looking at the different programming efforts ongoing, and in particular the two types that we presented, we can see an increase in agents' capacities of filtering different aspects of the data presented to them. They can be more and more sensitive to variations of strategy, as in the example of the reiterated cooperation games, and they can develop a series of moral standards in response to other agents' sensibilities. However, the roles that they play, and the worlds that they play in, are static in comparison with ours. Even as we increase the "noise" in information to which the agents are exposed, this is just a function to be controlled by those systems, and not something that constitutes the systems themselves. It is this sense we want to highlight. As much as we need efforts to develop agents in a way that increases their autonomy – no longer paradoxically we hope – then we need broader and more open universes where agents react and evolve, too.

In a certain sense, the two implementations we attempted to interpret seem, because of the way reasoning and context play out in them, to follow a correct step in the direction of a more extended and recognizable autonomy. This sense is one where their ability to provide reasons for certain procedures, their ability to play normative games, seems to close the gap in an important way, providing a better view on what ethics is for us. In ethics and in the theory of ethics, the challenge is to provide a picture of how agents can be able to deal with their situations, of how they can be able to deal with a reality that is not only demanding (this seems to be true for any form of life) but whose demands need to be fulfilled in a certain way, within a certain time. The reasons provided by those agents, when, for example, they commit errors – as in the act of killing another warrior because the agent erroneously supposes the promise of a "kingdom of ends" is equivalent to the execution of his will – are certainly presented in a form that is odd. It is difficult to identify his autonomy, even when understood in a framework that does not attempt to pinpoint it but to show how it is constituted as a dispersed continuum.

However, we should remember that the main function of our reason is the expression of our beliefs: to provide indications of external states of affairs, as much as to provide indications of the kinds of evaluators of these external states we ourselves are (like in the example of the moth of Dretske (1988, p.91s).

Recall the knight and princess. This first implementation provides not only a measure of good or bad partners in the game, but also a minimal logical framework articulating the evolution of others evolved in any situation similarly, i.e. socially, and where some players are recurrent in the life-span of the agent enabling trust, reputation and reciprocity. These seem to be our own, and this fact helps us to defend a project that is otherwise somewhat hard to support. This project is intuitively irrational: that of the sandboxed comprehension of our own moral lives, of our ideals and of our abilities to plan for futures which are in every case not guaranteed, since all of this – as the sandbox, itself – belongs to an open world.

That the limited and simple agents that these models permit us to play with, do build, almost as strong as necessity, something involving these aspects in the blind interplay of their games, provides an argument to reject the twine of those positions that regard this project as unauthentic (i.e. a mere fiction to support some more basic realm like the one of power). Such criticisms are not far from being rejected on the same basis as those conceptions that anchor morality in an apparent supernatural being, supposing that only such great power could support it. Some people attempt to sustain morality on a magical something (freedom) that hardly fits in our worldview of natural beings that have evolved. Our openness to a plethora of variables wider than those, should rather be considered a reason for the keeping of our, but only at the start, apparently jeopardized dignity.

Conclusion

In a nutshell, we propose an interpretation of two computational efforts for telling 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, whilst losing, in the process of that very same identification, a perspective that could have permitted a multitude of agents, with different degrees (and attending constraints) of autonomy, thereby providing a richer account of autonomy. We aimed to provide this critical conception of autonomy in a way that helped avoid forgetting the more humble starts of those values, our values. However, this result is collateral with regard to our main aims, since it was not our intention to provide here a lesson in the humbleness of what we value and its precariousness in our world. If so wished, this can be attained within a philosophical discussion, for example, in the concluding remarks of Bernard Williams's essay on moral luck (Williams, 1981), where the author highlights the circumstantial character of ethics and, therefore, the fact that these cannot be that higher ground inaccessible to agents – Kant's "bright jewel" – even human ones. Only a better understanding of how we get to value such things could achieve this.

With this in mind, our aims have been to provide this better comprehension. Our goal was to contribute to a better understanding of autonomy and, following up on the repercussions of our first example, how autonomy can be interpreted as having evolved, by tracing its evolution in the field of artificial agents. Certainly, models are just models, but the twin tasks of our inquiry, of developing more legitimate moral agents and a better comprehension of ourselves, remain open. We believe that our work is carried out at the nexus of these two tasks, and thus more efficiently responds to these two challenges than does the common a priori evaluation. In fact, our position is that better models reach better theories, once they permit the elimination of some of the limitations of armchair exercises, which, when modeled, reveal counterintuitive results and unexpected difficulties. We sustain that models are theory laden and biased, in the sense that the limitations concerning what we think should be the case, do actually regulate – recognizing here a dimension of heteronomy – what we eventually get as results. This supports the necessity of a tight co-evolution of modeling and theory.

Margaret Boden (especially in Boden, 1998) established a similar frame for the comprehension of autonomy in the field of artificial life. Her concerns were to dismiss the notion that in a deterministic world (one like those simulated in our models, like the ones she describes or, following some results of science, one like our own), the emergence of supernatural agents came, mainly, to confirm our delusional character, with the correlated implication of a denial of our freedom. Our aim has been narrower, showing that we have computational tools that do not depend only on our intuitions, to investigate the concept of autonomy and Ethics more generally, thereby offering a chance at updating our own moral codes and meta-reasoning over those codes. And thus, our inquiry parallels hers in that it aids in identifying the shapes kicking up the cloud that has covered over the concept of autonomy, at once clearing errors due to delusions of untutored intuition. If this point is received intact, it is now clear that only with wider and better models can we achieve this result. After all, freedom follows understanding, and it too can have evolved (Dennett, 2004).

Acknowledgement

Fernando da Costa Cardoso acknowledges the support of Conselho Nacional de Desenvolvimento cientifico e tecnológico (CNPQ/Brazil)

References

- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), pp. 251-261.
- Anscombe, G. E. (1958). “Modern moral philosophy”. *Philosophy 33 (124)*, pp. 1-19.
- Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge, Cambridge University Press.
- Arkin, R. (2009). *Governing lethal behavior in autonomous robots*. Boca Raton, CRC Press.
- Arpaly, N. (2004). *Unprincipled virtue: An inquiry into moral agency*. Oxford, Oxford University Press.
- Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). “Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action”. *Adaptive Behavior*, *17*(5), pp. 367-386.

- Boden, M. A. (1998). "Autonomy and artificiality". *Cognitive Architectures in Artificial Intelligence: The Evolution of Research Programs*, *2*, pp. 300-312.
- Boden, M. A. (2008). "Autonomy: What is it?" *Biosystems*, 91(2), pp. 305-308.
- Cheney, D. & Seyfarth, R. (2007). *Baboon Metaphysics*. Chicago, Chicago University Press.
- Dancy, J. (2004). *Ethics without principles*. Oxford, Oxford University Press.
- Danielson, P. (2010). "Designing a machine to learn about the ethics of robotics: the N-reasons platform". *Ethics and information technology*, *12*(3), pp. 251-261.
- Danielson, P. (1992). *Artificial morality: virtuous robots for virtual games*. London, Routledge.
- Darwall, S., Gibbard, A., & Railton, P. (1997). *Moral discourse and practice.* New York, Oxford University Press.
- Dell'Acqua, P., Mattias Engberg, & Pereira, L. M. (2003). "An Architecture for a Rational Reactive Agent" in: Moura-Pires, F., & Abreu, S. (eds.), *Progress in Artificial Intelligence*, Procs. 11th Portuguese Intl.Conf. on Artificial Intelligence (EPIA'03), p. 379-393, Springer, LNAI, Beja, Portugal.
- Dell'Acqua, P., & Pereira, L. M. (2004). "Common-sense reasoning as proto-scientific agent activity", *Journal of Applied Logic,* 2(4): pp. 385-407.
- Dennett, D. C. (2004). *Freedom evolves*. London, Penguin UK.
- Doris, J. (2010). *The moral pshychology handbook*. Oxford, Oxford University Press.
- Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT press.
- Dworkin, G. (1988). *The Theory and Practice of Autonomy*, New York, Cambridge University Press.
- Floridi, L. (2007). "A look into the future impact of ICT on our lives". *The information society*, *23*(1), pp. 59-64.
- Lopes, G. & Pereira, L. M. (2006). Prospective Programming with ACORDA, in: Empirically Successful Computerized Reasoning (ESCoR'06) workshop at The 3rd International Joint Conference on Automated Reasoning (IJCAR'06), Seattle, USA.
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), pp. 389-400.
- Guyer, P. (2003). "Kant on the theory and practice of autonomy*"*. *Social philosophy and policy*, 20(02), pp. 70-98.
- Haidt, J. (2007). "The new synthesis in moral psychology". *Science*, *316*(5827), pp. 998-1002.
- Han, T. A., & Pereira, L. M. (2013). "Intention-based Decision Making via Intention Recognition and its Applications", in: Guesgen, H., 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. (2011). "Intention Recognition Promotes The Emergence of Cooperation". *Adaptive Behavior*, 19(3), pp. 264-279.
- Han, T. A., Pereira, L. M., & Santos, F. C. (2012). "Corpus-based Intention Recognition in Cooperation Dilemmas". *Artificial Life*, 18(4) pp. 365-383.
- Han, T. A., Pereira, L. M., Santos, F. C., & Lenaerts, T. (2013). "Good Agreements Make Good Friends". *Sci. Rep.*, 3, doi: 10.1038/srep02695.
- Han, T. A., Pereira, L. M., & Lenaerts, T. (2015). "Avoiding or Restricting Defectors in Public Goods Games?". *J. Royal Society Interface*, 12:2014 pp. 1203.

- Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. New York, Oxford University Press.
- Kant, I. (2002). *Groundwork for the Metaphysics of Morals*. New Haven, Yale University Press.
- Kowalski, R. (2011). *Computational logic and human thinking: how to be artificially intelligent*. Cambridge, Cambridge University Press.
- Lopes, G., & Pereira, L. M. (2010). Prospective storytelling agents. In M. Carro, & R. Peña (Eds.), *Proceedings of the Twelfth International Symposium on Practical Aspects of Declarative Languages (LNCS)* (Vol. 5937, pp. 294-296). Berlin: Springer-Verlag.
- Lin, P., Abney, K., & Bekey, G. A. (2011). *Robot ethics: the ethical and social implications of robotics*. Cambridge, MIT Press.
- Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge, Cambridge University Press.
- Nowak, M. A., & Sigmund, K. (2005). "Evolution of indirect reciprocity". *Nature*, *437*(7063), pp. 1291-1298.
- Pacheco, J. M., Santos, F. C., & Chalub, F. A. C. (2006). "Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity". *PLoS computational biology*, *2*(12), e178.
- Pereira, L. M. (2014). Can we not Copy the Human Brain in the Computer? In *"Brain.org"* (pp. 118-126). Lisbon: Fundação Calouste Gulbenkian.
- Pereira, L. M., Han, T. A., & Santos, F. C. (2014). "Complex Systems of Mindful Entities -- on Intention Recognition and Commitment" in: Magnani, L. 2014 (ed.), Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues, pp. 499-525. Berlin, Springer-Verlag.
- Pereira, L. M. (2012). Evolutionary Tolerance. In L. Magnani, & L. Ping (Eds.), *Philosophy and Cognitive Science—Western & Eastern Studies (SAPERE)* (Vol. 2, pp. 263-287). Berlin: Springer-Verlag.
- Pereira, L. M., & Saptawijaya, A. (2011). Modelling Morality with Prospective Logic. In M. Anderson and S. L. Anderson (Eds.), *Machine Ethics* (pp. 398-421). New York, NY: Cambridge University Press.
- Pereira, L. M., & 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, & Lopes, G. (2007). Prospective Logic Agents, in: J. M. Neves, M. F. Santos, & J. M. Machado (eds.), Progress in Artificial Intelligence, Procs. 13th Portuguese Intl. Conf. on Artificial Intelligence (EPIA'07), pp.73-86, Guimarães, Springer.
- Petersen, A. C. (2012). *Simulating nature: a philosophical study of computer-simulation uncertainties and their role in climate science and policy advice*. Boca Raton, CRC Press.
- Regan, T. (1986). *The Case for Animal Rights*. Berkeley, University of California Press.
- Rushton, J. P. (1975). "Generosity in children: Immediate and long-term effects of modeling, preaching, and moral judgment". *Journal of Personality and Social Psychology*, 31(3), pp. 459.
- Saptawijaya, A., & Pereira, L. M. (in press). The Potential of Logic Programming as a Computational Tool to Model Morality. In R. Trappl (Ed.), *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations* (Cognitive Technologies). Berlin: Springer-Verlag.

- Scanlon, T. M. (2014). *Being realistic about reasons*. Oxford, Oxford University Press.
- Schneewind, J. B. (1998). *The invention of autonomy: A history of modern moral philosophy*. Cambridge, Cambridge University Press.
- Seth, A. K. (2007). Measuring autonomy by multivariate autoregressive modelling. In *Proceedings of the 9th European conference on Advances in artificial life*, pp. 475-484. Berlin, Springer-Verlag.
- Seth, A. K. (2010). “Measuring autonomy and emergence via Granger causality”. *Artificial life,* 16(2), pp. 179-196.
- Singer, P. (1993). *Practical ethics*. Cambridge, Cambridge University Press.
- Simão, J., & Pereira, L. M. (2003). “Neuro-Psychological Social Theorizing and Simulation with the Computational Multi-Agent System Ethos”, Invited paper in: Procs. Congresso em Neurociências Cognitivas, Évora, Portugal, November 2003.
- Strawson, G. (1994). “The impossibility of moral responsibility”. *Philosophical Studies*, 75(1), pp. 5-24.
- Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford, Oxford University Press.
- Williams, B. (1981). *Moral luck: philosophical papers 1973-1980*. Cambridge, Cambridge University Press.

Key Terms and Definitions

Abduction: A reasoning method whereby one chooses from available hypotheses those that best explained the observed evidence, in a preferred sense.

Artificial Morality: The emerging field that aims to use programming language in order to either test ethical theories and aspects of the moral dimension or to embedded autonomous computer systems with moral aspects.

Autonomy: indicates the capacity of an agent to make un-coerced decision in a context. As defined in the present chapter it is an emerging property and has a three dimensional frame.

Computational Logic: An interdisciplinary field of enquiry that employs the techniques from symbolic logic to reason using practical computations, and typically achieved by means of computer supported automated tools.

Counterfactual: A concept that captures the process of reasoning about a past event that did not occur, namely what would/could/might have happened, had this alternative event occurred; or, conversely, to reason about a past event that did occur, but what if it had not.

Dual-Process Model: A model that explains how a moral judgment is driven by an interaction of two different psychological processes, namely the controlled process (whereby explicit moral principles are consciously applied via deliberative reasoning), and the automatic process (whereby moral judgments are intuition-based and mostly low-level, not entirely accessible to conscious reflection).

Evolutionary Game Theory: An application of game theory to systematically study the evolution of populations, typically by resorting to simulation techniques under a variety of conditions, parameters, and strategies.

Logic Programming: A programming paradigm based on formal logic that permits a declarative representation of a problem and reasoning about this representation, that reasoning being is driven by a specific semantics.