A Computational Model for Simulation of Moral Behavior

Fernanda M. Eliott and Carlos H. C. Ribeiro
Informatics, Technological Institute of Aeronautics,
Praça Marechal Eduardo Gomes, São José dos Campos, São Paulo, Brazil

Keywords: Biologically Inspired Architecture, Artificial Moral Machine, Reinforcement Learning.

Abstract: The extension of our integration to technologies brings about the possibility of inserting moral prototypes into artificial agents, no matter if they are going to interact with other artificial agents or biological creatures. We describe here MultiA, a computational model for simulating moral behavior derived from changes over a biologically inspired architecture. MultiA uses reinforcement learning techniques and is intended to produce selective cooperative behavior as a consequence of a biologically plausible model of morality inspired from a perusal of empathy. MultiA has its sensorial information translated into emotions and homeostatic variable values, which feed cognitive and learning systems. The moral behavior is expected to emerge from the artificial social emotion of sympathy and its associated feeling of empathy, based on an ability to internally emulate other agents internal states.

1 INTRODUCTION

How to design an autonomous artificial agent able to socially interact and deal with conflicting tasks that require emotional guidance to be solved? A computational agent that incorporates artificial emotional and moral intelligences can lead to ways of producing consensual action between artificial creatures or both biological and artificial ones. And then, if we succeed in developing an artificial moral agent (AMA), would it be more useful the guidance from an immoral or moral behavior? In this work, we intend to develop an architecture that rudimentarily mimics moral behavior (through a simulation of empathy) and test it in a testbed cooperative game(Wang et al., 2011). Our hypotheses are a) if a moral or immoral agent can work better within an artificial group, and b) if a hybrid agent that can trigger both moral and immoral behavior might autonomously activate more easily moral action policies with biological creatures, and immoral actions otherwise. An artificial agent able to simulate a moral behavior may be important in general social or domestic assignments, *e.g.* taking the role of monitoring highly dangerous criminals, people in quarantine or in other situations, where there are social dilemmas to deal with. Moreover, the artificial empathy from a moral system could be used as a resource in argumentation-based negotiation in multi-agent systems (MAS), likewise, it might be useful to improve the responses to general MAS issues stressed by (Wooldridge, 2009), such as how to bring up cooperation in societies of self-centered agents; how to recognize a conflict and then encounter an agreement; or, as highlighted by (Matignon et al., 2012), the challenges to coordinate the agents activities in order to cooperatively conquer goals. In (Damásio, 1994) there is an explanation of the crucial involvement of emotions during the process of intelligent decision making. Also, through the *Somatic Marker Hypothesis*, emotion participation in filtering data and awakening our attention to what matters the most is stressed. Putting forward the vital role of emotion and feelings in rational decisions, social emotions such as sympathy (and its associated feeling of empathy) are described as taking into account social interaction and homeostatic goals (Damásio, 2004). Just as emotions and feelings aid human being in taking fast and intelligent decision spending less time and reducing the computational burden (Damásio, 1994), the simulation of moral behavior (through the embodiment of social emotions and feelings) might produce relevant results to the computational process of decision making. In case an artificial computational agent somehow succeeds in estimating the state of another agent and acts in response to it, the outcome may be advantageous. Gadanho (Gadanho, 1999) proposed a bioinspired behavior-based architecture fed by basic robotic sensor data (obstacle proximity and light intensity) in the context of a single agent. The data is translated into sensations, feelings and basic

M. Eliott F. and H. C. Ribeiro C..
A Computational Model for Simulation of Moral Behavior.
DOI: 10.5220/0005139002820287
In *Proceedings of the International Conference on Neural Computation Theory and Applications* (NCTA-2014), pages 282-287
ISBN: 978-989-758-054-3

emotions, which represent homeostatic goals that the agent is supposed to learn to keep above a minimum level. Influenced by (Damásio, 1994), the bioinspired design of (Gadanho, 1999) includes a simulation of the Somatic Marker Hypothesis: it helps determining when learning should take place and if the behavior selection should be reevaluated as new sensations take place. The architecture was further improved in (Gadanho, 2002), and finally in (Gadanho and Custódio, 2002), (Gadanho, 2003) it received a cognitive system and the acronym ALEC (*Asynchronous Learning by Emotion and Cognition*). ALEC inherits the biological inspiration from the previous architecture plus the influence of Clarion (Sun, 1998), an architecture to model cognitive processes through a psychological perspective.

Our proposed model uses ALEC as a starting point, but changes it for simulating moral behavior in MAS tasks. It will be outfitted with the artificial feeling of empathy — the sensitivity to the situation of another agent — through a system responsible for simulating, to a minor degree, mirror-neurons (Damásio, 2004). The model is called MultiA since it was inspired by the ALEC architecture, and will be tested in the context of more than one agent. We expect that, by responding to the feeling of empathy, MultiA shall be capable of producing artificial moral behavior and choosing cooperative action policies.

1.1 Related Work

According to (Wallach and Allen, 2008) Artificial Moral Agents (AMAs) require the ability to ponder diverse options and perspectives to properly perform under the human moral setting. It is mentioned the expectation about AMAs not deforming the moral ecology and, although that could be a delusive hope, it would be relevant to add moral freedom to the design of AMAs, whether or not it is consistent with determinism (even there, ethic behavior would concede choices with an unpredictable ending). The apprehension about AMAs behaving negatively is also present in (Bringsjord et al., 2006), where it is regarded that deontic logic, thanks to its possibility of formalizing a moral code, allows the script of theories and dilemmas in a declarative mode. That could enable specialists to analyze and restrict behavior in ethically sensitive environments, adding matter to the debate about Lethal Autonomous Systems, as pointed out in reflections by (Arkin, 2013) and (Asaro, 2012). In (Bello and Bringsjord, 2012), there is a concern about including restrictive commands on the machine, and that those should be related to the moral human cognition. Also, the moral common sense is highlighted and presented in a modified model (the original is in (Bello et al., 2007)) of mindreading. From the results, it is concluded that AMAs need to have something that resembles a common moral sense to productively interact with humans.

Computational simulation of moral has also been considered. To exemplify, we mention three models. First, the LIDA Model (Wallach, 2010) (Wallach et al., 2010), influenced by the Global Workspace Theory (GWT) and by the *Pandemonium* Theory (Jackson, 1987) for the automation of action selection. An AMA under LIDA would be designed to be a practical solution to a practical problem: how to take into account the maximum possible ethically relevant information within the time available for selecting an action. The ETHEL Model (Anderson and Anderson, 2011), whose application field is related to *prima facie* duties, was implemented and tested within the notification context: an analysis of when, how often, and whether to run a notification about a medicine to a particular patient. Finally, in order to reflect about the Moral Theory vis-à-vis the conflict Generalism *versus* Particularism, Guarini (Guarini, 2006), (Guarini, 2012) draws insights from (Dancy, 2010): if the moral reasoning, including learning, could be done without the use of moral principles. If so, models of artificial neural networks (ANN) could provide indications of how to do it, given the fact that ANNs would be able to generalize new cases from those previously learned - and do it without principles of any kind. Thereby, ANNs are modeled to classify and reclassify cases with a moral purport, being the output (acceptable or not) an answer to moral dilemmas attached to the questions *kill* or *let die*. According to the author, the results suggested that the classification of non-trivial cases from the absence of queries about moral principles would be more plausible than might be supposed at first sight, although important limitations suggested the need of principles. Regarding a reclassification, which would be an important part of the reasoning in humans, simulations indicated the need for moral principles. Both (Wallach et al., 2010) and (Guarini, 2006) underline the value of the Theory of Mind and cognition into the subject of morality.

1.2 Background

The simulation of moral behavior provided by artificial emotions and feelings may, as in humans, aid a computational system to take faster and more intelligent decisions, or may prove itself important to prevent the execution of undesirable actions and for finding an agreement during the process of decision making in MAS environments.

Our purpose is to develop an architecture able to simulate moral behavior during social interaction. Regarding the exercise of empathy, individuals are divided into three groups: moral, immoral and amoral. Unpretentiously but in a simple approach, the formers have the social feeling of empathy properly functioning; and the immoral perform actions that somehow hurt the established moral code of his/her community. The latter can be interpreted as moral or immoral, depending on his/her social behavior. The amoral is thus characterized by absence of a mechanism that allows the individual to put himself/herself in the place of the other, and be sympathetic to his/her circumstances. We stress that there is neurophysiological basis for this classification: according to (Kandel et al., 2000) the lateral orbitofrontal cortex seems to participate in mediating empathetic and socially appropriate responses, thus damage to this area would be associated with failure to respond to social cues and produce lack of empathy. The mechanism that allows the existence of empathy is described in (Damásio, 2004) through the cognitive perspective but, as in (Proctor et al., 2013) and (De Waal, 2009), on the account of the emotional standpoint.

In (Damásio, 2004) there is the consideration that the brain can internally simulate certain emotional states, establishing a ground for emotionally possible outcomes and emotion-mediated decision making. There is also the thought that internal simulation takes place during the process along which sympathy emotion turns into the feeling of empathy. Regarding the social interaction, this is produced via mirror-neurons that can, for example, make our brain internally simulate the movement that others do while in our field of vision. That kind of simulation would allow us to predict the movements that would be necessary to establish communication with the other, which will have its movements mirrored. Finally, the internal simulation about our own body could be as well related to the mirror-neurons. Mirror-neurons were discovered in the premotor cortex area of macaque monkeys by (Di Pellegrino et al., 1992), (Rizzolatti et al., 1996). In (Gallese and Goldman, 1998) there is a reflection regarding the human aptitude of simulating the mental states from others, and thus understanding their behavior, assigning to them intentions, goals or beliefs. It is suggested that what might have evolved to such a capacity is an action execution/observation matching system; also, that a class of mirror neurons would play its role on that. Moreover, a possible activity of the mirror-neurons would be to promote learning by imitation. It is stressed that there is now the agreement that all normal humans develop the capacity of representing mental states from others (the system representation oftentimes receives the name folk psychology). Finally, there is the consideration that fitness could be availed from such ability, as detecting another agents goals and inner states can help the observer to predict the other future actions, which can be cooperative or not, or even threatening.

To summarize, the social emotion of sympathy feeds the feeling of empathy. But the social emotions benefit from the internal simulation improved by mirror-neurons that internally mirror the situation of the other. A possible activity of the mirror neurons could be promoting learning by imitation. The feeling of empathy will be less or more intense depending on the importance of a particular other agent (Damásio, 2004). Our research and corresponding premises are being guided by the differentiation of three types of agents regarding the feeling of empathy: the moral, immoral and amoral. The three will have different action policies, as their social interactions will be guided by a model that tries to simulate a pattern of morality. The moral tries not to take advantage from the others; more than that, tries to cooperate even if that is a good option only to the other; the immoral takes advantage from the others more easily and cooperates less often. The amoral imitates the social behavior of the others.

2 PROPOSED ARCHITECTURE

According to (Damásio, 2004) we seek to maintain negative emotions in low levels and, the positive ones in high levels. So the purpose of homeostasis would be to product a state of life better than neutral, to accomplish what we identify as well-being. MultiA will establish its preferences considering its own and peers well-beings (WB). Damásio (Damásio, 2004) defined social emotions using the concept of moral emotions by (Haidt, 2003), we will do the same while designing the artificial emotions. (Haidt, 2003) explains emotions as responses to a class of events perceived and understood by the self and so the emotions usually provoke tendencies of actions. It is particularly important to differentiate social emotions from other emotions: social emotions trigger action tendencies during situations that did not represent direct harm or benefit to the self (disinterested action tendencies), other emotions are more self-centered. Notwithstanding, (Haidt, 2003) also points out that all social emotions are likely to indirectly benefit the self.

The general scheme of MultiA is illustrated in Figure 1. It is composed by four systems: Perceptive System (PS), Cognitive System (CS), Learning System (LS) and Decision System (DS). The PS receives from the environment the current number of neigh-

bors; the reinforcement from the interaction with a neighbor; and an identifying index of the neighbor. It then updates its artificial sensations, emotions and feelings. The CS helps that operation through providing past data (memories): a history of the agent's action selection and preferences. Beyond that, the CS updates its ANN (called ANN_n), which mimics a mirror-neuron network, thus internally simulating the neighbor's current well-being. Therefore, ANN_n plays its role on updating the social emotion of sympathy that will feed the feeling of empathy (from PS). Finally, the agent's well-being (WB) will be measured taking into account its current feelings. Through these feelings, the WB represents how suitable has been the action selection, taking into account the agent itself but also the utility of the neighbor. The empathy can be more or less stressed depending on this utility and on the sympathy, both feed the empathy feeling which, on its turn, refeeds the social emotion of sympathy. The latter is also fed by the output of the ANN_n. The LS has one ANN for each action, and uses the Q-Learning algorithm (Watkins, 1989) to estimate the utility value for the paired current feelings (input space from PS) and action. Each ANN is trained according to the outcome driven by the execution of its corresponding action (Lin, 1993), employing the agent's well-being as the target value. Thus, the output (Q-value) from an ANN represents the WB that will result to the agent if, in response to the current feelings, the agent selects the action represented by the ANN. Note that as the empathy feeling will be part of the input space, learning will contemplate a WB that takes into account the impact of the agent's action to the neighbor. The DS will consider the Q-values acquired from the LS and choose an action: during the beginning of a simulation, the LS uses a high exploration rate for the state-action space.

2.1 Experimental Setting

MultiA will be tested and examined in a task and environment defined by (Wang et al., 2011), where three network topologies (lattice, scale-free and small-world) were used as interaction models for a generalized version of the Prisoner's Dilemma (PD) where a cascading failure effect (due to multiple agent defections in opposition to cooperation) takes place, with the aim of simulating the cascading effect of economic crises with multiple bankruptcy, or even disappearance of agents in a short time span. The cascading effect is obtained trough the elimination of agents (nodes and its connections), and occurs whenever the agent does not succeed in getting enough (that is, above its survival tolerance) cooperative actions from its neighbors. The higher the profit for defecting in response to the neighbor's cooperation, the higher is the probability of spreading the defection strategy to the network as the agents have the possibility of imitating a random successful neighbor strategy. Thus, because of the cascading node defection process, the network structure co-evolves with the interactions of the agents. Due to the elimination mechanism, cooperation would become the optimal strategy and ultimately overcome defecting as successive agent interactions take place. Thus, if not all agents are eliminated, a pure state of cooperation can emerge. (Wang et al., 2011) conclude that a process of cascading failure may take place when there are defecting strategies and vulnerable agents, and their results suggest that rational agents could survive and make profit through cooperation, naturally solving the social dilemma of profit *versus* cooperation.

The option of imitating a random neighbor action selection will be available to MultiA through the use of mirror-neurons that internally simulate the neighbor's well-being caused by the history of action selection. As our goal is to simulate the emergence of moral behavior from action selection influenced by the feeling of empathy, it is important to have hypotheses that guide the interpretation of the results. First of all we must reflect about our main inspiration, the human moral behavior: would the human be a naturally social creature or that condition would have emerged only for survival? Would the human be the *Zõon Politikón* from (Aristotle, 50BC), the *bon sauvage* from (Rousseau, 1762), or the human nature would have the tendency to fear all against all (Hobbes, 1651)? Relating to the moral behavior, which would be the human universals? Those questions relate to the origin or maintenance of moral behavior, but one more issue which deserves attention is about the definition of what could be more adequate to model as an *artificial* moral behavior - especially as sometimes what one get is not what one intended it to be. Would it be more convenient to develop a decision process that tends to something more like (Machiavelli, 1532) or certain attitudes, independently of the finality, would be unacceptable? In face of the vast quantity of information and uncertainty, if the action selection of an AMA tend to utilitarian parameters, would it really produce the greatest possible good for the greatest number of people (Bentham, 1907)? Still remain doubts about what tendencies should be designed in AMAs. The hypotheses guiding our research are:

1 What kinds of agents will exist? They will be designed according to three sets of premises: moral (MA), immoral(IA) and amoral agents (AA). The MA

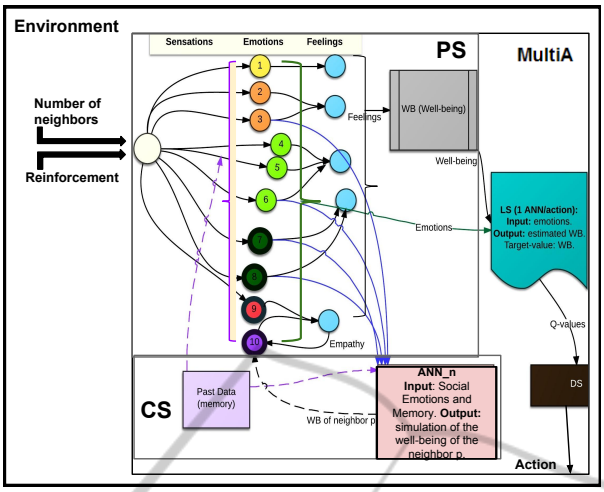

Figure 1: The general scheme of MultiA.

strongly cares about a neighbor, even though that may bring its own elimination. But it also may defect with the aim of isolating a constantly defecting neighbor. IA also care about neighbors, but is concerned with the profit it can get through social attachment: it will cooperate mostly with its group of IAs, but can decide to cooperate with others if it is getting isolated (to prevent complete isolation and elimination). Thus, MA usually cooperates; IA will cooperate (if so) mostly with some IA and AA can imitate both.

2 For different kinds of agents, what is the meaning of defection? It can be executed by all of them, the difference is the goal behind that action: if it is to make profit (IA and AA); if it belongs to an exploratory phase (MA; IA and AA); if it is to prevent elimination (MA; IA and AA) or, even, to eliminate a neighbor from the group (MA; IA and AA). If the agent can observe and differentiate its neighbors, it can learn to respond differently to them and to stop cooperating with defectors. Both IA and AA will isolate defectors more easily than MA. Moreover, AA, in order to survive and keep neighbors, will mirror MA and IA according to convenience.

3 Relating to the kind of agent, will it lead to relevant difference to the network structure? As MAs are naturally cooperative, they are supposed to keep as neighbors IA and AA. Therefore, the final population of a moral majority will contemplate a reasonable number of IAs and AAs. Accordingly, IA, in a society of immoral agents, will easily isolate a defecting neighbor by not caring about it (elimination) and only about the advantage, if any, of maintaining the neighborhood. Thus, the final population will remain with a large proportion of cooperative immoral agents, as the defecting IA will be easily excluded. As the AAs will imitate a neighbor, they will add uncertainty as they change strategy.

4 What would we expect from artificial empathy? Would it be convenient to develop a decision process that tends to something Machiavellian? If the action selection of one AMA tend to utilitarian parameters, would it really lead to the better good for all? What is best: to maintain a defecting neighbor in order to not lose it or just eliminate it? Should the AMA be morally hybrid: immoral towards agents that fail or delay the task and moral while interacting with living creatures? As an example, consider artificial agents having to coordinate activities and priorities in order to formulate a traveling plan for a human or complete the task of finding an object in a certain environment, if one agent from the group stops working or fails, it might be better to isolate it from the group. This is a cut off the artificial empathy feeling about that one agent, so the agent that simulates moral behavior will have the tendency to cooperate but, if it is required, it could also act in a different direction.

3 FINAL REMARKS

From the biological basis provided by (Damásio, 2004), our purpose is to modify ALEC (Gadanho, 2003) to obtain a rudimentary AMA, called MultiA architecture. Morality will emerge from the action selection policy (cooperate or defect). The temptation to defect will have different impact on the agent's goals according to its feeling of empathy and the simulation of mirrors-neurons (supposed to give as output the well-being of the other). On the experimental setup, defection will represent a way of getting profit but, also, of isolating someone from the group. Subsequent to the conclusion of the design of MultiA we

consider as future work bringing reputation (Brigatti, 2008) as an influential aspect for moral behavior.

ACKNOWLEDGEMENTS

The authors thank CNPQ and FAPESP for the financial support.

REFERENCES

Anderson, S. and Anderson, M. (2011). A prima facie duty approach to machine ethics and its application to elder care. In *Human-Robot Interaction in Elder Care*.

Aristotle (2013 (350BC)). *Aristotle's Politics*. Chicago U.

Bello, P., Bignoli, P., and Cassimatis, N. (2007). Attention and association explain the emergence of reasoning about false beliefs in young children. pages 169–174.

Bello, P. and Bringsjord, S. (2012). On how to build a moral machine. *Topoi*, pages 1–16.

Bentham, J. (1907). *An introduction to the principle of morals and legislation*. Oxf U. Press.

Brigatti, E. (2008). Consequence of reputation in an open-ended naming game. *Physical Review E*, 78(4):046108.

Bringsjord, S., Arkoudas, K., and Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *Intelligent Sys., IEEE*, 21(4):38–44.

Damásio, A. (1994). Descartes' error (new york: Putnam).

Damásio, A. (2004). *Looking for Spinoza: Joy, sorrow, and the feeling brain*. Random House.

Dancy, J. (2010). Can a particularist learn the difference between right and wrong? In *The Procs. of the twentieth world congress of philosophy*, volume 1, pages 59–72.

De Waal, F. (2009). *The age of empathy: Nature's lessons for a kinder society*. New York: Harmony.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental brain research*, 91(1):176–180.

Gadanho, S. (1999). *Reinforcement learning in autonomous robots: an empirical investigation of the role of emotions*. PhD thesis, U. of Edinburgh. College of Science and Engineering. School of Informatics.

Gadanho, S. (2002). Emotional and cognitive adaptation in real environments. In *In the symposium ACE*2002 *of the 16th European Meeting on Cybernetics and Sys. Research*. Citeseer.

Gadanho, S. (2003). Learning behavior-selection by emotions and cognition in a multi-goal robot task. *The Journal of Machine Learning Research*, 4:385–412.

Gadanho, S. and Custódio, L. (2002). Asynchronous learning by emotions and cognition. In *Procs. of the seventh Int. Conf. on simulation of adaptive behavior on From animals to animats*, pages 224–225. MIT Press.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501.

Guarini, M. (2006). Particularism and the classification and reclassification of moral cases. *Intelligent Sys., IEEE*, 21(4):22–28.

Guarini, M. (2012). Moral cases, moral reasons, and simulation. *AISB/IACAP World Congress*, 21(4):22–28.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences*, pages 852–870.

Hobbes, T. (1969 (1651)). *Leviathan*. Scolar Press.

Jackson, J. V. (1987). Idea for a mind. *ACM SIGART Bulletin*, 101:23–26.

Kandel, E., Schwartz, J., and Jessell, T. (2000). *Principles of neural science*, volume 4. McGraw-Hill New York.

Lin, L. (1993). Reinforcement learning for robots using neural networks. Technical report, DTIC Document.

Machiavelli, N. (1985 (1532)). *The Prince*. U. of Chicago.

Matignon, L., Laurent, G., Le Fort-Piat, N., et al. (2012). Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge E. Review.*, 27(1):1–31.

Proctor, D., Brosnan, S., and De Waal, F. (2013). How fairly do chimpanzees play the ultimatum game? *Communicative & integrative biology*, 6(3):e23819.

Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive brain research*, 3(2):131–141.

Rousseau (1971 (1762)). *The Social Contract*. Penguin.

Sun, R. ; Peterson, T. (1998). Autonomous learning of sequential tasks: experiments and analysis. *IEEE Transactions on Neural Networks*, 9(6):1217–1234.

Wallach, W. (2010). Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics and information technology*, 12(3):243–250.

Wallach, W. and Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxf U. Press.

Wallach, W., Franklin, S., and Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, 2(3):454–485.

Wang, W.-X., Lai, Y.-C., and Armbruster, D. (2011). Cascading failures and the emergence of cooperation in evolutionary-game based models of social and economical networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3):033112–033112.

Watkins, C. J. (1989). *Learning from delayed rewards*. PhD thesis, Kings College, UK.

Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.