On the Intelligence of Moral Agency

Helder Coelho[1] and Antônio Carlos da Rocha Costa[2]

[1] LabMAg, Faculdade de Ciências da Universidade de Lisboa, 1749-016 Lisboa, Portugal, hcoelho@di.fc.ul.pt

[2] Escola de Informática, PPGINF, Universidade Católica de Pelotas, 96.010-000 Pelotas, RS, Brazil, rocha@atlas.ucpel.tche.br

Abstract. More advanced and complex applications, such as serious games, where physical and virtual environments interchange with human and artificial agents and along heavy social simulations, require another sort of architectures. With the enlarging autonomy comes an increased need to ensure that their behaviour is in line what we expect from them. Therefore, a combination of intelligence and ethics becomes mandatory, and this means new design principles and technical requirements for the social agency and the presence of trust and confidence. Mentality, before the sole key concern, is now mixed with morality and within the social spaces where autonomous agents act on our behalf. In order to model new agent behaviours with qualities we need other kind of more intricate mental models, able to support moral reasoning capabilities. Today, the pressing quest is which are the crucial building blocks and mechanisms for those inovative agents.

Keywords: moral agency, moral agents, norm innovation.

1 Introduction

"The function of Moral is to guide action intuitively and unconsciously."
J. Greene, 2007.

Environment, cognition, emotions, peer pressure, values, pride and the social relations all influence our decisions between choosing right over wrong. Any decision an agent makes when it comes to prefer a good or a bad behaviour reveals his true character. This implies also agents must have an explicit conception about the outcomes of their actions and the capability to classify and assess them accordingly.

Agency is only the capacity of an agent to act in a world, yet moral agency is the responsability for making moral judgements about the acting choices, and morality refers to a certain code of conduct and a system of actions and reactions directed to keep everyone behaving according to it. In brief, Moral refers to those explicit and implicit rules and actions able to govern agents´s social behaviour.

A clear understanding of how cultural changes interact with individual agent

actions is central to informing democratically and humanely guided efforts to influence cultural evolution. The example of norm innovation can help us to understand the complex 2-way dynamics of sociality (how new conventions spread in social systems), because norms emerge at the aggregate level (and immerge into the minds of agents) to fix the future behaviour of agents and the whole functioning of the society.

Moral systems are composed of four kinds of regulative elements: moral norms, moral values, moral judgements and moral actions. Norms are (conventional, social) rules or patterns of behavior, serving to maintain order and to guarantee social regulation. But, norms (and institutions) have a short life. Harmonization occurs and social order is restored. Norm innovation depends on the mechanisms by which new norms are conceived, the conditions under which they are spread, the extent to which they evolve as they are distributed through all the society, the circumstances under which they become institutionalized, and, the process through which they decay, are lost and replaced by new ones.

Moral values are of two kinds (reference and assessment values), and serve the purpose to set standards of quality and direction to the agent behaviours, and they are both closely connected to moral norms. Reference values are high level values that the society adopts in order to characterize itself in general terms (democracy, liberty, progress, adherence to heritage, religiousness, etc.). Reference values tend to be the defining elements of norms, in the sense that norms are conceived and adopted to control behaviours so as to keep the society adherent to those reference values.

Assessment values are operational values with which behaviours are dynamically evaluated, as a consequence of their compliance or not with norms. Behaviours that comply with norms are assessed positively, and behaviours that do not comply are assessed negatively. The intensity with which the behaviours comply or deviate from the norms are reflected in the magnitude of the assessment value assigned to the behaviours (bad, very bad, good, very good, etc.).

Moral judgements are rational opinions with which agents classify each other´s behaviours, according to their compliance or not to the current set of moral norms. As a result of a moral judgement, a behaviour is marked as compliant or not to the set of moral norms, and a moral norm value of the assessment kind is assigned to it. Moral judgements may be combined with other kinds of rational judgements to form moral reasoning, which are the special kinds of social reasoning through which the agents decide and/or justify the moral actions that they take.

Moral actions are regulative control actions (meta-actions) that the agents emit in order to influence each other about the adequacy of other behaviours to the current set of moral norms. Such regulative actions are of two special kinds, either punishments or rewards, and tend to assign additional costs (punishments, interdictions) or to supply new resources (rewards, permissions) to the agents´s actions, according to the

moral evaluation (praise or blame) made about them, and to the moral values associated to them.

Concretely, moral actions may take either the form of behaviour control, affecting the possibility of the agents´s actions, or the form of organizational control, affecting the way agents adopt social roles to each other. How do moral systems evolve? How are they represented in the mind of an agent? How are moral actions concretely realized and become effective in a certain context?

Any society can be viewed as organized along two main levels: on the bottom, the economical-material infrastructure, and at the top, the moral-cultural superstructure. There is an ongoing flow between the two levels (micro-macro) of norms and values: moral values of reference (high-level values) and moral values of evaluation (low-level or operational values). The arrival of new norms and the renewal of existing ones are related with the adaptation of reference moral values to the current working of a society. Norm innovation is guided by the evolution of reference values (with moral-cultural character) which are chosen as a consequence of political-economical dispute around the economical-material values.

More research and experimentation is necessary on questions of transmission, transformation and contribution of the mental constructs to understand the dialectical relation between social structures and individual agency and collective interaction, say the dynamics of sociality. This new knoeledge will have an influence on how the artificial mind of an agent may be architectured.

2 State of the art

The topic of moral agents became hot in recent years (see IEEE Intelligent Systems Journal, July/August 2006), due to the original scientific contributions coming from Cognitive Neuroscience, Evolutionary Psychology, or even Philosophy. Damasio´s group, at the University of Southern California, covered the social spaces (individuals in relation to others), the physiological roots of some social emotions (happiness, pride, compassion). Hauser, at Harvard University, suggested our moral judgements are derived from unconsciousness, intuitive processes that operate over the causal-intentional structure of actions and their consequences (Koenigs et all 2007). He believes we have a moral organ, a sort of faculty, able to embed a universal moral grammar (Mikhail 2007), a tool to build up specific moral systems, able to generate judgements about permissible and forbidden actions prior to the involvement of our emotions and systems of conscious, rational deliberation. Accoding to Hauser, moral rules have two ingredients, a prescriptive theory or body of knowledge (social conventions, norms, ceremony manners) about what one ought to do, and an anchoring set of diverse emotions.

Within Artificial Intelligence, Cognitive Science and Social Psychology, several authors in the last decade (Bazzan et all 1999), (Bordini et all 2000), (Allen et all

2000), (Ribeiro and Costa 2003), (Floridi and Sanders 2004), (Dimuro et all 2005), (Wiegel et all 2005), (Allen et all 2006), (Kowalski 2007), (Anderson et all 2006), (Guarini 2006), (Moor 2006), (Bringsjord et all 2006), (Wiegel 2006), (Costa and Dimuro 2007), (Savarimuthu and Purvis 2007), (Franco et all 2008), (Hegselmann 2008), (Lotzmann et all 2008), (Will 2009) start to advance ideas about Hume (sentiments), Hume/Kant (sentiments and reasoning) or Rawls (action grammars) moral creatures, building up artificial virtues (enforcement agencies). Around all these contributions we can also find a diversity of agent´s architectures devised to engender specific characters and personalities, namely recent work by (Wiegel 2006) on the necessary building blocks of an agent, (Lorini and Castelfranchi 2007) on the cognitive structure of surprise, and (Mascarenhas e all 2009) on cultural agents.

3 Moral machine

The mixture of reasoning and emotion is behind the generation of moral behaviours and judgements about gains and losses of an agent (Koenigs et all 2007). Meanwhile, emotions work as weights, pushing us more for one side that the other. The same occurs with mentality, where each mental state (eg. a belief or a desire) is constrained by a set of attributes and values (Antunes 2001), (Corrêa and Coelho 2004). The generation of actions will be responsible for producing acts, linked to utilitarian (focus on consequences) or deontological (focus on rules) judgements.

The discussion on the precise building blocks of a second generation of moral agents was advanced for the first time by (Wiegel 2006), around the construction of the SophoLab Project and it was supported on the BDI model and the JACK agent language. He clarified the general structure and organization, the design principles and the technical requirements, in particular the emergent behaviour, the redundant degrees of freedom, the absence of any central director of action, and the local scope of control. The first generation moral agents supported forms of interconnectedness (team work), multipurpose goals, but it was limited in embodiment and epistemology, the two crucial features to be explored later on. Research done around moral grammars by (Mikhail 2007) and (Hauser 2006) was not yet conclusive, from the point of view of the agent´s autonomy and its place in the heart of the organization.

At the same time, the attempt to construct artificial agents, governed by norms, was also one of the aims of the EEC EMIL (Emergence in the Loop, Simulating the Two Way Dynamics of Norm Inovation) project, started in 2006 (Andrighetto et all 2007), with the main focus on norm innovation (Lotzmann et all 2008). So, the agent design of EMIL-A was guided by the norm formation process (information transfer structure). The functional description, around the first prototype in NetLogo, was unable to reveal the mental side and to explain how autonomy and potency were not taken into full account (Lotzmann 2008). The planning (making capability) was very simple and the decision-taking module was no more than a trivial utility function. Other experiments by (Andrighetto et all 2008) adopted also simple agent

architectures and did not reflect upon the ideas based upon the dialogue between cognition and affection.

Proposals, done in the pasto n drives and will inclusion were not taken into care, and even no explanation was given on how cognitive and affective states may interact (through layers) to engender a moral reasoning capability, a hint defended by Hauser and Damasio (Koenigs et all 2007) and by Cognitive Neuroscience at large (Greene 2005).

Agent autonomy depends heavily on the power-of ability and it is dependent on the will mental state (Coelho and Coelho 2009), a missing point of BDI model to allow a kind of insurgent agents with direct action potency. The introduction of a deontic element by (Wiegel 2006), extending the standard BDI model through the deontic-epistemic-action logic (DEAL framework), will not be sufficient to capture the whole flavour of a moral agency, and the same argument can also be applied to the use of BOID model (Broersen et all 2001), more keen to engender personalities. The moral conduct of an agent requires more than the means-ends analysis, the so-called planning capability of the BDI model.

A social space, in progress, associated with a serious game for managing human resources, requires a moral agency with much more advanced features. Before, in the serious game around the management of natural resources (say, water), the simplified BDI was adopted (Adamatti et all 2009), and as a consequence social interactions among agents were of poor quality. In other experimentation conducted by (Costa and Dimuro 2007) and (Franco et all 2008) moral sentiments were not taken into account because the selected scenario was not demanding in ethical considerations. On the contrary, in this particular case study, the personality (temper, character) of the agents was one of the major concerns. This year a proposal for the design of cultural agents, by (Mascarenhas 2009), mixed several sub-systems of a general agent for synthetic characters and it was near of the basic ethics machine.

4 Case study: norm-innovation

Let us select the case study of norm innovation in a society to explain how the morality Works in general. Norms play two roles, the constitutive one, to generate emergent social behaviour, and the regulative one, to be a source of social order (engender the social structure). The example of road traffic and pedestrian crossing, implemented by (Lotzmann et all 2008), show how norms modify behaviours.

A society, or even a small social community, can be seen as a complicate system of agents. Complex systems are composed of many different interacting autonomous elements and governed by social laws, with non-linear relations and network structures. So, complexity can be classified as physiological, as we look to the size or as social as we look to interactions. Let us zoom now on these systems composed of many interacting intelligent autonomous agents (AA´s).

AA´s are able to adapt and evolve with a changing environment, making independent decisions. They form new mental objects and processes, consequent to the emergent behavior of the whole system they are a part of, and act on the basis of these particular objects and processes. When a change happens at some level, it requires some further change in agents´s individual behaviours and beforehand in their minds, in order to allow the new pattern spread over that system. Norm creation and behaviour evolution work together, and the micropsychology selects behaviour, institutions and evolution to be looked at.

Individual mental change allows agents to modify their behavior accordingly. Yet, innovation in social systems is a bi-directional process: 1) bottom-up, emergence of new entities or phenomenons at the aggregate level and from the interactions among agents impose creation of structures; and, 2) top-down, immergence of entities or phenomenons in the minds of the agents, ie. the insurgence in their minds of a new mechanism, representation or process that leads the agents to modify their behaviours in conformity with the emerged effect. Sociality is viewed along two loops: the emergence of structures at the macro level (social relations, groups) and the immergence of norms at the micro level (individual interaction).

The behavior regulation is done by norms (rules, conventions, patterns) which are central in the role theoretic concept of individual action and decision taking. There are two main types of norms, those for coordination, and those for obligations, prescriptions and directives on commands. The regulation is made by 1) change of action and 2) roles (defined by attributes, behavior and social relations). And, the fine tuning is achieved by 1) processes among agents, 2) normative transmission, and 3) transformation.

Moral agency is the agent´s responsibility for making moral judgements and taking actions that comport with morality. Moral decisions (foundation of morality) are triggered by reason (evidence, facts) and also by emotion. Both cooperate for generating moral sentiments.

Moral judgement uses a utilitarian calculus for choosing between right and wrong. Behind a moral decision there is always an interplay of thought, emotion, prevision, empathy, anguish and ambivalence.

How is normative governance effected? Societies (groups) are regulated by different sorts of mechanisms associated to 1) decaying and spreading of norms, and also to 2) transformation and internalization of norms.There are three main classes of normative agent-based simulation models. The point of departure was done by (Axelrod 1986) along the game-theoretic approach. It is still good for explaining the dynamics of norms, and the strategic adaptation of agents to changing environmental conditions.

The second direction of simulation models is the AI approach. It follows (Conte and Castelfranchi, 1995) and demonstrates the effects of norms. It includes norms that exceed strategic adaptation, which can be interpreted as an internalized property (Castelfranchi et all 1998), (Castelfranchi 1999). The weak point: agents are normative automata and no mechanisms for transmission and transformation problems are allowed.

Actualy, there is a convergence of both traditions: the models of (Verhagen 2001) and (Savarimuthu 2007) include elements of the other line of thought, and a partial answer to the questions of transformation, transmission and contribution. Both models replicate not only the findings, but also the shortcomings of the classical role theory.

5 Prospect for a moral agent

The interplay of mentality, sociality and morality reveals the definite anatomy of those smart creatures able to think about, to interact with others in a society, and also to decide upon good and evil. Which is the most suitable architecture for an agent with these three qualities?

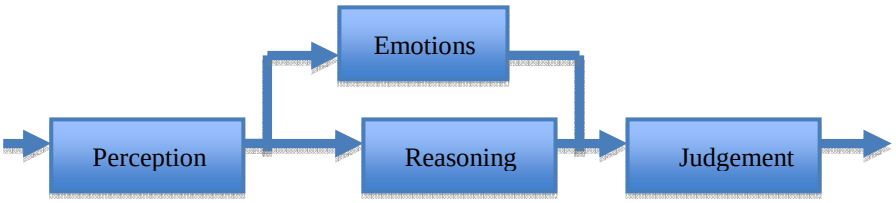

Fig. 1 Simplified sketch of a moral creature influenced by Hume and Kant ideas.

Agents can be reduced to simple bit strings when genetic programming is adopted in social simulation of complex scenarios. In what concerns symbolic programming, an agent can be more elaborated than a decision (utility) function. For example, in a risky environment, (Castelfranchi et all 2006) adopted a one-layer structure by mixing an extended BDI with an emotion manager for modeling cautious agents. In the serious game (mixing human and artificial agents) of water management (Adamatti et all 2009), an artificial agent has a behavioural profile linked to one or more strategies regarding a certain role (BDI model was simplified), no learning and planning modules were available, and only reduced decision making skills were offered, and again a one-layer structure was adopted. In another serious game on participatory management of protected areas (Briot et all 2008), conflict dynamics was taken care and a more advanced decision capability was implemented, but agents had no mentality and affective power. When designing cultural agents, (Mascarenhas 2009) updated an old architecture of social intelligent agents for educational games and proposed to combine a memory store with a reactive device and a deliberative machine, without forgetting the motivational states of the other agents. However, serious games require moral agents in order to be acceptable (serious) by users.

A moral agent, as it was defended by Hauser (see figure 1) and Green, is a mix of cognitive and affective capabilities, but no architecture was till today presented as the definite one, despite several design attempts from Hume, Kant or Rawls, and the trials made by the community of Agents. Several questions needed to be answered: What makes a moral (norm-abiding, virtuous, conventional) agent? By what mechanisms can moral behaviour (abstract values) spread or decay from one agent to another (like memes)? How are explicit morals implemented and added to the overall architecture?

The human mind is not completely rational in order to win when facing reality. Therefore, an artificial mind is also forced to avoid traps and to take into account other skills apart of reason, such as sensorial perception, intuitions, emotions and socio-cultural constructs. For example, prefer a short cut when facing discomfort, assign a value based upon some preconceived opinion, or go on with some previous agreement can disturb good decisions. The answer is to adopt a set of heuristics, like to maintain the long term plan, delay action in order to think about several alternatives, and to perform as an outsider when looking to that issue for the first time. Be less moralist when analyzing what is or not right, use those issues involved in the decision (more cooperative), and follow devil´s lawyer to embody relevant data (otherwise absent) is a recipe to avoid catastrophes.

A moral agent needs to get a more intricate way of thinking (Kowalski 2007) than a simple reactive (assimilate observations of changes in the environment) or a proactive one (reduce goals to sub-goals and candidate actions). Why? It is not sufficient to embody a goal-based or a value-based model. We need a mix of intuitive (low level) and deliberative (high level) processes, and also the ability to think before acting (pre-active) when choosing between right or wrong, ie. capability to think about the consequences of the candidate actions (generate logical consequences of candidate actions, helping to decide with heuristics or decision theory between the alternatives). The classic component based on the observe-think-decide-act cycle (present in the BDI model) is unable to deal with morality because we get different kinds of goals (achievement, maintenance) and, at the same time, preferences and priorities are requested. The one-layer structure is no longer the solution because we arrive at our ultimate moral (utilitarian, where results maximize the greatest goods, or deontological, where any moral evaluation is independent of consequences) judgements by a mix of emotions and conscious reasoning. As a matter of fact, emotions drive behaviours like weights, and play a critical mediating role in the relationship between an action´s moral status and its intentional status. A moral ability may be seen as a set of rules (a grammar according to Hauser) to constrain the behaviour of the agent: each rule having two ingredients, the body of knowledge and the set of anchored emotions, which are going to interplay.

Every decision an agent makes, when it comes to choosing between right or wrong, reveals his true character (subjectivity): Humean model with emotions behind judgements, or Rawlsian model, with emotions and reasons after judgements have only one layer and trade-offs are not allowed. There is always a sentiment of

avoidance in violating what seems to be reasonable, ie. the possibility to have access to the outcomes (classifications) of the agent actions.

A moral agent associates always reason with emotion, social values and cultural-situational knowledge before making a decision. Therefore, its more-than-one-layer architecture, integrating micro and macro levels, requires an extended (with will and expectations) BDI model, the addition of emotional machinery to deal with sentiments, a library of contexts to situate any evaluation, heuristics to avoid wrong decisions (mind traps), a sort of universal moral grammar (Mikhail 2007) to fix any sort of moral system and action generation, and also modules concerning decision taking, constraint satisfaction (reinforcement) learning and planning. The organization with interconnected multiple layers seems inevitable on account of the balance between reasoning and emotion and the assembling/tuning of composite judgements (embedded in preference criteria).

6 Conclusions

The research and experimentation around the intelligence of moral agency is betting on understanding and managing complexity in social systems with autonomous agents. The selection of norm innovation, as a topic, is also helping us to comprehend now how new conventions and principles of right (and wrong) action emerge and spread in those systems to get social order. Norms are complex social artifacts because of the role in connecting emergence and immergence, through a movement between micro and macro levels. Applications such as regulation of e-communities or realistic serious games for managing human capital are eager of new agent models and architectures with ethical concerns and some sort of subjectivity.

Several open questions frame our current research: How do actors produce and are at the same time a product of social reality? How an idea (memes) for a behavior that becomes a norm gets invented in first place? Which ideas are accepted and which are rejected driven by adaptation and evolution? How many are slowly assembled from diverse data in a single mind? Answers, from Cognitive Neurosciences, Moral or Evolutionary Psychology, point to a strong focus on a context sensitive approach to agency and structure, the interplay of which leads to emergent phenomena, underlining the generative paradigm of computational social science. Agent-based modeling and simulation can be of great help in order to allow a better comprehension of this sort of complexity.

References

Adamatti, D., Sichman, J. and Coelho, H. An Analysis of the Insertion of Virtual Players in GMABS Methodology Using the Vip-JogoMan Prototype, Journal of Artificial Societies and Social Simulation, JASSS in press, 2009.

Allen, C., Varner, G. and Zinser, J. Prolegomena to any Future Artificial Moral Agent, Journal of Experimental Theoretical Artificial Intelligence, Volume 12, pp. 251-261, 2000.

Allen, C., Wallach, W. and Smit, I. Why Machine Ethics? IEEE Intelligent Systems, Volume 21, Number 4, July/August, 2006.

Anderson, M. and Anderson, S. L. Machine Ethics, IEEE Intelligent Systems, Volume 21, Number 4, July/August, 2006.

Anderson, M., Anderson, S. L. and Armen, C. An Approach to Computing Ethics, IEEE Intelligent Systems, Volume 21, Number 4, July/August, 2006.

Andrighetto, G., Campenni, M., Conte, R. and Paolucci, M. On the Immergence of Norms: a Normative Agent Architecture, Proceedings of AAAI Symposium, Social and Organizational Aspects of Artificial Intelligence, Washington DC, 2007.

Andrighetto, G., Campenni, M., Ceccone, F. and Conte, R. Conformity in Multiple Contexts: Imitation vs. Norm Recognition, Proceedings of the World Congress of Social Simulation, Fairfax VA, July, 2008.

Antunes, L. Agents with Decisons Based in Values (in Portuguese), PhD Thesis, Faculty of Sciences, Universidade de Lisboa, 2001.

Antunes, L., Balsa, J., Urbano, P. and Coelho, H. The Challenge of Context Permeability in Social Simulation, Proceedings of the Fourth European Social Simmulation Association Conference, Toulouse, September, 2007.

Antunes, L., Balsa, J., Urbano, P. and Coelho, H. Exploring Context Permeability in Multiple Social Networks, World Congress on Social Simulation (WCSS-2008), George Mason University Fairfax (USA), July 14-17, 2008.

Axelrod, R. An Evolutionary Approach to Norms, American Political Science Review, Volume 80, 1986.

Bazzan, A. L. C., Bordini, R. H. and Campbell, J. Moral Sentiments in Multi-Agent Systems, in Intelligent Agents V, Springer-Verlag LNAI Nº 1555, pp. 113.131, 1999.

Bordini, R. H., Bazzan, A. L. C., Vicari, R. M. and Campbell, J. Moral Sentiments in the Iterated Prisoner´s Dilemma and in Multi-Agent Systems, Brazilian Journal of Economics, Volume 3, Number 1, 2000.

Bringsjord, S., Arkoudas, K. and Bello, P. Towards a General Logicist Methodology foe Engineering Ethically Correct Robots, IEEE Intelligent Systems, Volume 21, Number 4, July/August, 2006.

Briot, J.-P., Vasconcelos, E., Adamatti, D., Sebba, V., Irving, M., Barbosa, S., Furtado, V. and Lucena, C. A. Computer-Based Support for Participatory Management of Protected Áreas: The SimParc Project, Proceedings of XXVIIIth Congresso f Computation Brazilian Society (CSBC´08), Belém, Brazil, July 2008.

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z. and van der Torre, L. The BOID Architcture: Conflicts between Beliefs, Obligations, Intentions and Desires, Proceedings of the Fifth International Conference on Autonomous Agents and MultiAgent Systems (AAMAS01), Montreal, Quebec, Canada, pp. 9 – 16, 2001.

Castelfranchi, C., Conte, R. and Paolucci, M. Normative Reputation and the Costs of Compliance, JASSS, Volume 1, Number 3, 1998.

Castelfranchi, C. Prescribed Mental Attitudes in Goal, Adoption and Norm-Adoption,

Artificial Intelligence and Law, Volume 7, pp. 37-50, 1999.

Castelfranchi, C. The Theory of Social Functions: Challenges for Computational Social Science and Multi-Agent Learning, Journal of Cognitive Systems Research, Volume 2, pp. 5-38, 2001.

Coelho, F. and Coelho, H., Meta Agency and Individual Power, Web Intelligence and Agent Systems: An International Journal (WIAS), in print, 2009.

Conte, R. and Castelfranchi, C., Cognitive and Social Action, UCL Press, 1995.

Corrêa, M. and Coelho, H. Collective Mental States in an Extended Mental States Framework, International Conference on Collective Intentionality IV, Certosa di Pontignano (Italy), October 13-15, 2004.

Costa, A. R. and Dimuro, G. P. A Basis for an Exchange Value-Based Operational Notion of Morality for Multiagent Systems, Springer-Verlag LNCS, 2007.

Dimuro, G. P., Costa, A. R. and Palazzo, L. A. M. Systems of Exchange Values as Tools for Multi-Agent Organizations, Journal of Brazilian Computer Society, 2005.

Floridi, L. and Sanders, J. W. On the Morality of Artificial Agents, Minds and Machines, Volume 14, pp. 349-379, 2004.

Franco, M., Costa, A. R. and Coelho, H. Simulating Argumentation about Exchange Values in Multi-Agent Interactions, Proceedings of the 1st Brazilian Workshop on Social Simulation (BWSS 2008), 19th Brazilian Symposium on Artificial Intelligence (SBIA2008), Salvador. October 26-30, 2008.

Grau, C. There is No "I" in "Robot": Robots and Utilitarism, IEEE Intelligent Systems, Volume 21, Number 4, July/August, 2006.

Greene, J. D. The Terrible, Horrible, No Good, Very Bad Truth about Morality and what to do about It, PhD Dissertation, Princeton University, 2002.

Greene, J. D. Cognitive Neuroscience and the Structure of the Moral Mind, in The Innate Mind: Structure and Contents, S. Laurence, P. Carruthers and S. Stich (Eds.), Oxford University Press, 2005.

Guarini, M. Particularism and Classification of Moral Codes, IEEE Intelligent Systems, Volume 21, Number 4, July/August, 2006.

Hauser, M. D., Chomsky, N. and Fitch, W. T. The Faculty of Language: What is it, Who hás it, and How did it evolve), Science, Volume 298, 22 November, 2002.

Hauser M. D. The Liver and the Moral Organ, SCAN, Volume I, pp. 214-220, 2006.

Hauser, M. D. Moral Minds: How Nature Designed our Sense of Right and Wrong, Ecco/Harper Collins, 2006.

Hegselmann, R. Modelling Hume, a draft for HUME1.0, February, 2008.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. and Damasio, A. Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements, Nature, March 21, 2007.

Kowalski, R. The Logical Way to Be Artificially Intelligent, Lecture Notes in Computer Science, Springer-Verlag, 2007.

Lotzmann, U., Möhring, M. and Troitzsch, K. Simulating Norm Formation in a Traffic Scenario, Proceedings of the Fifth Annual Conference of the European Social Simulation Association, Brescia, September, 2008.

Lotzmann, U. TRASS, A Multi-Purpose Agent-Based Simulation Framework for Complex Traffic Simulation Applications, A. L. C. Bazzan and F. Klügl (Eds.), Multi-Agent Systems for Traffic and Transportation, IGI Global Press, 2008.

McLaren, B. M. Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions, IEEE Intelligent Systems, Volume 21, Number 4, July/August, 2006.

Mikhail, J. Universal Moral Grammar: Theory, Evidence and the Future, Trends in Cognitive Science, Volume 11, Number 4, 2007.

Moor, J. H. The Nature, Importance and Difficulty of Machine Ethics, IEEE Intelligent Systems, Volume 21, Number 4, July/August, 2006.

Nichols, S. Norms with Feeling: Toward a Psychological Account of Moral Judgment, Cognition, Volume 84, pp. 221-236, 2002.

Nichols, S. Sentimental Rules, Oxford University Press, 2004.

Pereira, L. M. and Saptawijaya, A. Computational Modelling of Morality, Working Report, 2009.

Powers, T. M. Prospects for a Kantian Machine, IEEE Intelligent Systems, Volume 21, Number 4, July/August, 2006.

Ribeiro, M. R. and Costa, A. R. Using Qualitative Exchange Values to Improve the Modelling of Social Interactions, Springer-Verlag LNCS, 2003.

Savarimuthu, B. T. R. and Purvis, M. Mechanisms for Norm Emergence in Multiagent Societies, Proceedings of the 6th International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS-07), Honolulu, Hawaii, 2007.

Urbano, P. Decentralised Consensus Games (in Portuguese), PhD Thesis, Faculty of Sciences, University of Lisbon, 2004.

Verhagen, H. Norms and Artificial Agents, Proceedings of the Sixth Meeting of the Special Interest Group on Agent Based Systems, 2001.

Verhagen, H. Simulation of the Learning of Norms, Social Science Computer Review, 2001.

Wiegel, V., van der Hoven, M. J. and Lokhorst, G. J. C. Privacy, Deontic Epistemic Action Logic and Software Agents, Ethics and Information Technology, Volume 7, pp. 251-264, 2005.

Wiegel, V. Building Blocks for Artificial Moral Agents, Working Report, Proceedings of EthicalALife06 Workshop, 2006.

Wiegel, V. Sopholab; Experimental Computational Philosophy (Simon Stevin Series in the Philosophy of Technology). Delft University of Technology, PhD thesis, 2007.

Will, O. Modelling Hume´s Moral and Political Theory, An Agent-Based Model on the Evolution of Trust in Strangers and Division of Labour Among, Working Report, 2009.