Decision Making for Agent Moral Conducts

Helder Coelho [1], António Carlos da Rocha Costa [2] and Paulo Trigo [3]

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

[2] Centro de Ciências Computacionais, PPGMC, Universidade Federal do Rio Grande,
96.201-900 Rio Grande, RS, Brazil, ac.rocha.costa@gmail.com

[3] LabMAg, Instituto Superior de Eng. de Lisboa, DEETC, 1959-007 Lisboa, Portugal,
ptrigo@deetc.isel.ipl.pt

Abstract. Looking to the operation of an agent architecture, ie. its goal generation and maintenance processing, is relevant to understand fully how a moral based agent takes appropriate and diverse decisions within social situations of serious games. How decision does happen is a complex issue and the major motivation of this paper, and our answer, the proposal of a new architecture, is supported on the clarification of the organization and structure of an agent, ie. the interpretation of agent actions (moral-driven behaviour) under the pressure of severe constraints.

Keywords: moral architecture, values and norms, behaviour regulation, morality reconsideration.

1 Introduction

> "Synthesis and simplification are the essential issues in architecture".
> Alejandro Aravena, Revista Única Sep. 19, 2009.

Recently, we have been discussing the proposal of an overall architecture of moral based agents, embedded in a social multitude, by facing two major issues, intelligence (Corrêa and Coelho, 2010) and complexity (Coelho and Costa, 2009; Coelho et al, 2010). This line of research is different from the logical programming direction, and it is more akeen to Sloman and Minsky bet on an architecture for cognitive diversity. The follow-up of a recent R&D EEC project, EMIL (2007-09), help a lot to clarify differences between norm-governed and moral agents. Norm processing is almost trivial as compared with moral decision. Several conjectures were put forward:

C1: A moral agent, like any cognitive agent, is defined by an hybrid architecture.
C2: The key, and central question, concerns how its decision policy is managed, because the environment is pro-active and it requests a certain complexity of social behaviour.

INForum 2010 - II Simpósio de Informática, Luís S. Barbosa, Miguel P. Correia (eds), 9-10 Setembro, 2010, pp. 721–732

C3: The architecture has many layers, at least four systems (cognitive, emotional, moral, and esthetical), and there are many interactions (feed-forward and feed-back flows) among its modules before an appropriate (moral) decision is attained.
C4: Choice and preference reconsideration (action selection) is mandatory. Moral agents have different individual cultures and values and must be cautious and respectful in order to avoid inappropriate behaviours (generation of social conflicts).
C5: Behaviours are ruled by norms which depend on values. Norms are means in order behaviours be compatible to moral values.
C6: Moral global behaviour is the result of many informed local decisions, taken by different modules and along n-layers, of feed-back and feed-forward moves, and the negotiation among those modules is often required to support the final decision.

2 Morality

A vision of morality, conformity to the rules of right (moral, virtuous) conduct, the so-called moral character (evaluation of particular individual qualities) is strongly connected to what is socially defined as normal or appropriate (March and Olson, 2009), true, right or good, in spite of the necessary calculus of consequences and expected utility. How agent conduct is engendered, according to values, rules, codes and principles is only a fraction of what is necessary. Any agent acts because it pursues to achieve a purpose or satisfy a desire, and it seeks also adequate actions in defence of its interests and, often, anticipates future consequences following criteria of similarity and congruence, rather than likelihood and value. Appropriateness reflects learning of some sort from personal history, but it does not guarantee technical efficiency or moral acceptability.

When moral judgments (weighing reasons for and against affective attitudes and moral intuitions) are given, in face of some non-trivial situation (eg. switch dilemmas in trolley problems), an intriguing contrast emerges between the intuitive opinions from those considering the scenarios. The respective acts may be evaluated differently and the choices ("sacrifice one life in order to save five") are unexpected for similar events. The interpretation can be explained by emotional arousal and by the importance attributed to intuitions. So, the utilitarian or deontological views are in danger to be good candidates for supporting moral judgment. We are convinced that somewhere in between lies the sound solution.

Are moral issues just a matter of taste or culture? Are moral judgments provoked by expressions of affective states on which reasons have little influence? Or, should more attitudes be justified, ie. some moral and cultural judgments are wrong, and others right, because they relate to facts of moral relevance in adequate or inadequate ways. What ought we do? Ignore all our ordinary moral judgments, and do what will produce the best consequences, or follow what we were told to do!

Morality is the respect for the other, and it is not a monolithic concept with sharp boundaries. Really, what happens in moral judgment, is mostly a part of typical response patterns (the so-called moral signature, full character of an agent), because there are aspects of acts and/or situations that are relevant to take moral decisions.

3 Moral character

When studying moral agents we are attracted by a diversity of feelings, a kind of pro-social sentiments, of guilt, compassion, empathy, anguish or ambivalence, triggered by states of affiliation or sadness. Moral decisions may be very complex because they entail the cooperative interplay of several systems, namely of thought, emotion, empathy or foresight, and along layers of importance. How can we design such an agent able to make judgements, by juggling evidence and emotions, reasons and sentiments? How may we envisage a moral mind? Three directions are possible: 1) With a set of mathematical formulae used to make predictions about behaviour? 2) With a computer program to simulate thinking? or, 3) With a description (operation) of mechanisms that explain observed mental phenomena? Our conjecture is: with a decision apparatus, and following the third trend.

Morality is more than simple utilitarianism or deontology, as some authors defend (Hauser, 2006), focusing on what actions are morally, right or wrong. It is not only a utility function with some devious calculus of importance, because it requests emotional regulation, according to recent findings of Cognitive Neuroscience, and the full cooperation between reason and emotion, at least.

We judge our actions by imagining what the future looks like, and we act because we would like to achieve a purpose, preserving a set of qualities. Imagination is essential to empathy, in order to comprehend the full moral dimension of a situation, and, in point of fact, to be an agent with moral virtues of character it is necessary to have more than general principles of rationality. And, get a direct answer about how the mind works implies to get closer about the description of several mechanisms that explain the mental phenomena at large (Minsky, 2006).

The most successful theoretical explanation in cognitive science has been mechanistic in the sense elucidated by philosophers of science. A mechanism is a system of parts whose interactions produce regular changes. Therefore, the idea of composing the architecture of a moral agent (our proposal in this paper) is debatable, but it allows to design how an agent achieve a variety of conflicting aims (components), such as: deliberation, advance goals and act on commitments that must be revisable; action guided by context-sensitive judgement; ability to be sensitive to the requirements of particular circumstances; emotional connection, or sensitivity of moral concepts (moral attention), imagination and self-reflection (Singh and Minsky, 2005). Such an architecture requires all of the skills we associate with general intelligence and common sense reasoning, namely 1) reasoning ability, ie. making logical inferences, synthesizing and interpreting information, or recognizing similarities and differences; 2) getting vision of the situations, judging and doing accurate predictions; 3) moral perception with intuitive skills of situations embedded in social customs, personal and relation histories (social interactions); 4) correcting and revising power (truth maintenance capacity) in order to guide further/future judgements (Ethics versus centred on wanting something other than what exists); and, 5) emotional intelligence in order to preserve the value of options not acted and to guide the agent through the practical reasoning process.

4 Moral totality

Usually, a moral decision does not follow a single criterion. It requires comparison of different points of view, some in favour and some against, and an algebra to take care of multiple criteria and trade-offs. So, amalgamating the multidimensional aspects of the decision situation into a single scale of measure is no longer the way out. Prudential calculus is made by characters served by a set of qualities, which implies taking into consideration some personalities of an agent to cover the skills akin to morality (personal, cultural, affective, anticipatory).

Decision is often defined as an objective function like a single point of view (profit or cost index) representing the preference (or not) of the considered actions, which is maximized (or minimized). In moral contexts, this is very simplified and unnatural, because any decision is always related to a plurality of points of view, and the pros and cons (relevance) are to be taken in due account. So, it is advisable to make the aggregation of individual preferences into collective ones when choosing, ranking or sorting the actions (solutions, alternative courses…).

Our intuition, about the good choice, is on multiple criteria decision analysis (MCDA) because there are several dimensions involved apart of ethics, it is suitable to structure the complex evaluation and to include both qualitative and quantitative criteria. This choice favours a behaviour that will increase the consistency between the evolution of the process, the objectives and the values. By attending each pertinent point of view separately, independently from the others, it is generally possible to arrive at a clear and common elicitation of preferences regarding the single point considered. This also leads to associating a specific criterion to each point of view.

5 Around the design of a moral agent

The interplay of mentality (cognition), sociality (collective regulation) and morality (norm/value guidance) 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 features?

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 al, 2006) adopted a two-layers structure by mixing an extended BDI with an emotion manager for modelling cautious agents. In order to design social and normative agents (Castelfranchi et al, 2000) defended norms as meta-goals on the agent´s own processes, around a BDI kernel and two levels of process abstraction. In the serious game (mixing human and artificial agents) of water management (Adamatti et al, 2009), any artificial agent had a behavioural profile linked to one or more strategies regarding a certain role (the 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 other serious games on participatory management of protected areas (Briot et al, 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, to be acceptable by users.

A moral agent, as it was defended by Hauser and by Green, is a mix of cognitive and affective capabilities, but no architecture was till today presented as the definite one, despite several design attempts (Wiegel, 2006; Andrighetto et al, 2007), without any explanation on the operation of the moral decision machinery. Several questions needed yet to be answered: What makes a moral (norm-abiding, virtuous, conventional) agent? By what mechanisms and layers can abstract moral principles and values spread or decay from one agent to another (like memes)? How are explicit morals implemented and added to the overall architecture to generate aims and desires, and, later on, to fix conducts? What is the function of moral reasoning, of perceiving a new detail in a situation, or of understanding the moral relevance of what we see? Which is the specific role of the (cognitive, moral, ethical) values?

A moral agent needs to get a more intricate way of thinking 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 correct 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 as 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. See our proposal for the architecture of a moral agent in figure 1.

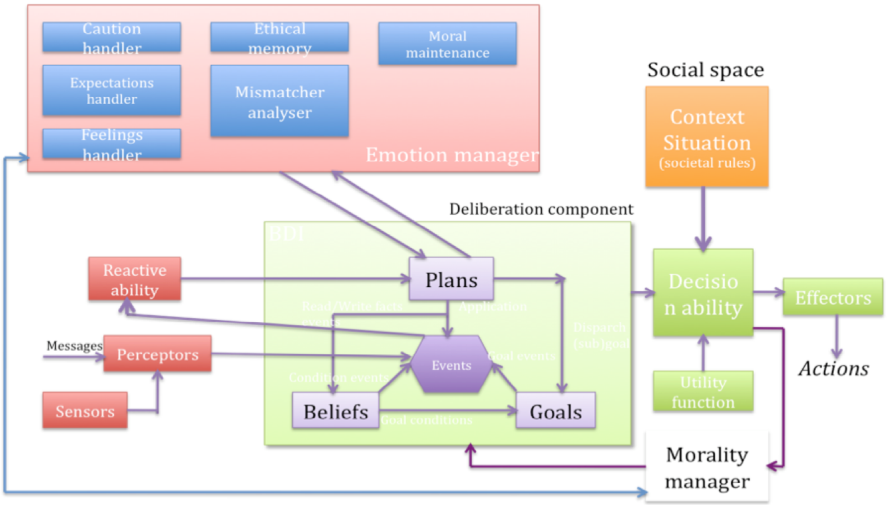

Fig. 1. Proposal of a moral agent kernel architecture.

This tentative proposal of a highly modular and hybrid moral architecture is composed by three layers, as opposed to two of the deliberative normative architecture of (Castelfranchi et al, 2000): 1) the first, for the classical cognitive flow, based upon deliberation (BDI), 2) the second, for the moral system with judgement (upon choices) and decision, a moral maintenance system, an ethical memory, and the morality (including a moral grammar and a moral learning module) manager, and 3) the third one for the emotional system containing the emotion manager (including three handlers for caution, expectations, and feelings, and a mis-matcher analyser). The two managers interact heavily between them and, also, each one with the BDI and decision modules.

The architecture of figure 1 has a high-level (the moral reasoning), mainly concerned with how the agent manages its currently available best options for diverse social situations, ie. how it orchestrates the choices together into a moral coherent behaviour. Such a structure allows the moral agent to be flexible enough in changing social environments and to adapt graciously. And, a low-level (moral reaction): a moral judgement is the consequence of a rational process (based upon moral rules) applied to a certain situation or of a simpler reactive process. The moral agent's decisions are not rigid ones but rather well balanced decisions, weighing preferred options or choices, with the aid of a morality manager (the white box in figure 1). The involved mixture of intuitive and deliberative processes embody also a question of power: who is in charge of the higher or lower levels?

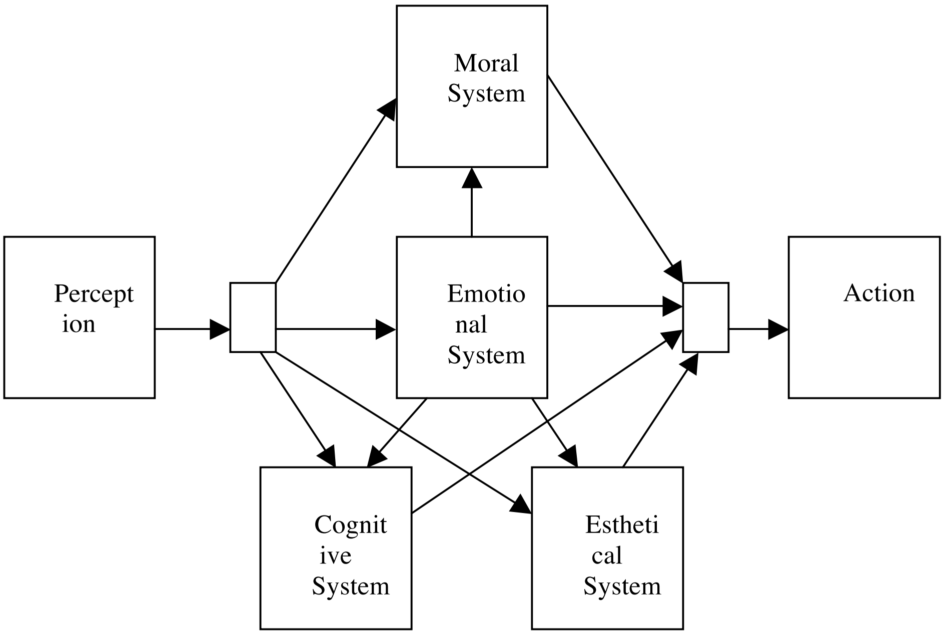

Fig. 2. Interplay of diverse systems

According to the social intuitionist model by (Green and Haidt, 2002), there is more one layer (Esthetics), because moral judgment is similar to esthetical one: when we listen to a story or look to some action/behaviour we get an instantaneous feeling (intuitions with some affective value) of approval/disapproval. In figure 2, we sketch an extended moral architecture with four layers, where the esthetical layer is appropriate to support the agent social reasoning, which involves, as a matter of fact all the four layers (each one with its own objectives and values)! Such an extension aims to model adequately the phenomena behind moral decision working, including the processing of all causal mechanisms.

We adopted an idea from (Dignum et al, 2001) by using desires (self purposes) as links between the cognitive layer and the other ones. Desires are generated by the non cognitive layers and work as factors (to be mixed with the normative factors) and capable to influence the agent deliberation. In (Castelfranchi et al, 2000), norms were mental representations (objects) entering the mental processing and the interaction, in several ways, with beliefs, goals and plans in order to fix the agent´s behaviour. This is also an interesting idea, adopted in the operation of our 4-layer architecture to allow an agent may follow or violate norms. In (Corrêa and Coelho, 1998) we proposed a table of mental states of an agent, facilitating the inclusion of other mental objects (eg. expectations, hopes), and extending easily the BDI classical architecture (Cascalho, 2007) with mental states through attributes (a kind of weights), laws of composition and control mechanisms).

Those desires are relevant to constrain moral judgments. By adopting influence diagrams, we may connect them to judgments by arcs, where each one has a weight according to the rules associated with the agent objective, emotional or cognitive situations. A moral judgment can be positive or negative, depending on deductions made by the moral rules, having a certain intensity/importance given by the sum of the weights of those factors (very high, high, average, low, very low).

Every decision an agent makes, when it comes to choosing between right or wrong, reveals his true character (subjectivity and identity): the Humean model with emotions behind judgements or the Rawlsian model with emotions and reasons after judgements have only two layers, where the main processing flow is done sequentially in one-layer, and the trade-offs are not allowed. In a 4-layer architecture, the interactions among layers, systems and components (eg. emotional vs. moral systems) make the personality of an agent. 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.

An effective decision should be based on the achievement of objectives. Criteria (universal principles, values, beliefs) and objectives (purposes, aims, desires) are used to measure how well we achieve our goals. Decision making is always difficult because trade-offs must be made among competing objectives. In order to consider trade-offs, we must be able to evaluate and measure each aspect of the decision, some quantitative, some qualitative, some very important and some not so important. Uncertainties and competing interests among the components (deliberation, emotion, morality, decision) also add to the complexity of the overall decision making.

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 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 An illustrative experiment

The interplay of cognition, collective regulation and norm/value guidance is better described by an example that justifies the components of our proposal. The usual purpose of a fairy tale (fable) is to provide a context for some general moral interpretation. Although the global message is usually very clear, a deeper reading of some fable details often reveals ambiguity even at the morality level interpretation.

We consider the well-known “Jack and the Beanstalk” fable (1807, British unknown author). The story tells of Jack, a very poor boy, whose lack of common sense exasperates his widowed mother. She sends him to the market to sell their only possession, a cow, but along the way, Jack meets a stranger (adult) who offers to buy the cow for five "magic beans". Jack is thrilled at the prospect of having magic beans, so he makes the deal. When he arrives at home with his beans, his mother, in despair that they were ruined, throws the beans out of the window, and the two go to sleep without eating any supper. The story proceeds with several adventures but, in the end, the boy and his mother get very wealthy because the beans turned out to be really magic.

The story fragment of the "cow for beans' trade" illustrates some interactions between goals, plans, beliefs, desires, social norms and moral values. We named the two agents J and B, respectively, referring to Jack (a child) and the adult owner of the (magic) beans, so we have Ag = { J, B }. The set of available resources may be described by Rs = { cow, beans, food, money }. The "possess" relation, p: Ag → Rs, describes an agent's belongings, thus p (J) = { cow }, and p(B) = { beans, food }. Each agent's goal is described by g: Ag → Rs, therefore g (J) = { money, food } and g(B) = { money }. According to the story, a general plan for each agent may be devised as follows: plan(J) = [get(cow), exchange-for(cow, money), buy(food)] and plan(B) = [get(beans), exchange-for(beans, money_1)].

Additionally, a social norm underlying the whole story is that "an adult always *negotiates honestly* with a child". This norm holds two important concepts: a) the negotiation, and b) the honesty. The "negotiation" calls for utility based reasoning and the "honesty" resorts to the moral interpretation of one's motivations. We know that the utility for a cow is much higher than the utility for five beans, i.e., util(cow) >> util(beans). But, how does the "honesty" concept integrates the overall formulation? One alternative is to interpret "honesty" as a moral evaluation of some subset of the agent beliefs, i.e., moralEval: 2^{Bel} → [0,1], where Bel represents the belief set and 0 (zero) and 1 (one) represent, respectively, the least and the most adherence to the moral principles underlying the corresponding belief subset. Additionally, the "moral signature" of each agent relates each moral concept, e.g., "honesty", with a subset of beliefs, i.e., moralSig: Mc → 2^{Bel} where Mc is the set of moral concepts (within a certain domain). Cleary, this moral signature relation, moralSig, already expresses some of the "moral guides" behind the (human) designer of the relation. Thus, a complex domain requires a (human) designer sensibility and expertise to be tuned in the process of interacting with other (human) designers.

Let us describe agent J as follows: "util(cow) >> util(beans)" ∈ Bel_J, the agent moral evaluation is $\text{moralEval}_\text{J}$({ util(cow) >> util(beans) }) = 0 and its moral signature is moralSig_J(honesty) = { util(cow) >> util(beans) }. So, agent B, after meeting agent J, refines its original plan into a new plan'(B) = [get(beans), exchange-for(beans, cow), exchange-for(cow, money_2)]. From a purely utilitarian perspective, money_2 must be higher than money_1 (above a threshold) in order for agent B to pursue this new more complex plan. But "util(cow) >> util(beans)" ∈ Bel_B so agent B is willing to drop the original plan and adopt the new one (plan'). But, at the same time, the negotiation environmental context includes a child so the "*negotiate honestly*" norm (cf. above) becomes active. Now the agent must apply its moral signature, moralSig, regarding "honesty". The agent uses a machinery to combine the moral evaluation, moralEval, with additional parameters, such as the utilitarian added-value for the new plan. Here, we simplify and decide just upon the moralEval; in this scenario we have $\text{moralEval}_\text{B}$({ util(cow) >> util(beans) }) = 0 so the agent B proceeds and moves to the new plan.

Now, under plan'(B) agent B must find a way to convince agent J that trading a cow for beans is a fair trade. Therefore, B must raise J's expectations regarding the beans, and B believes that a way to raise a child's expectations is to invoke directly its "magic world". Hence a new plan is generated plan''(B) = [get(beans), inform(beans, "magic"), exchange-for(beans, cow), exchange-for(cow, money_2)]. The new plan takes B to convince J to trade its cow for the beans; B gets the cow's money and J makes a plan to get a lot of food and richness from the beans.

This illustrative scenario shows agent B moving forth and back among plans, beliefs, social norms and moral signatures and values. But, if one is not able to follow all the internal reasoning details is it possible to know whether agent B was following norms and moral principles? Let us assume that agent B believes that the beans are really magic and that they would provide a huge fortune to its owner. Then a completely different scenario would arrive and the only difference (from the previous one) could be that "util(cow) << util(beans)" $\in \text{Bel}_B$. In this new scenario agent B exhibits an aesthetically altruistic behaviour. In fact such a high order altruism also resorts for some degree of divine power over disgrace and poorness. But, on the other hand if we were to dissect the child's beliefs and (utilitarian and moral) reasoning we could find that a social norm such as: "trading extremely differently valued assets is not a fair trade" is also active. Usually this is a norm that a child knows about in order to prevent him from bargaining with a much younger child. In this new context the child also ends up bypassing its "fairness" moral signature along with the associated social norm.

The above reasoning scenarios were drawn from a deeper analysis of the internal processes of two agents in the context of an apparently innocuous fairy tale.

7 Conclusions

> "Agents are a way of thinking, a conceptual frame for modelling
> active, distributed, complex, and layered phenomena."
> C. Castelfranchi in IEEE Internet Computing, March-April 2010.

The research and experimentation around the sketch of an architecture for moral agents is supported on the belief that moral decisions are very complex processes. 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. We invested, for more than a decade, in heavy experimentation about agent models and architectures, for individual and collective decision making (large scale disasters, electric energy markets, semantic web spaces), trying in each step forward, to increase the number of interactions and relations among components of the next architecture.

The character of a moral agent is dependent on its architecture, namely on the interactions (for negotiation) and on the relations (global complexity) among its components. Any architecture reveals also a mix of high level (deliberative, moral reasoning) and low level (reactive, intuitive) processes, where some one of them is in power to support the acting.

Our agent design ideas were based on the understanding of the semantic operation of morality, rethinking computing and knowledge in terms of interaction and social processing, but several open questions frame still our current research: How can we operationally verify all the interactions behind a moral agent architecture? How do actors produce and are at the same time a product of social reality? 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 modelling 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.

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.

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 Areas: The SimParc Project, Proceedings of XXVIIIth Congress of Computation Brazilian Society (CSBC´08), Belém, Brazil, July 2008.

Cascalho, J. The Role of Attributes for Mental States Architectures, PhD Thesis (in Portuguese), University of Açores, 2007.

Castelfranchi, C., Dignum, F, Jonker, C. M. and Treur, J. Deliberative Normative Agents: Principles and Architectures, Proceedings of 6th ATAL Conference (1999), Intelligent Agents VI, Springer LNCS 1757, 2000.

Castelfranchi, C., Falcone, R. and Piunti, M. Agents with Anticipatory Behaviours: To Be Cautious in a Risky Environment, ECAI, 2006.

Coelho, H. and Costa, A. R. On the Intelligence of Moral Agency, Proceedings of the Encontro Português de Inteligência Artificial (EPIA-2009), October 12-15 Aveiro (Portugal), in L. S. Lopes, N. Lau, P. Mariano e L. M. Rocha (eds.), New Trends in Artificial Intelligence, pp. 439-450, 2009.

Coelho, H., Costa, A. R. and Trigo, P. On the Complexity of Moral Decision, FCUL and DI Working Report, 2010.

Corrêa, M. and Coelho, H. From Mental States and Architectures to Agents´ Programming, Proc. Of the 7th Iberoamerican Congress on Artificial Intelligence

(IBERAMIA98), Lisbon 6-9, Springer-Verlag LNAI 1484, pp. 64-85, 1998.

Corrêa, M. and Coelho, H. Abstract Mental Descriptions for Agent Design, Intelligent Decision Technologies (IDT), an International Journal, IOS Press, 2010.

Costa, A. R. and Dimuro, G. Moral Values and the Structural Loop (Revisiting Piaget´s Model of Normative Agents), PUC Pelotas Working Report, 2009.

Dignum, F. Kinny, D. and Sonenberg, L. From Desires, Obligations and Norms to Goals, Utrecht University, 2001.

Green, J. and Haidt, J. How (and Where) does Moral Judgment Work? In Trends in Cognitive Sciences, Academic Press Volume 6, Issue 12, December 2002.

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

March, J. G. and Olsen, J. P. The Logic of Appropriateness, Arena Centre for European Studies Working Papers WP 04/09, University of Oslo, 2009.

Mascarenhas, S. F. Creating Social and Cultural Agents, IST MS.C. Thesis, 2009.

Minsky, M. The Emotion Machine, Simon & Schuster, 2006.

Trigo, P. and Coelho, H. Decision Making with Hybrid Models: The Case of Individual and Collective Motivations, Proceedings of the EPIA-07 International Conference (New Trends in Artificial Intelligence), pp. 669-680, Guimarães, 2007.

Trigo, P. and Coelho, H. Decisions with Multiple Simultaneous Goals and Uncertain Causal Effects, in Artificial Intelligence in Theory and Practice II, IFIP Volume 276, Springer-Verlag, pp. 13-22, 2008.

Trigo, P. and Coelho, H. Simulating a Multi-Agent Electricity Market, in Proceedings of the 1st Brazilian Workshop on Social Simulation (BWSS-08/SBIA-08), Bahia, October 26-30, 2008.

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