Agents with a Moral Dimension

(Doctoral Consortium)

Cristina Battaglino
Dipartimento di Informatica
Corso Svizzera 185, Universitá degli Studi di Torino, Italy
battagli@di.unito.it

ABSTRACT

As argued by [9], moral decision making entails considering alternatives and assessing the pros and cons of their possible consequences for self and others. From the area of affective neuroscience the concept of moral emotions has been introduced [9] and neurobiological findings [7] show that moral emotions are used to judge the adequacy of actions and are central to moral behavior, decision making and learning. My aim is to build a computational model for moral emotions in order to enable intelligent agents [2] to understand moral consequences of actions through moral emotions. The agent is able to compare alternative scenarios and to decide what course of actions and goals to pursue in order to show a morally driven behavior. Moral emotions are useful when the agent is engaged in a social interaction with a user or other agents, because (i) moral emotions may lead the agent towards the compliance with (shared) moral values (ii) the agent is equipped with moral emotions which make her potentially emphatic to others.

Categories and Subject Descriptors

I.2.m [**Artificial Intelligence**]: Miscellaneous

General Terms

Languages, Theory

Keywords

moral emotions, moral values, intelligent emphatic agents

1. MORAL EMOTIONS

Based on [9, 13] we argue that moral emotions are complex emotions involving cognitive processes. Given [9], we identify the following moral emotions : Pride, Self-reproach, Reproach, Admiration, Gratification, Gratitude, Anger and Remorse. In the field of Artificial Intelligence, we observed an increased interest in studying computational models of emotions; many computational models have been modeled [12], but few of them take into consideration moral emotions. EMA (EMotion and Adaptation) [10] considers Joy, Distress, Guilt, Anger, Hope and Fear as emotions. FLAME [8] (Fuzzy Logic Adaptive Model of Emotions) and EM (Emotion Model) [14] take into consideration moral emotions but they don't provide a independent process to derive them and don't enable the agent to understand the moral consequences of her actions. For example, in FLAME, the problem of deriving the evaluation of actions is overcome through the user feedback on actions via a learning algorithm.

In [4] we defined an independent approach to derive moral emotions. Our work is based on appraisal theories, in which emotions arise from the evaluation of a situation according to some appraisal variables defined in the theory itself. The evaluation of actions is a core concept for the generation of moral emotions. In our model, the agent evaluates the moral consequences of her actions with respect to moral values. Moral values are enclosed in a BDI agent [2]: our agent features a set of moral values organized in a scale of values and she is constrained to respect them [16]. Each value is associated with a set of conditions; when one of the conditions holds in the state of the world the value is put at stake, otherwise the value is in an equilibrium state. An action is a *right moral action* if it re-establishes a value at stake while it is a *bad moral actions* if it puts at stake a value. Through the independent appraisal of actions, we enabled the agent to have a moral dimension and to understand the moral consequences of her actions (and other agents' actions) in a domain-independent way. Note that, we related the praiseworthiness and blameworthiness of an action with the compliance with moral values.

2. MORAL VALUES AND CONFLICTS

In [1] we fashioned an architecture for virtual characters able to face off a moral dilemma. Values are used as a motivation for the agent: when the agent realizes that one of her values is put at stake, she forms a value-dependent goal with the aim of re-establishing the value at stake. In the reasoning cycle, during a phase called *deliberation phase*, the agent has to choose what goal she wants to commit to. We based the agent's choice on expected emotional reward utility of plans that the agent forms to achieve her goals. We only considered Pride and Self-reproach in the deliberation phase and we left out the generation of emotions because the focus is on representing and detecting moral dilemmas. When the character is in a moral dilemma [11], she has to choose between different values that are incompatible to each other and one of them must be sacrificed. We explicated the concept of moral dilemma as a *chiasmus conflict* between plans: two plans π_1 and π_2 are in conflict if the plan π_1 re-establishes a value v_i and puts at stake a value v_j while the plan π_2 re-

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

establishes a value v_j and puts at stake a value v_i. When two plans are in conflict, the probability of bringing the value in an equilibrium state is set to zero, due to the fact that they are incompatible to each other [11, 16] and the agent feels strong emotions.
In [6] we extended the architecture in order to include individual selfish goals (not only value-dependent goals as in [1]) and a generation phase for emotions. We also defined independent rules for the generation of Joy and Distress emotions, necessary to model moral compound emotions. Following the work in [14], we based the evaluation of events on goal processing. Given the agent's goals, the expected reward of the plans, devised to achieve them, is calculated by taking into consideration the potential for Shame, Pride, Joy and Distress emotions. The expected reward utility of plans is based on a conflict between plans: a plan can achieve a goal g_i, threaten another goal g_j, re-establish a value at stake v_n or put at stake a value v_m. Conflicts are not modeled in an explicit way as in [1], they are hard-wired in plans. Consequently, the agent doesn't detect moral dilemmas in an explicit way. After the agent chooses the goal she wants to commit to, she executes one action of the plan, then she monitors the world to perform the emotional appraisal according to domain-independent rules defined in the model.

3. FUTURE WORK AND IMPROVEMENT

By now, the model starts to sound implementable and we intend to evaluate the generation of moral emotions with a real implementation. My next steps are: (1) unify the work on moral dilemmas [1] to include a conflict detection phase also in the agent's reasoning cycle presented in [6] so that the agent is able to detect conflicts between plans, to understand the consequences of her actions and to cope with moral conflicts through emotion deliberation; (2) choose an application domain to evaluate the model and, consequently, a method of evaluation; (3) include mood and decay functions of emotions and (4) implement a consequent model of emotions to cope with moral conflicts.
Conflict detection between plans has a high computational cost and we assume a continuous planner in which abstract actions are considered and detailed out during the execution of the plan as in [3]. This solution allows simplifying the task of prospect reasoning, but comparison between plans to detect conflicts as in [1] still remains a hard task. A solution can be inspired by works on conflict between plans [15] and goals [17]. For example, we can attach a procedural and declarative knowledge to goals as in [17] or use external procedures, called Semantic Attachment, in order to compute the valuations of state variables at planner run-time as in [5].
In conclusion, my thesis aims to develop agents with a moral dimension that are able to deal with moral conflicts and to participate in dynamic social environments, in which agents can be engaged in social interactions, showing a proper moral dimension that drives their behavior. For example, such a model can be employed in an empathetic virtual agent that interacts with the user to understand what action can be appraised as beneficial (or harmful) by the user. Finally, the model can be employed in interactive entertainment applications, in which the engagement of the user is increased through moral dilemmas.

4. REFERENCES

[1] C. Battaglino and R. Damiano. Emotional appraisal of moral dilemma in characters. In *Proc. of the 5th int. conf. on Interactive Storytelling*, ICIDS'12, pages 150–161, Berlin, Heidelberg, 2012. Springer-Verlag.

[2] M.E. Bratman. *Intention, plans, and practical reason.* Harvard University Press, Cambridge Mass, 1987.

[3] Michael Brenner and Bernhard Nebel. Continual planning and acting in dynamic multiagent environments. In *Proc. of the 2006 int. symp. on Practical cognitive agents and robots*, PCAR '06, pages 15–26, New York, NY, USA, 2006. ACM.

[4] L. Lesmo C. Battaglino, R. Damiano. Moral appraisal and emotions. In *Workshop EEA - Emotional and Empathic Agents, AAMAS*, AAMAS '13, 2012.

[5] T. Keller S. Trüg M. Brenner C. Dornhege, P. Eyerich and B. Nebel. Semantic attachments for domain-independent planning systems. In *in Proc. of ICAPS, 2009.*

[6] L. Lesmo C.Battaglino, R. Damiano. Emotional range in value-sensitive deliberation. In *Autonomous Agents and Multi-Agent Systems*, 2013.

[7] A. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain.* Harper Perennial, 1995.

[8] Magy Seif El-Nasr, John Yen, and Thomas R. Ioerger. Flame (fuzzy logic adaptive model of emotions). *Autonomous Agents and Multi-Agent Systems*, 3:219–257, September 2000.

[9] Velez Garcia, Alicia, Ostroskysolis, and Feggy. From morality to moral emotions. *International Journal of Psychology*, 41(5):348–354, October 2006.

[10] Jonathan Gratch and Stacy C. Marsella. A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.

[11] John F. Horty. Moral dilemmas and nonmonotonic logic, 1994.

[12] Stacy C. Marsella, Jonathan Gratch, and Paola Petta. Computational models of emotion. In *A blueprint for an affectively competent agent.* Oxford University Press, Oxford, 2010.

[13] A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions.* Cambrigde University Press, 1988.

[14] W. Scott Reilly and Joseph Bates. Building emotional agents, 1992.

[15] John Thangarajah, Lin Padgham, and Michael Winikoff. Detecting avoiding interference between goals in intelligent agents. In *Proc. of the Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pages 721–726. Academic Press, 2003.

[16] B. van Fraassen. Values and the heart's command. *Journal of Philosophy*, 70(1):5–19, 1973.

[17] M. Birna van Riemsdijk, Mehdi Dastani, and John-Jules Ch. Meyer. Goals in conflict: semantic foundations of goals in agent programming. *Autonomous Agents and Multi-Agent Systems*, 18(3):471–500, 2009.