IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Emotional range in value-sensitive deliberation

This is the author's manuscript

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/147231 since 2016-11-30T12:15:38Z

Publisher:

IFAAMAS

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

24 April 2024

Emotional Range in Value-sensitive Deliberation

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

Rossana Damiano
Dipartimento di Informatica
Corso Svizzera 185,
Università degli Studi di
Torino, Italy
rossana@di.unito.it

Leonardo Lesmo
Dipartimento di Informatica
Corso Svizzera 185,
Università degli Studi di
Torino, Italy
lesmo@di.unito.it

ABSTRACT

This paper presents a model of an agent's behavior that takes into account emotions and moral values. In our proposal, when the description of the current situation reveals that a moral value is 'at stake', the moral goal of re-establishing the threatened value is included among the active goals. The compliance with values generates positive emotions like pride and admiration, while the opposite brings to shame and self-reproach.

During the deliberation phase, the agent appraises her plans in terms of the emotional reward they are expected to yield, given the trade off between moral and individual goals. In this phase, the emotional reward affects the agent's choices about her behavior. After the execution phase, one's and others' actions are appraised again in terms of the agent's values, giving rise to moral emotions.

The paper shows how emotional appraisal can be coupled with the choice among possible lines of action, presenting a mapping between plans and emotions that integrates and extends preceding proposals.

Categories and Subject Descriptors

I.2.m [**ARTIFICIAL INTELLIGENCE**]: Miscellaneous; I.2.1 [**ARTIFICIAL INTELLIGENCE**]: Knowledge Representation Formalisms and Methods

General Terms

Languages, Theory

Keywords

emotions, moral values, empathic agents

1. INTRODUCTION

Notwithstanding the efforts put by scholars in trying to understand the complex mechanisms underlying moral reasoning, in fields that range from agent theories [14, 9] and cognitive science [11] to deontic logics [21], most approaches fall short of explaining the complex interplay of moral behavior and emotions. This interplay becomes very relevant when it comes to designing virtual agents who must display a believable emotional behavior. Although an artificial agent may be programmed to always choose the action that achieves the right goal in a moral sense, the capability of reasoning on moral values is important for designing virtual agents who act as companions, counsellors, and so on. More importantly, for such agents, it is important that they also have the capability of understanding the emotions that accompany moral choices, since they need to show empathy with their human partners. According to Susan Stark, in fact, "it has been argued that emotions play a crucial role in moral epistemology and that agents simply cannot have complete and accurate moral perception and cannot make reliable moral judgment absent emotional engagement with moral situations." [31] (p. 355).

The relevance of a moral dimension in the implementation of autonomous agents was highlighted in [1]. In that paper, it was also pointed out that it is important to have 'explicit' models of ethical behavior, where the behavior of the agent is driven by a declarative representation of moral values. The alternative is the use of implicit models where the moral behavior is hard-wired in the implementation. The advantage of an explicit model is that the agent can reason about its values and choose the best line of behavior in a flexible and context-dependent way. The claim of this paper is that moral principles affect the choice of actions in a way that respects the general organization of an autonomous agent. There is a complex interplay between the agent's individual goals and her moral values, which mainly derive from the group or society she belongs to. In this perspective, a paradigmatic situation is given by moral dilemmas, which are a particular case of the issue of choosing the best action in an environment where the agent can have multiple, incompatible goals [2].

Another reason for adopting an explicit model of ethical behavior is that it enables us to model the relationship between ethically-driven actions and emotions. According to [12], the desire to attain positive emotions constitutes one of the strongest forces that produce behavior. In case of decisional oppositions, an agent's tendency to preserve her moral well-being (often perceived as a value per-se in everyday life) can help her to choose a line of behavior, especially when the opposition concerns individual goals and moral concerns. Consider, for example, two agents, Max and Mary: Max wants to have a chocolate candy, but the only

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.*
Copyright © 2013, International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

way to satisfy this desire is to steal Mary's candy. If honesty, for Max, is more important than his hunger for chocolate, he will drop the goal; otherwise, he will steal the chocolate candy. But, if his desire is very strong, Max may eventually steal the chocolate candy despite his strong sense of honesty, (mis)lead by the anticipation of the pleasure he would feel by tasting the delicious taste of the chocolate.

In this paper, we pursue the following research goals:

- Integrating the notion of moral values in agent deliberation, in a way that be compatible with practical architectures; as shown by the example above, the goal of this model is not to build an idealised model of moral deliberation, but to provide a way to simulate practical reasoning about values.
- Acknowledging the role of emotions during and after the process of value-sensitive deliberation; the model grasps the role of emotions in the choice among competing (and incompatible) moral alternatives, allowing an agent to "feel" and express the emotional range that accompanies moral deliberation.

The paper is structured in the following way: after surveying the related work, in Section 3, we illustrate the background assumptions and motivations for our model, described in Section 4. In Section 5, we propose an agent architecture based on this model. Discussion and conclusions end the paper.

2. RELATED WORK

Regarding the field of computational model of emotions, many works tried to integrate an emotional component in a cognitive architecture for intelligent agents [29, 16, 24]. Although different theories of emotions have been proposed (including physiological and dimensional models), most computational models are based on appraisal theories, in which cognitive processes are involved in the generation of emotions [26, 23, 30]. According to appraisal theories, cognitive processes have the function of building a mental representation of the situation in which a person is involved (following [25] we call this representation "person-environment relation"). Emotions arise from appraisal of the person-environment relation according to appraisal dimensions that are defined in the theory (i.e. desirability of an event, likelihood of an event, causal attribution for the agent, etc.). Different computational models refer to different appraisal theories. For example, many computational models [16, 29, 15, 17] are inspired by the OCC (Ortony Clore Collins) model [26], while Lazarus's [23] and Scherer's theory [30] are implemented respectively by [18] and [3]. In the following, for the sake of brevity, we compare computational models of emotions chiefly on the basis of which appraisal derivation and affect derivation model are used.

The appraisal derivation model specifies how the appraisal variables are derived from the representation of the person-enviroment relation [25]. Then, the appraisal variables are used by the affect derivation model to generate the emotional response for the agents. Most computational models use domain-independent rules to construe the emotional response in affect-derivation model [18, 3, 29, 15, 17] but few computational models are based on an independent approach to derive the appraisal variables in the appraisal derivation model, the major are: EM (Emotion Model) [29], FLAME (Fuzzy Logic Adaptive Model of Emotions) [15] and EMA (EMotion and Adaptation) [18].

EM [29] employs the goal processing to assess appraisal variables in an independent way to evaluate outcomes of events (i.e. the goal failure and success determine the evaluation of an event as desiderable or undesiderable). But, regarding the evaluation of actions, the appraisal derivation model is not domain-independent and it is limited to assign credit or blame to the success or failure of goals. As noted by the authors in [29], this is not feasible in order to model all possible evaluations of actions with respect to moral values (standards). Besides, goals must be decoupled from the evaluation of actions, in order to make possible the generation of emotions regarding only the evaluation of actions (i.e. Pride, Shame, Admiration and Reproach). In FLAME [15] appraisal and affect derivation model are based on domain-independent fuzzy rules, but, FLAME uses an heuristic approach based on user feedback to model the evaluation of actions. The user can perform in the system action as "bad agent" or "good agent" so the agent can associate a value "v" to actions that represent the goodness level of the action. Moral values are not modeled in an esplicit and, although FLAME uses an independent approach in the appraisal derivation, the agent is not endowed with a inner model that she can adopt to evaluate the moral consequences of her actions. In EMA [18], inspired by Lazarus's appraisal theory [23], appraisal is formed by a set of independent processes that operate on a plan-based representation of person-enviroment relation, named *causal interpretation*. With respect to moral emotions, EMA models only Anger and Guilt [20] and doesn't encompass other moral emotions. The authors argue that modeling moral values, as in the OCC model, is an approach too constrained and only credit or blame are assigned to actions as positive and negative utility over states. But, as authors noted in [18] an explicit representation of moral values is a requirement to make distinctions between certain emotional states (e.g. guilt from shame) and to inform dialogue or coping strategies.

Being posited at the junction of values and emotions, moral emotions have received less attention than goal–related ones. We believe that the notion of value can be an effective tool to model emotions characterized by a prominently moral nature. The design of an agent who reacts to moral values requires the interaction of a rational and an emotional component. Both components should embed some notion of moral value, to let the agent reason about the compliance with values at the behavioral level, and feel moral emotions [20] when her values are put at stake or re-established.

We avail ourselves of the OCC model (as EM [29] and FLAME [15] did) and, inspired by EMA [18], our model take into account syntactic information of the plan-based representation. Also, we adopt the goal processing approach of EM [29] to assess some appraisal variables as "desirability".

With respect to other computational models, we model moral values in an explicit way and provide domain independent rules to encompass the appraisal derivation of actions with respect to moral values. We rely on our previous work [10] to establish an explicit link between moral values, moral emotions and the appraisal of action with domain independent rules based on moral values. Further, we use emotions in the deliberation phase of the agent.

3. BACKGROUND AND MOTIVATIONS

In the OCC theory [26], the person-environment relation is represented by goals, standards and attitudes; appraisal dimensions are represented by *desirability* (or undesirability) of an event, *praiseworthiness* (or blameworthiness) of an action, *liking* (or disliking) of an object. In OCC, emotions are divided in four basic classes:

- *Event-based* emotions, that arise from reactions to events (i.e. being pleased (or displeased) about the event with respect to goals).
- *Attribution* emotions, that arise from reactions to agents (i.e. approval (or disapproval) of an action performed by an agent with respect to standards).
- *Attraction* emotions, that arise from reactions to objects (i.e. liking (or disliking) of an object with respect to attitudes).
- *Compound* emotions, that arise emotions that arise when the same situation is appraised at the same time as an action and an event

According to this model, the agent's "standards" (i.e. the agent's "beliefs in term of which moral and other kinds of judgmental evaluations are made") affect the evaluation of self and others' actions. While in the OCC model goals are conceived as states of affairs that one wants to obtain, standards concern the state of affairs that one believes she ought to obtain. Standards represent the beliefs in terms of which moral and other kinds of judgmental evaluations are made, such as *you ought to have tried harder* or *you ought not to do things that upset other people* citeOrtony88. Actions that meet the agent's standards are deemed praiseworthy, and their execution triggers emotions like pride and admiration. Conversely, blameworthy actions trigger emotions like shame and reproach.

Following [10], we define the *praiseworthiness* of an action on the basis of the goal that motivates the action itself: an action is praiseworthy if it is motivated by a value–dependent goal, and the value the goal depends on is acknowledged as such by the appraising agent. In other words, the praiseworthiness is not an intrinsic property of the action, but resides in the motivations that determine the agent's intention to execute it, i.e., the commitment to a goal. The role of values is relevant not only for the appraisal of an agent's own actions, but also for the appraisal of other agents' behavior. When the appraising agent differs from the agent of the appraised action, the appraising agent evaluates the other agent's actions according to her own values, praising that action only if she can ascribe to the other agent the value–dependent goal to re–establish a value at stake (that she shares with the other agent). Conversely, the *blameworthiness* of an action is defined on the basis of the effects it brings about in the state of the world. If an action puts at stake a value of the appraising agent, it is considered blameworthy, independently of the motivation of the action's agent to execute it.

We also consider the four compound emotions that are characterized by the conjunction of Attribution emotions and Event-based emotions, like Joy and Distress.

Following an established line in emotion modeling [29], the desirability of an event can be defined with respect to its consequences for the appraising agent's goals. If an event brings about a state of affairs in which a goal of the appraising agent is satisfied, the event is desirable; the event is undesirable if its effects are in conflict with the satisfaction of a goal of appraising agent.

Emotional appraisal plays a twofold role in our agent architecture and it is split into two distinct phases: in the deliberation phase, the agent feels "anticipatory" emotions, which help her trade off her individual goals with the goals that she has formed as an effect of her values; after the execution phase, the agent updates her beliefs and feels a certain emotion according to the emotional appraisal of the obtained state of the world.
Similarly to anticipatory reasoning on actions, which allows agents to envisage the consequences of their possible behaviors, anticipatory emotional appraisal allows the agent to choose a line of behavior in the light of the emotional states it would determine in the agent (*emotional reward*). Since the OCC model acknowledges a distinction between positive and negative emotional states, the agent will tend to prefer the lines of behavior that are more likely to make positive emotional stases arise in her. Attribution emotions work in favour of the compliance with values; conversely, Well-being emotions tend to privilege individual goals, but the two types of emotions, each of which encompasses a positive and a negative polarity, compensate each other in a way that depends on the context (subjective values and beliefs, available plans, etc.).

In particular, moral appraisal requires the agent to account not only for the intrinsic morality of her goals, but for the notion of 'responsibility' that is related with the side effects of one's behavior [7]. Since an agent is held responsible not only for the intended effects of her actions, but also for the unwanted effects which negatively affect her goals and values (including those of the society she is part of), it is necessary for the agent to assess the compliance with values not only by examining her own goals, but also by considering the practical actions she may perform to achieve them. So, in the model we propose, emotional appraisal is conducted by expanding the agent's goals into the plans (to a limited degree, for cognitive and computational reasons) and by assessing the consequences of these plans on the agent's goals and values in order to foresee possible interference. Notice that this mechanism is necessary to deal with the paradigmatic case of moral dilemmas, where an agent is faced with values which, although not ideally in contrast, in the given context mutually exclude each other since any possible way to attain one makes impossible the attainment of the other [2].

Since the model of emotional appraisal we propose relies on the evaluation of the plans the agent forms to achieve her goals (both individual and value–based ones), some relevant qualities of these plans are accounted for, such as the *probability of success*, the effort *required* and their *importance* for the achievement of the agent's goals. We use the syntactic structure of plans to derive the affect intensity model. For example, the importance of goal is multiplied with the probability of success to calculate the intensity of goal-based emotions. The effort (i.e. the cost of the plan) is a measure

that reduces the intensity of goal-based emotions.

Finally, after the execution of actions, the agent monitors the environment and updates her emotional states by assessing the actual consequences of her own actions, and the changes that have spontaneously occurred or have been brought about by other agents. While in anticipatory appraisal only the rules for Well-being emotions and attribution emotions are applied, in this phase Event–based emotions are also considered.

We claim that, when the agent translates her goals into practical lines of behavior, the projection of these lines of behavior must also encompass the evaluation of the agent's own emotional states, such as shame or pride, that contribute to orientate the agent's choice towards value-compliant courses of actions. The advantage of this integration is that the agent not only forms her goals based on the compliance with her values, but moral emotions become relevant when conflicting goals (and plans) are formed and the compliance with values (or with one's own private goals) must be traded off.

4. VALUE-SENSITIVE AGENT MODEL

In a previous work by [13], a BDI agent is extended with the notion of values and emotions. In this paper we further extend the agent model by defining an agent as a pair *(MS,E)* where MS is the agent mental state and E is the agent emotional state. The mental state of an agent includes the agent's Beliefs B, a set of Goals G, a set of Values V and a set of Values at stake $V_{atStake}$ (Fig. 2).
The beliefs base B is a collection of literals in traditional logic programming style.

The agent's values are organized in a scale of values [32]. A value $v \in V$ is represented by a tuple $v(r,d,Vc)$: r is a real number with range $[-1,1]$ and denotes the importance of the value, d is a real number with range $[-1,1]$ and represents the degree with which the value is shared with the society [2], Vc is the set of violation conditions. If one or more conditions $vc_i \in Vc$ hold in the agent's beliefs the value is put at stake.

Following the approach described in [33], a goal g is defined by a tuple $g(Ac, Sc, Fc, S, \Pi)$. Ac is the set of adoption conditions: when one or more conditions $ac_i \in Ac$ hold in the beliefs base, the goal (which usually is in a sleeping state) is adopted by the agent (i.e. the goal becomes a desire, following the BDI model). Sc is the set of success conditions: when one or more conditions $sc_i \in Sc$ hold in the belief base, the goal is achieved and dropped. Fc is the set of failure conditions: when one or more conditions $fc_i \in Fc$ hold in the belief base, the goal is unachievable (so it is dropped). The success and failure conditions are useful for decoupling goals and plans. A plan can fail, but if the failure condition is not true in the agent's beliefs the goal is not dropped, since the agent can find another plan to reach the goal. S is the goal state (Fig.1). A goal is in an adopted state when the agent considers it as an available options. An adopted goal can be in a suspended state or in an active state. Being in an active state means that the agent's focus is on the active goals[1]. A dropped goal can be achieved or unachieved, while a goal is in a sleeping state when it is not in the agent's options. Finally, Π is the set of plans associated with the goal.
A goal can be an individual or a moral goal. The set of individual goals is named G_I, while the set of moral goals is named G_S. We consider a moral goal as a value-dependent goal with a general adoption, success and failure conditions related to the re-establishment of a value (i.e. a moral goal is in the form $re\text{-}establish(v_x)$).

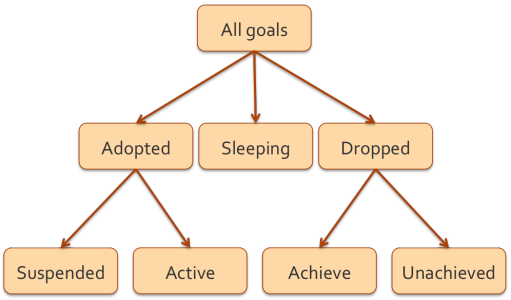

Figure 1: Taxonomy of goals states

The emotional state is represented as the set of current emotions E that the agent feels(Fig. 2). An emotion is represented by a tuple *e(Type, Int, Obj, Ag)*: *Type* is the emotion category according to OCC model, *Int* is the intensity of emotion (a real number), *Obj* is the cause of emotion (i.e. the goal or value from which emotion rises), Ag identifies who is the target of the emotion (self or another agent).

Finally, the agent has a means-ends reasoning capability. A plan is represented by a tuple $\pi(G, T, u)$ where G is the set of goals that the plan may look for, T is the hierarchical structure decomposition of the plan and u is the emotion-based plan utility. We adopt the paradigm of probabilistic propositional planning [5] in which operators are specified in an extended STRIPS-like notation called PSO (Probabilistic STRIPS Operator) [22]. A classical STRIPS operator is defined by precondition and a set of effects. The former identifies the set of states in which the action can be executed, and the latter describes how the environment changes as a result of taking the action. A PSO operator associates actions to stochastic effects, a list of variable values with a probability attached. Using a planner as Buridan [22] we can calculate the probability that a plan reaches a goal state.

5. MORAL EMOTIONS IN DELIBERATION

In (Fig. 2) the agent's architecture is depicted with her mental state described in Section 4. In addition, the agent maintains in her memory the optimum plan π_{opt} and the set *UASet* (*Updates Action Set*) formed by the pairs (*updates, action*). The element *updates* identifies a set of changes in the belief base B. Every update can be either an addition or a deletion of one of the conditions associated to a goal (i.e. Ac, Sc, Fc) or an addition or deletion of one of the conditions Vc associated to a value. The element *action* is the action that caused the change (it may be empty if the agent

[1]An high-level goal and its subgoals can be active at the same time. This is not a problem, because in practical architectures a structure is maintained to keep the high-level goals in relation to their subgoals.

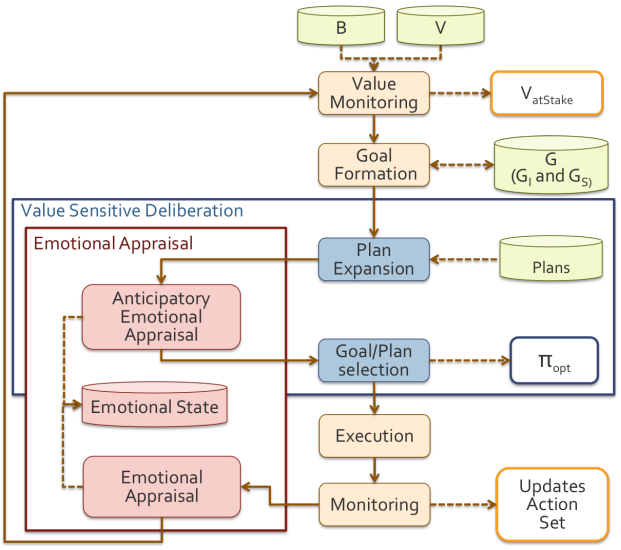

Figure 2: The agent's architecture

has no knowledge of the nature of the update, i.e. the agent could not know the cause of the update). Note that using a data structure to trace the changes in the beliefs base is in line with current architecture for BDI agent as Jadex [28] and Jason [4]. In the following we describe the agent's reasoning cycle depicted in (Fig. 2). Before the agent starts the reasoning cycle, she observes the **Initialize** phase in which she perceives the enviroment and updates her belief base according to what she has perceived

Value Monitoring: if a condition vc of a value $v_i \in V$ holds in B then the value is at stake and added to the list.

Goal Formation: every individual goal in the *sleeping* set is checked to verify if any of its adopting conditions ac holds in the belief base. If so, it becomes an adopted and suspended goal. Then, zero or more moral goals g_i motivated by the values at stake v_i are adopted by the agent.

Value-sensitive Deliberation: The agent starts to find plans in order to satisfy her goals. For every goal $g_i(Ac, Sc, Fc, S, \Pi) \in G$, each plan $\pi_{i,k}$ in Π is expanded[2]. According to BDI model [6], the agent has to choose the goals to commit to (which becomes active). This phase is based on anticipatory emotional appraisal and value-sensitive deliberation (see *Anticipatory Emotional Appraisal* and *Goal PlanSelection* in Fig. 2): the agent performs a prospect reasoning to detect conflicts between goals and values. Then, she calculates the emotional reward for every goal $g_i \in G$ and for each plan $\pi_{i,k} \in \Pi$. Joy, Distress, Pride and Self-reproach are the emotions considered in the deliberation process.

Anticipatory Emotional appraisal: In order to evaluate the emotional reward utility, we make use of the following functions: I, P and E. The I function returns the importance of a goal. The function P returns the probability of success of the plan, while the function E returns the cost of the plan. The expected emotional reward (EER) of a plan $\pi_{i,k}$ is based on the conflict between goals and values [2] that the agent derives by the prospect reasoning. A plan can achieve one or more individual goals but, at the same time, can make some other adopted individual goals unachievable. Further, a plan $\pi_{i,k}$ can re-establish a value at stake $v_i \in V_{atStake}$ and at the same time can threaten another value $v_j \in V$. As a consequence, the emotional reward derives from: (1) the intensity of the joy that the agent feels if she reachs the individual goal $g_i \in G_I$ through the plan $\pi_{i,k}$; (2) the distress intensity that the agent feels if, executing the plan $\pi_{i,k}$, some other adopted goals $\in G_I$ has become unachieved; (3) the pride intensity that the agent feels if she re-establishes a value at stake v_i through the plan and reaches the related moral goal $g_s \in G_S$; (4) the self-reproach intensity that the agent feels if the plan $\pi_{i,k}$ threatens some other values v_j *in* V. Given a plan $\pi_{i,k}$, we noted with G_A the set of individual goals satisfied by the plan, with G_T the set of individual goals threatened, with V_B the set of values re-established and with V_T the set of values put at stake. The intensity of anticipatory emotions Joy $EER_J(G_A, \pi_{i,k})$, Distress $EER_D(G_T, \pi_{i,k})$, Pride $EER_P(V_B, \pi_{i,k})$and Self-reproach $EER_{SR}(V_T, \pi_{i,k})$) are:

Algorithm 1 Anticipatory Emotional Appraisal

```
π_opt = nil;
u_opt = minPossibleValue;
for all g_i ∈ G do
  for all pi_{i,k} ∈ Plans do
    u(π_{i,k}) = calculateUtility(π_{i,k});
    if u(π_{i,k}) > u_opt then
      u_opt = u(π_{i,k});
      π_opt = π_{i,k};
    end if
  end for
end for
return π_opt;
```

$$EER_J(G_A, \pi_{i,k}) = \frac{P(\pi_{i,k}) * \sum_{g_a \in G_A} I(g_a)}{E(\pi_{i,k})} \quad (1)$$

$$EER_D(G_T, \pi_{i,k}) = \frac{P(\pi_{i,k}) * \sum_{g_t \in G_T} (I(g_t))}{E(\pi_{i,k})} \quad (2)$$

$$EER_P(V_B, \pi_{i,k}) = \frac{P(\pi_{i,k}) * \sum_{v_b \in V_B} (r(v_b) + d(v_b))}{E(\pi_{i,k})} \quad (3)$$

$$EER_{SR}(V_T, \pi_{i,k}) = \frac{P(\pi_{i,k}) * \sum_{v_t \in V_T} (r(v_t) + d(v_t))}{E(\pi_{i,k})} \quad (4)$$

where, in the equations 3 and 4, the priority of values and their degree shared with the society are taken into consideration. Finally, the overall plan utility based on emotional reward is computed as:

$$u(\pi_{i,k}) = (EER_J + EER_P) - (EER_D + EER_{SR}) \quad (5)$$

[2]In this phase, for computational reasons, the plan is only partially expanded. Following the work in [8], plans can include 'abstract actions' that enable the agent to postpone part of the planning process at execution phase.

For example, Max has the goal to eat a chocolate candy. In order to satisfy his goal, the chocolate candy must be stolen from Mary but the 'steal' action makes the violation condition of the value 'honesty' true. So, if Max executes his plan, the emotional reward utility is derived from the Joy intensity and the Self-reproach intensity. Let us consider another plan, in which Max asks Mary to give him the chocolate candy. In this case no value is put at stake and the emotional reward utility is derived from the Joy intensity only. If the value 'honesty' is very important for Max, he chooses the plan to ask Mary the chocolate candy, even if the plan can have a less probability of success.

Goal/plan selection: Once the emotion utility is calculated for every goal $g_i \in G$ (and every plan $\pi_{i,k} \in \Pi$), the agent will choose the plan with the best reward utility. The goal related to this plan becomes a goal with an active state (i.e. the current agent intention) and the plan is chosen for execution by the agent.

Execution: following the work in [8], the π_{opt} plan is further expanded. The agent starts to execute an action of the optimal plan π_{opt}.

Monitoring: the agent perceives the world and updates her beliefs *(update(B))*.

Emotional Appraisal: The agent checks the UASet in order to assess what changes have occurred and their cause. According to the type of updates (goals achieved or failed and values re-established or at stake) the agent feels a certain emotion type. In Table 1 we recorded the rules with which the emotions are generated and how their intensity is computed. The 'Category' column specifies the emotions category, the 'Eliciting condition' column explains the rules used to generate the emotions, the 'Agent' column specifies who performed the action and, finally, the 'Intensity' column shows the formulas used to compute the emotion intensity. Our intensity model is based on the *expect utility model* [19] but, while in the expected utility model are taken into account probability and utility of goals attainment, we also include an effort cost. Concerning the intensity of moral emotions, we propose equations with more parameters than other computational model [29, 15] that include moral emotions in their work. Clearly, we have to assess the validation of our proposal with a test to figure out if our prediction are in line with human behaviour as suggested by [19].
Inspired by the work on 'literary feelings' by [27], in the following we refer to examples taken from the clay-animated film 'Mary and Max' by Adam Elliott (Australia, 2009). Mary and Max are friends and they are going for a walk in Central Park in New York city. In the examples reported below, if the UASet contains a pair <*updates, action*> in which the updates element can be one of the following:

- An addition of a success condition of a goal $g_i \in G_I$, the agent feels a **Joy** emotion *e(Joy, I_J, g_i, self)* because the goal g_i is reached (regardless of who performed the action). The intensity depends on the importance (intrinsic and extrinsic) of the goal achieved and on the effort that the agent made to achieve her goal.
  Example: Max is happy because he satisfied the goal of eating a vanilla ice cream.

- An addition of a failure condition of a goal $g_i \in G_I$, the agent feels a **Distress** emotion *e(Distress, I_D, g_i, self)* because the goal g_i is unachieved, regardless of who performed the action. The intensity depends on the importance of the goal and the effort, as Joy.
  Example: Max is sad because he has no money and his goal of eating a vanilla ice cream cannot be satisfied.

- A deletion of a violation condition of a value $v_i \in V_{atStake}$ and the action is an action performed by the agent, the agent feels a **Pride** emotion *e(Pride, I_P, v_i, self)* because the value is re-established. The intensity depends on the value rank, the importance of the moral goal, the effort made and the society value degree. The latter is considered as a measure of the admiration that the agent receives from others.
  Example: a robber assaults a little old lady in order to steal her necklace. Max, who is very brave, helps the old lady out and stops the robber by immobilizing him. Max is very proud of himself, because he re-established the value 'honesty' put at stake by the robber.

- An addition of a violation condition of a value $v_i \in V$ and the action is an action performed by the agent, the agent feels a **Self-Reproach** emotion *e(Self-reproach, I_{SR}, v_i, self)* because the value is put at stake. The intensity depends on the value rank, the effort made and the society value degree. The latter is considered as a measure of the reproach that the agent receives from others.
  Example: Max is ashamed because he put the value 'honesty' at stake by stealing a chocolate candy from Mary.

- A deletion of a violation condition of a value $v_i \in V_{atStake}$ and the action is an action performed by another agent, the agent feels an **Admiration** emotion *e(Admiration, I_A, v_i, other)* because the value is re-established. The emotion is directed towards the agent who performed the action. The intensity depends on the value rank and on the importance of the moral goal. Note that, if the agent has noticed that a value is at stake, then she must have a moral goal g_i motivated by the value. The intensity of admiration emotions depends on how the agent herself believes the value important.
  Example: Mary is proud of Max because he re-established the value at stake 'honesty', that both share, by stopping the robber.

- An addition of a violation condition of a value $v_i \in V$ and the action is an action performed by another agent, the agent feels a **Reproach** emotion *e(Reproach, I_R, v_i, other)* because the value is put at stake. The emotion is towards the agent who performed the action. The intensity of the Reproach emotion depends on the value rank and on the importance of the moral goal.
  Example: the old lady feels a reproach emotion directed toward the robber because he put at stake her value 'honesty'.

- An addition of a success condition of a goal $g_i \in G_I$, a deletion of a violation condition of a value $v_i \in V_{atStake}$ and the action is an action performed by the

Table 1: Emotion generation rules

Category	Eliciting condition	Agent	Intensity
Joy	a goal $g_i \in G_I$ achieved	self/other	$Imp(g_i) * E(\pi_{i,k})$
Distress	a goal $g_i \in G_I$ not achieved	self/other	$Imp(g_i) * E(\pi_{i,k})$
Pride	a value $v_i \in V_{atStake}$ balanced	self	$(r(v_i) + Imp(g_i) + d(v_i)) * E(\pi_{i,k})$
Sef-reproach	a value $v_i \in V$ at stake	self	$(r(v_i) + Imp(g_i) + d(v_i)) * E(\pi_{i,k})$
Admiration	a value $v_i \in V_{atStake}$ balanced	other	$r(v_i) + Imp(g_i)$
Reproach	a value $v_i(r, C) \in V$ at stake	other	$r(v_i) + Imp(g_i)$
Gratification	a goal $g_i \in G_I$ achieved, a value $v_i \in V_{atStake}$ balanced	self	I(Joy) + I(Pride)
Gratitude	a goal $g_i \in G_I$ achieved, a value $v_i \in V_{atStake}$ balanced	other	I(Joy) + I(Admiration)
Remorse	a goal $g_i \in G_I$ achieved, a value $v_i \in V$ at stake	self	I(Self-reproach) + I(Distress)
Anger	a goal $g_i \in G_I$ not achieved, a value $v_i \in V$ at stake	other	I(Reproach) + I(Distress)

agent, the agent feels a **Gratification** category emotion *e(Gratification, I_G, v_i, self)* because the individual goal is reached and the value re-established. Following the OCC model [26], Gratifications emotion is a compound emotions and the intensity is the sum of the intensity of Pride Emotion and Joy Emotion.
Example: the robber tries to steal the wallet to Max. Max is gratified about himself because he stopped the robber and saved his own wallet (the robber is not so lucky in anyone of these examples).

- An addition of a success condition of a goal $g_i \in G_I$, a deletion of a violation condition of a value $v_i \in V_{atStake}$ and the action is an action performed by another agent, the agent feels a **Gratitude** emotion *e(Gratitude, I_G, v_i, other)* because the individual goal is reached and the value re-established. The emotion is towards the agent who performed the action. The intensity of Gratitude emotion is the sum of the intesity of Admiration emotion and Joy emotion.
Example: the old lady is grateful to Max because he stopped the robbing and saved her necklace.

- An addition of a failure condition of a goal $g_i \in G_I$, an addition a violation condition of a value $v_i \in V$ and the action is an action performed by the agent, the agent feels a **Remorse** emotions *e(Remorse, I_R, v_i, other)* because the goal is unachieved and the value is put at stake. The intensity of Remorse emotion is the sum of the intesity of Self-reproach emotion and Distress emotion.
Example: Max feels remorse because he stole the chocolate candy from Mary and, as a consequence, Mary didn't want to eat dinner with him. The value 'honesty' is put at stake and the goal of eating with Mary is unsatisfied.

- An addition of a failure condition of a goal $g_i \in G_I$, an addition of a violation condition of a value $v_i \in V$ and the action is an action performed by another agent, the agent feels a **Anger** category emotion *e(Anger, I_A, v_i, other)*. The emotion is towards the agent who performed the action. The intensity of Anger emotion is the sum of the intesity of Reproach emotion and Distress emotion.
Example: Mary is angry with Max because he stole her chocolate and put at stake her value 'honesty' and he theatened the goal of eating dinner together.

6. DISCUSSION AND CONCLUSIONS

In this paper we presented the integration of emotional appraisal into a value–sensitive agent model. In our proposal, emotional appraisal plays the two fold role of driving the selection of goals and plans, and determining the generation of emotional states into the agent. Agents who feel (and can express) moral emotions as part of their deliberative processes and can appraise their own and the others' behavior in terms of moral emotions can be positively employed in applications that range from virtual companions to education and training.

The model we propose meets a set of requirements put forth by the research on emotions, such as the generation of moral emotions, necessary to create empathic agents, the generation of emotional states, which may be useful for the communication with other agents, the synchronisation with the values shared by a community via the drive provided by moral emotions in deliberation.

Some aspects of the model have not been explored into depth. For example, the model does not allows distinguishing goal failure from the simple inability to make the success condition of the goal true (for wrong execution, accidental failure, etc.). Intuitively, at the emotional level, this difference has to do with negative feelings such as frustration and, again, with the notion of responsibility (for example, in case an agent has undertaken the execution of an action based on a wrong assessment of her own capabilities). We don't model yet the behavior and cognitive consequent model that we reserve for future work.

Future work also includes the extension of the agent architecture to include different model of emotions. For example, we plan to include a richer representation of the emotions (in order to generate different emotional states within the same emotion category), a decay function and an overall mood state that can influence cognitive and behavioral aspects of the agent as in [15, 18]. Finally, we plan to implement our model and to test it in order to assert the validation of the model and of the equations with which the emotional reward utility of agent's plans and the intensity of emotion are computed in the Anticipatory appraisal phase and in the Emotional Appraisal phase. Some future decisions, as noted by [25], can be influenced by the domain application. Our aim is creating dramatic believable characters, situated into a virtual enviroment, being able to resolve moral conflict and to create a meaningful story in a storytelling system.

7. REFERENCES

[1] Michael Anderson and Susan Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15–26, 2007.

[2] Cristina Battaglino and Rossana 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.

[3] Christian Becker-Asano. *WASABI: Affect Simulation for Agents with Believable Interactivity*. PhD thesis, Faculty of Technology, University of Bielefeld, 2008. IOS Press (DISKI 319).

[4] Rafael Bordini and Jomi Hübner. Bdi agent programming in agentspeak using jason. In Francesca Toni and Paolo Torroni, editors, *Computational Logic in Multi-Agent Systems*, volume 3900 of *Lecture Notes in Computer Science*, pages 143–164. Springer Berlin / Heidelberg, 2006.

[5] Craig Boutilier, Thomas Dean, and Steve Hanks. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, 11:1–94, 1999.

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

[7] M.E. Bratman. What is intention. *Intentions in communication*, pages 15–32, 1990.

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

[9] J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, and L. van der Torre. The boid architecture: conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the fifth international conference on Autonomous agents*, pages 9–16. ACM, 2001.

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

[11] R. Conte and C. Castelfranchi. Norms as mental objects. from normative beliefs to normative goals. *From reaction to cognition*, pages 186–196, 1995.

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

[13] R. Damiano and V. Lombardo. An Architecture for Directing Value-Driven Artificial Characters. *Agents for Games and Simulations II: Trends in Techniques, Concepts and Design*, pages 76–90, 2011.

[14] F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1):69–79, 1999.

[15] 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.

[16] Clark D. Elliott. *The affective reasoner: a process model of emotions in a multi-agent system*. PhD thesis, Northwestern University, Evanston, IL, USA, 1992.

[17] Patrick Gebhard. Alma: a layered model of affect. In *Proc. of the fourth int. joint conf. on Autonomous agents and multiagent systems*, AAMAS '05, pages 29–36, New York, NY, USA, 2005. ACM.

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

[19] Jonathan Gratch, Stacy C. Marsella, Ning Wang, and Brooke Stankovic. Assessing the validity of appraisal-based models of emotion. In *Proc. of the Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, Amsterdam, The Netherlands, 2009.

[20] Jonathan Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, October 2001.

[21] J.F. Horty. Moral dilemmas and nonmonotonic logic. *Journal of philosophical logic*, 23(1):35–65, 1994.

[22] Nicholas Kushmerick, Steve Hanks, and Daniel Weld. An algorithm for probabilistic planning, 1995.

[23] Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, USA, August 1991.

[24] S. Marsella and J. Gratch. EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, March 2009.

[25] 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.

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

[27] D. Pizzi, F. Charles, J.L. Lugrin, and M. Cavazza. Interactive Storytelling with Literary Feelings. *ACII2007, Lisbon, Portugal, September*, 2007.

[28] A. Pokahr, L. Braubach, and W. Lamersdorf. Jadex: a BDI Reasoning Engine. *Multiagent Systems, Artificial Societies and Simulated Organizations*, 15:149, 2005.

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

[30] K. R. Scherer. The role of culture in emotion-antecedent appraisal. *Journal of Personality and Social Psychology*, 73:902–922, 1997.

[31] Susan Stark. Emotions and the ontology of moral value. *The Journal of Value Inquiry*, 38:355–374, 2004. 10.1007/s10790-005-1341-y.

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

[33] M.B. van Riemsdijk, M. Dastani, and M. Winikoff. Goals in Agent Systems: A Unifying Framework. In *Proceedings of AAMAS'08*, 2008.