Agents making moral decisions

Jaspreet Shaheed and **Jim Cunningham**
Department of Computing,
Imperial College London,
London SW7 2BZ, UK
{jss00,rjc}@doc.ic.ac.uk

Abstract. As the environments that intelligent agents operate in become more reflective of the real world, agent's decision-making processes must become more nuanced. In this paper, we present a decision-making model for an intentional agent which has been inspired by *Kohlberg's theory of moral development* and the *appraisal theory of emotion.* Agents utilising this model anticipate how undertaking actions will make both themselves and other agents feel, with the agent's sense of right and wrong helping to determine which emotions are evoked in which circumstances. We proceed to present some initial findings from runs of our agent implementation over situations from well known children's stories.

1 Introduction

As the environments that intelligent agents operate in become more reflective of the real world and our expectations of agents become greater, agents decision-making processes must become more nuanced. In the case of computer games, we are likely to feel much more empathy for characters in game worlds whose decision-making we intuitively understand.

In this paper we present a decision-making model for intentional agents[1] [11] which has been inspired by *Kohlberg's theory of moral development* [7], an ontology of moral reasoning which may help to explain the differing behaviours and priorities of game world characters. We describe the theory briefly in section 2. We follow this in section 3 with an introduction to *appraisal theory* [5] and the OCC model [8] which we use to 'ground' our use of Kohlberg's theory by taking morality to be feeling the right emotions in the right circumstances.

Section 4 represents the main contribution of this paper– a description of our agent which has been equipped with the decision model. A system which utilises our agents has been implemented in Qu-Prolog [1], a multi-threaded extension of Prolog which provides high-level communication between threads, processes and machines. Two scenarios representing the stories of "The three little pigs" [4] and "The pied piper of Hamelin" [6] have been created and we present some initial findings from these scenarios. Next in section 5, we take a brief look at some related work before finally presenting some conclusions and ideas for future work in section 6.

2 Kohlberg's theory of moral development

Kohlberg [7] interviewed people of different ages, telling them stories and posing them moral dilemmas based upon them. He found that whilst interviewees in the same age bracket might differ on the course of action they might suggest that characters of a story should take, the factors they took into account and the way they reached decisions were often similar. He classified responses, in so doing, identifying six distincts stages (or levels) of moral reasoning. The first two he termed *'pre-conventional'*, levels 3 and 4 *'conventional'* and levels 5 and 6 *'post-conventional'*[2]. He found that whilst respondents at higher levels would understand the reasoning of lower levels they would find them inadequate for responding to certain moral dilemmas and prefer the reasoning of the level they had reached.

The pre-conventional levels of moral reasoning (stages 1 and 2) are especially common in the youngest children although adults too can sometimes exhibit this kind of reasoning.

Stage 1: Individuals focus on the *direct consequences* that their actions will have for themselves. An action is perceived as right/wrong if the person who undertakes it is rewarded/punished and the better/worse the reward/punishment the better/worse the act must have been.

Stage 2: Right behaviour is defined by what is in one's own best interest. Concern for others is not based on loyalty or intrinsic respect but only to a point where it might further one's own interests. Less significance is attached to reward and punishment with punishment, for instance, being now regarded as an occupational hazard.

The conventional levels of moral reasoning are typical of adolescents and adults. Conventional reasoners judge the morality of their actions in comparison to societal views and expectations:

Stage 3: Individuals are receptive of approval or disapproval from other people and try to be a 'good boy' and 'good girl' and live up to other's expectation having learnt that there is an inherent value in doing so. Level 3 reasoners now take into account relationships (and their maintenance) when judging the morality of their actions.

[1] Intentional agents are guided by their plans and not just reacting to events.

[2] Kohlberg's theory has attracted some criticism for seeming to be biased towards certain types of societies, but we will not explore these issues here.

Stage 4: Individual now begin to take account of laws, dictums and other social conventions not for the approval of others (as in stage 3) but because of a belief in their importance in maintaining a functioning society (including, the belief that society's needs may often transcend one's own).

In the post-conventional levels of moral reasoning, an individual's sense of justice may lead them to hold critical views of laws or norms.

Stage 5: Individuals are regarded as having different values. Laws are no longer regarded as rigid dictums and where they do not promote the general welfare should be changed so as to meet the greatest good for the greatest number of people.

Stage 6: In Stage six, moral reasoning is based upon universal ethical principles. Laws are valid only insofar as they are grounded in justice and a commitment to justice carries with it an obligation to disobey unjust laws. Whilst Kohlberg insisted that stage 6 exists, he had difficulty finding participants who consistently demonstrated it.

In our research to date, we have focussed on stages 2 – 4. The reasoning of stage 1 seems particularly suited to the representation of child-like characters. Meanwhile, to represent the reasoning of stages 5 and 6, mechanisms different to the ones we will outline throughout the rest of this paper are likely to be needed: different deontic operators for social norms, laws and the beliefs an agent themselves holds about justice.

3 Appraisal theory and the OCC model

Appraisal theory [5] has recently become the 'predominant psychological theory of emotion' [12]. In appraisal theory, stimuli elicit emotions because of a person's subjective evaluation or *appraisal* of them. The questions of which criteria perceived stimuli are appraised against and which reactions are triggered have been explored by a number of researchers. One of the most applied models, the OCC model [8] was proposed by Andrew Ortony, Gerald Clore and Allan Collins and is shown in figure 1.

In the OCC model, emotions are seen as reactions to three types of stimuli: events, agents and objects. Central to appraising events is their desirability with respect to goals; central to appraising agents is the praiseworthiness of their actions with reference to standards; and central to appraising objects is their appealingness as determined by attitudes.

Of the OCC emotions, the ones that we are currently using (and the mechanisms by which they are evoked) are shown in table 1. An extension to the OCC set are the emotions 'being_admired' and 'being_reproached'– which represent the emotions of an agent which believes other agents hold these emotions towards it. The need for these emotions will be seen shortly.

In our research we use the OCC emotions to 'ground' our use of Kohlberg's theory by taking morality to be the feeling of the right emotions in the right circumstances. Table 2 shows how the OCC emotions may 'map' to Kohlberg's levels.

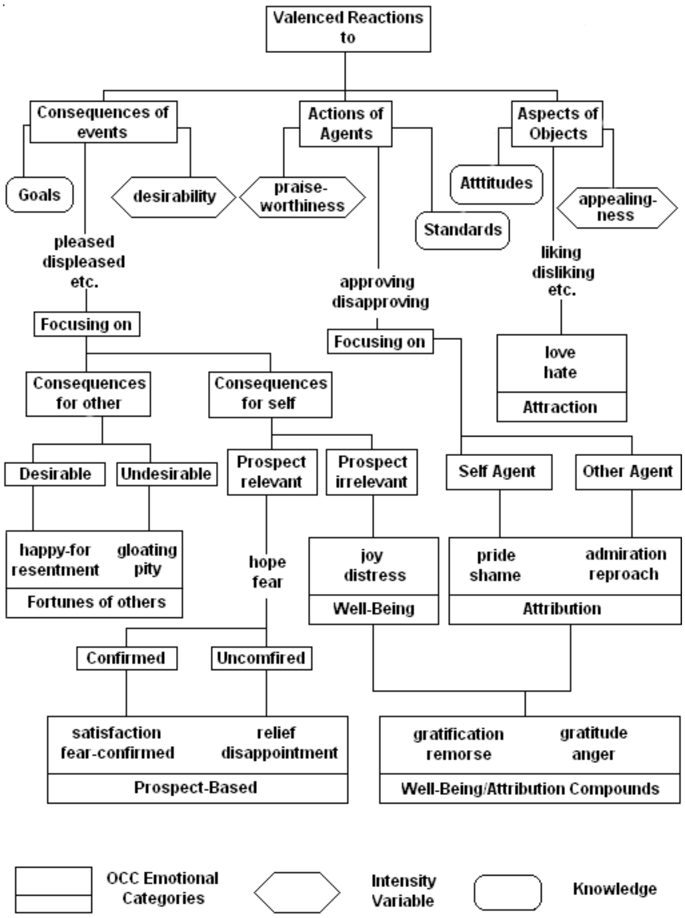

Figure 1. The OCC model of appraisal

Emotion	Cause
joy, distress	Evoked directly by particular states of affairs/happenings
pride, shame	Evoked by comparing ones own actions against standards (norms, laws, etc.)
admiration, reproach	Evoked by comparing other agent's actions against standards (norms, laws, etc.)
being_admired, being_reproached	Evoked when an agent believes other agents feel admiration/reproach respectively towards it
relief, disappointment	Evoked when the agent finds themselves in a better or worse than expected state respectively
anger	Evoked when feeling distress and reproach, or disappointment and reproach as a result of a particular agent's actions

Table 1. Implemented emotions

Stage	Emotions	
2	joy, distress	The agent is only concerned with the mental states of other agents where they lead them to undertake actions which will ultimately cause the agent joy or distress.
3	being_admired, being_reproached	The agent wants to avoid the reproach of other agents, instead wanting to earn their admiration.
4	pride, shame	The agents own judgement of its actions is now the most important determiner of the morality of an action.

Table 2. The emotions of different levels

4 Implementation of the decision model

Prolog pseudocode for the agent is shown in figure 2, it is similar to that of many BDI agents (such as AgentSpeak(L) [11]) but has been augmented with emotions.

```
agent_cycle :-
        get_percepts,
        update_beliefs,
        update_emotions,
        update_mode,
        ((
          reacting = true,
          execute_intention
        );(
          reacting = false,
          form_plans,
          select_plans,
          execute_intention
        )),
        agent_cycle.
```

Figure 2. The agent cycle

After perceiving the environment (which is represented using the event calculus [9]), the agent updates its beliefs. Changes in belief lead to appraisal and an update of emotions. The *update_emotions* predicate not only updates the agent's own emotions but the emotions it believes other agents are experiencing. The criteria for an agent's appraisal includes the intrinsic pleasantness/unpleasantness of states, interference with goals/expectations/intentions (as represented by the agent's plans) and the conformance (or non-conformance) of actions to standards.

The appraisal process is parameterised by the agent's morality (and the morality it assumes of other agents) so, whilst an agent might be aware of a standard (for instance, 'don't lie') and even expect other agent's to adhere to it, it will only feel the emotions evoked by comparing its actions to the standard if it is a level 4 agent.

Many ontologies of emotion distinguish *fullblown* emotional episodes in which a protagonist may be consumed by an emotion and *underlying* emotions for which the relationship between emotion and action is less clear. In order to account for fullblown emotional episodes, the *update_mode* predicate can cease further deliberation in favour of the adoption of a particular plan which will be executed without appraisal.

Seeking a preferable emotional state drives the *form_plans* and *select_plans* predicates. Plans are formed through abduction (using [14]) with possible goals the removal of sources of unhappiness (distress) or bringing about conditions which cause happiness (joy). Appraisal of plans is central to their selection and a number of processes are involved:

- The beliefs and emotions of every agent at every state within a plan are identified– however agent's preferences and expectations affect the emotions felt (for an agent to feel disappointed, it must have expected to be in a different, more preferable state) so until a plan is annotated with preferences and expectations an incomplete picture is produced.
- Prefered states *for every agent* need to be identified, using estimates of the morality of other agents. An emotional state A is preferred over an emotional state B if $A - B$ contains a good emotion of a level higher than any in $B - A$ or if $B - A$ contains a bad emotion of a level higher than any in $A - B$. The emotional preference algorithm is inspired by Kohlberg's observation that respondents prefer the reasoning of higher levels– so the judgement of the highest level of moral reasoning is the most important in decision-making.
- Expectations are determined by using knowledge of which agents participate in the plan together with their preferences to identify a path through the plan preferable for all involved. The agent cannot expect co-operation in plans which are not preferable for all, since, the emotional preferences it estimates for other agents already account for their desire to earn its approval and conform to certain values.
- A state may prompt a *reaction* from agents. In our implementation to date, the agent only modifies the states of the plan so as to account for the strong emotions evoked in agents by the plan (and the actions they may subsequently take) but not other (non-reactive) deliberation or counter-planning.
- Ultimately, if the agent has no expectations with respect to the plan (representing a path through the plan preferable to all agents whose cooperation is needed) then the plan is abandoned, otherwise it may be adopted (subject to resource constraints/absence of preferable plans).

4.1 The three little pigs

> Three little pigs leave home to seek their fortunes. Two of the pigs build themselves flimsy homes and are eaten by a wolf that blows their houses down whilst one builds a sturdy brick house and ultimately foils the wolf.

The aspect of the three little pigs story that we are most interested in is the decision-making of the pigs at the start of the story– when they decide what kind of house to build. A scenario representing the story has been created. It consists of a description of the agents in the story: their names, their morality levels and their beliefs as to the levels of morality of other agents in the scenario. It also contains event calculus axioms describing the initial situation and the effects of actions and finally rules which describe under which circumstances particular emotions are evoked. One such set of rules is:

- Having any kind of home causes joy (for the pigs).
- Building a brick house takes a lot of effort, causing distress (more so than having a home)
- Building a brick house causes pigs pride (if their level of morality >= 4).
- Building houses other than a brick house cause pigs shame (if level of morality >= 4).
- Pigs admire other pigs that build brick houses (if level of morality >= 3).
- Pigs feel reproach towards other pigs that build houses other than brick houses (if level of morality >= 3).

Given these rules, pigs set as having a low morality level (2) always choose to build straw or stick houses, whilst pigs at a high morality level (4) will always choose to build brick houses. Interestingly though, pigs at an intermediate level (3) will choose which type of house to build according to their beliefs about the other agents in the scenario– in other words, if there is no-one around who they believe will judge them (another agent with morality level >= 3) they will build a straw or stick house.

This highlights scope for an interesting extension. Currently, the agents' estimations of other agents are fixed– but if agents were to assume a default level for others and then refine that through observation, a pig might observe another pig building a solid house and then build a solid house themselves to avoid feeling bad as a result of the judgement of that pig.

In addition, the story could equally have been represented through other sets of rules– perhaps highlighting the feeling of safety that having a brick house would provide. This might correspond to a younger child's understanding of the story: the pigs, having left home, no longer fear the punishment of their mother. The behaviour of level 1 agents would no longer be constrained (hence building straw/stick houses). A level 2 agent with a more refined/common-sense approach to self interest, more able to look after itself, might choose to build a brick house, not because of high level emotions but simply out of emotions like *fear* and *hope*.

4.2 The pied piper of Hamelin

In the story of the pied piper of Hamelin, a village is overrun with rats. An enigmatic stranger (the piper) offers to rid the town of rats for which the villagers promise to pay. After he fulfils his end of the bargain the villagers renege on the agreement. To punish the villagers, the piper leads away the children of the town.

Figure 3 shows the piper's representation of his plan to get money. State $s1$ is the initial state. In state $s2$ the piper has offered to remove the rats from the village. In state $s3$, the villagers agree whilst in $s7$ they don't. In state $s4$ the piper has led the rats away. In $s5$ the villagers keep the agreement whilst in state $s6$ they break it (which leads to the piper taking the children away in state $s17$).

The piper estimates both his and the villager's emotions in each of these states. If the piper's own morality and his estimation of the villager's morality is set at level 4, he predicts that state $s6$ will evoke the emotions shown in figure 4. He will be disappointed (because he expected to be in state $s5$), reproachful because the villagers have broken their obligation and angry as a result of the presence of the other emotions. Meanwhile, the villagers will be ashamed of their own actions both intrinsically (shame) and because of the damage it will done to their relationship with the piper (being reproached). Additionally, they themselves will feel disappointed because they too would have preferred state $s5$ where they might have less money but would feel better about themselves!

Table 3 describes the results of running the scenario with differing values for the piper's morality and piper's estimation of the villagers' morality. The villagers' morality and their estimation of the piper's morality is fixed at (4,4).

Currently, no parameter settings lead to the recreation of the actual story. Some of the factors that inhibit the story emerging include:

- Agents appraise every state of a plan before selecting one rather than employing a short term lookahead. The villagers may have intended to keep the agreement when they made it but the system does not accommodate this possibility.
- Agents assume that their morality is (correctly) known to

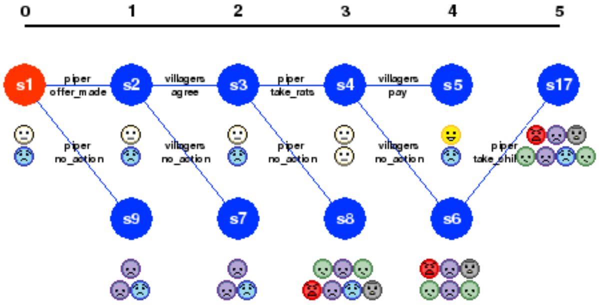

Figure 3. Pied piper's view

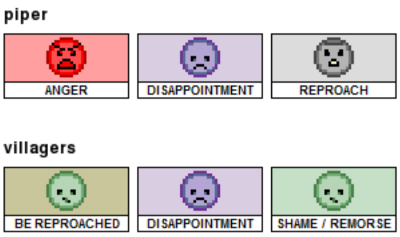

Figure 4. State s6

Morality	Results
(4,4), (4,3), (3,4), (3,3), (2,4)	Agreement made and kept, irregardless of the presence/absence of the 'take away children' action.
(4,2), (3,2)	Agreement made and kept only if the 'take away children' action is present in the scenario. Otherwise, the piper believes the villagers will be untroubled at breaking the agreement and he will have no recourse against them.
(2,3), (2,2)	Agreement not made or kept, irregardless of the 'take away children' action. The piper does not wish to make the agreement because he does not believe the villagers are concerned about his judgement. Additionally, since he wouldn't be surprised/reproachful/angry if the villagers did break the agreement- he doesn't even consider the possibility that he might end up taking the children of the village away and so realise that keeping the agreement would actually be in the villager's best interests.

Table 3. Pipers expectations at/assuming different levels of morality

other agents so even if the villagers did intend to break the agreement, they believe that the piper will see-through them.

5 Related work

We are not aware of any applications of Kohlberg's theory in AI, but there have recently been a number of applications of appraisal theory (because of an increasing interest in the use of emotions in computing [3], [2]). 'Double appraisal' in which an agent predicts how its actions will make another agent feel has been utilised in the FearNot! project, an educational program which models bullies and victims in a school setting [13]. Meanwhile, the creation of characters to populate 'story worlds' is the concern of [10], in which emotions may trigger behavioural switches.

Appraisal equipped agents tend to behave in a way more faithful to the psychological roots of appraisal theory whereas our approach combines appraisal and planning in a different way so that agents plan *in anticipation of* emotions as well as *because* of them.

6 Conclusions and future work

In this paper, we have presented a decision-making model for intentional agents which has been inspired by *Kohlberg's theory of moral development* and the *appraisal theory of emotion.* The agent's distinguishing feature is it's anticipation of how undertaking actions will make *itself and other agents feel*, a process parameterised by the *agent's sense of right and wrong*, and it utilising this knowledge as the basis for its decision-making. We have presented some initial findings from runs of the agent system over scenarios representing well known childrens' stories.

We believe our approach is promising, in the scenarios we have considered the differing representations that our model supports and extensions we envisiage seem to correspond to quite plausible understandings of the respective stories. However, at the moment our decision-model categorises only three types of agents, utilises relatively few of the OCC emotions and has only been run in small, tightly constrained environments. In future work, we plan to:

- Establish 'relationships' between the agents (by making admiration and reproach important to an agent only when the emotions come from particular agents).
- Utilise more of the OCC emotions, beginning with *'happy-for'*, *'pity'*, *'gratitude'* and *'remorse'.* These emotions are important to building relationships and may lead to agents (of level >= 3) adopting other agent's goals as their own.
- Create larger environments involving greater numbers of agents.

In addition, in order to evaluate our model, we plan to generate stories with morals by attempting to match runs of the system (with varying parameters) against a template for a moral story, in which a character of little moral virtue ends up unhappy as a (possibly indirect) consequence of their own actions. It is likely though, that some of the factors which inhibited the generation of the pied piper of Hamelin story may similarly inhibit the generation of other fable-like stories so these issues will need to be addressed.

REFERENCES

[1] Quprolog home page. http://www.itee.uq.edu.au/∼pjr/HomePages/QuPrologHome.html.

[2] e-circus, 2007. http://www.e-circus.org/.

[3] humaine, 2007. http://emotion-research.net/.

[4] Anon. The Three Little Pigs. http://www.pitt.edu/∼dash/type0124.html#halliwell.

[5] M. B. Arnold, *Emotion and personality*, Columbia University Press, New York, 1960.

[6] Robert Browning. The Pied Piper of Hamelin. http://www.pitt.edu/∼dash/hameln.html#browning.

[7] Lawrence Kohlberg, *The Development of Modes of Thinking and Choices in Years 10 to 16*, Ph.D. dissertation, University of Chicago, 1958.

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

[9] R.A. Kowalski and M.J. Sergot, 'A logic-based calculus of events', in *New Generation Comptuting*, (1986).

[10] Petta P. Rank, S., 'Appraisal for a character-based story-world', in *Intelligent Virtual Agents, 5th International Working Conference, IVA 2005*, (2005).

[11] Anand S. Rao, 'AgentSpeak(L): BDI agents speak out in a logical computable language', in *Seventh European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, pp. 42–55, Eindhoven, The Netherlands, (1996).

[12] Smith C. A. Roseman, I. J., *Appraisal Theory in K. Scherer, A. Schorr, T. Johnstone (Eds.) Appraisal Processes in Emotion: Theory, Methods, Research*, Oxford: Oxford University Press, 2001.

[13] Ruth Aylett Sandy Louchart and Joao Dias, 'Double appraisal for synthetic characters', in *Intelligent Virtual Agents*, (2007).

[14] M.P. Shanahan, 'An abductive event calculus planner', in *The Journal of Logic Programming*, volume 44, pp. 207–239, (2000).