A complete classification of ethical attitudes in Multiple Agent Systems

(Extended Abstract)

Matteo Cristani
Department of Computer Science
University of Verona
strada Le Grazie 15
I-37134 Verona, Italy
matteo.cristani@univr.it

Elisa Burato
Department of Computer Science
University of Verona
strada Le Grazie 15
I-37134 Verona, Italy
elisa.burato@univr.it

ABSTRACT

We provide a complete classification of agents' attitudes in performing actions that are evaluated in a three-valued space with respect to three evaluation contexts: personal advantage, social usefulness and legality. Moreover we found an algorithmical approach to approximatively solve the problem of abducing the agent's attitude by observing the actions she chose among her feasible ones. Knowing agents' behavioural rules is relevant in constituting agents' societies.

1. INTRODUCTION

Several recent investigation in Multiple Agent System and in Agent Cooperation have dealt with the problem of analysing agents' behaviour by means of their attitudes in performing actions. The majority of the scholars regarding to it aim at designing a MAS where individuals and groups are controlled by means of a general framework that realises a legal context. Nowadays it is an unrealistic approach since it presupposes the existence of an authority that centrally controls the activities of the agents; collaborative systems like Wikis, Content Management Systems, Forum etc. are open and not controlled. In such contexts it results interesting to apply those major concepts that have been studied in the perspective of a moral artificial agent as in [1]. Starting from a different viewpoint we assume that agents can enter the system without strict control mechanisms in front, so a formal punishment systems can be difficult to implement. If an agent that enters a system will be in charge of a specific duty we will assign her this duty only if we consider her reliable. Although some efforts in this direction have already been performed [2, 3], a general framework for treating attitudes as a basic component in MAS has not yet been carried out.

2. ETHICAL MULTIPLE AGENT SYSTEMS

In this paper we focus upon the problem of *attitude abduction*:

> *Given a set of actions* $X_{ag} = \{x_1, x_2, \ldots, x_n\}$ *and an agent* ag*, establish the minimal set of admissible attitudes that are compatible with the performance of the actions in* X_{ag}*.*

Cite as: A Complete Classification of Ethical Attitudes in Multiple Agent Systems, (Extended Abstract), Matteo Cristani, Elisa Burato, *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009),* Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10–15, 2009, Budapest, Hungary, pp. 1217–1218

The formalisation of the attitude abduction problem we propose here has been partially inspired by the work described in [3] and by the paper [2]. The starting point is to extend Multiple Agent System with an *ethical component* to enable agents in evaluating their feasible actions. We define an *Ethical MAS*(EMAS) as a tuple $\mathcal{M} = \langle A, X, C, \aleph(.), e(.,.)\rangle$, where A is the set of the agents, X is the set of the actions, C is the set of the evaluating contexts, $\aleph : A \rightarrow X$ is the function that maps an agent to her feasible actions and $e : A \times X \rightarrow S^{|C|}$ is the evaluation functions which represents how an agent ag evaluates an action x with respect to all the contexts in C and by assuming S as the evaluation space. In our framework we suppose that C contains three contexts L, S and I representing respectively the legality, the social utility and the personal advantage; moreover we assume that the actions are evaluated in a three-valued space $S = \{P, N, -\}$ stating that an action could respectively *promote*, *demote* or *be indifferent* with respect to an EMAS context. The meanings of the action evaluation in each context are:

Legality : an action may be a duty (P), an opportunity ($-$) or a felony (N);

Social utility : an action may be an obligation (P), a possibility ($-$) or an iniquity (N);

Personal Advantage : an action may be advantageous (P), fair ($-$) or disadvantageous (N).

We assume that the values in S are totally ordered as $P > - > N$.

3. AGENT ATTITUDES: DEFINITION AND CLASSIFICATION

The general approach in MAS literature defines the agent attitude as a *context ordering* that implements the agent preference relations in choosing actions to perform. Given an EMAS $\mathcal{M} = \langle A, X, C, \aleph(.), e(.,.)\rangle$ and an agent $ag \in A$, the *attitude* of ag is a total or partial order among the contexts in C and we denote it by $\Upsilon(ag) = \{(c_i, c_j) | c_i, c_j \in C$ and ag prefers c_i to $c_j\}$. For each attitude we build a graph of evaluations of actions that we call the *actions hierarchy*, indicated by $\mathcal{G}(\Upsilon(ag)) = (V, E)$ such that:

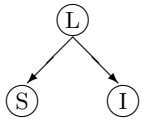
- $V = \{v | \exists x \in \aleph(ag) : \forall (c_i, c_j) \in \Upsilon(ag) : v = e(ag, x) \wedge v_i \succ v_j\}$
- $E = \{(v, w) \in V^2 | v \succeq w\}$

By assuming that the evaluations contexts are the obedience to the law, the social utility and the interest in personal advantage as said in the precedent section, and by having a three-valued evaluation space, there are several different attitudes an agent may adopt. Below we make a classification of the attitudes an agent may present and we explain them in a human readable form; by definition, each attitude corresponds to a contexts ordering. We exemplify human readable description for attitudes 1, 7, 10 and 17.

1. **legalistic/selfish** $\Upsilon = \{(L, I), (I, S)\}$:

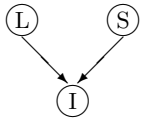
"In all the situations, choose the action that is a legal duty and not individually disadvantageous";

2. **legalistic/altruistic** $\Upsilon = \{(L, S), (S, I)\}$;
3. **altruistic/legalistic** $\Upsilon = \{(S, L), (L, I)\}$;
4. **altruistic/selfish** $\Upsilon = \{(S, I), (I, L)\}$;
5. **selfish/legalistic** $\Upsilon = \{(I, L), (L, S)\}$;
6. **selfish/altruistic** $\Upsilon = \{(I, S), (S, L)\}$;
7. **absolutely legalistic** $\Upsilon = \{(L, I), (L, S)\}$:



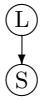
"In all the situations, choose the action that is a legal duty";

8. **absolutely altruistic** $\Upsilon = \{(S, I), (S, L)\}$;
9. **absolutely selfish** $\Upsilon = \{(I, L), (I, S)\}$;
10. **legalistic and altruistic** $\Upsilon = \{(L, I), (S, I)\}$:



"In all the situations choose the action that is a legal duty or a socially useful action even if it is not an individually advantageous one";

11. **legalistic and selfish** $\Upsilon = \{(L, S), (I, S)\}$;
12. **selfish and altruistic** $\Upsilon = \{(I, L), (S, L)\}$;
13. **legal over individual and socially detached** $\Upsilon = \{(L, I)\}$;
14. **individual over legal and socially detached** $\Upsilon = \{(I, L)\}$;
15. **social over individual and legally detached** $\Upsilon = \{(S, I)\}$;
16. **social over individual and legally detached** $\Upsilon = \{(I, S)\}$.
17. **legal over social and individually detached** $\Upsilon = \{(L, S)\}$:



"In all the situations, do not worry about your personal advantage and choose the action that is a legal duty; if you have no legal obligations to perform choose actions that are socially useful";

18. **social over legal and individually detached** $\Upsilon = \{(S, L)\}$;

The above enumerated attitudes can be organised in an attitude taxonomy that we do not provide here for the sake of space; however the is-a relation among the attitudes is easily readable: for instance a legalistic/selfish agent is also an absolute legalistic one.
Having the taxonomy of all agent attitudes we regard to the *coherence* of an actions to the agent attitude. An action x is *coherent* with the attitude $\Upsilon(ag)$ of the agent ag if it is in the action hierarchy of ag, $\mathcal{G}(\Upsilon(ag))$, and no other feasible action $y \in \aleph(ag)$ is preferred to x in $\mathcal{G}(\Upsilon(ag))$.
The problem of attitude abduction results to be solvable if the performed and the feasible actions are known. We found an algorithm that solves the attitude abduction problem and that is sound and complete. We do not provide the algorithm here for the sake of space but we give the main steps it follows:

1. Receive the admissible attitudes A of the agent ag;
2. For each action x the agent performed test if x is coherent with each of the admissible attitudes in A;
3. if there is an attitude $a \in A$ to which x is not coherent then $A = A \setminus \{a\}$;
4. if A is empty then output that ag is incoherently committed to her performed actions, otherwise output that ag has the admissible attitudes in A.

The algorithm is polynomial in the number of attitudes, in the size of performed action sequence and in the number of feasible actions.
Further developments will be explored that are based upon the introduction of *privileges* (namely actions reserved to certain agents) and *immunities* namely duties that some agents is not forced to do. This aspect needs to be related to the fundamental concepts in legal theories, as started to be discussed in the linguistic literature.

4. REFERENCES

[1] C. Allen, G. Varner, and J. Zinser. Prolegomena to any future artificial moral agent. *J. Exp. Theor. Artif. Intell.*, 12(3):251–261, 2000.
[2] K. Atkinson and T. Bench-Capon. Addressing moral problems through practical reasoning. *Journal of Applied Logic*, 6:135–151, 2008.
[3] M. Cristani and E. Burato. Approximate solutions of moral dilemmas in multiple agent system. *International Journal of Knowledge and Information Systems*, 18(2):157–181, 2009.