Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 47–53
ISBN 978-83-60810-14-9
ISSN 1896-7094

Utility-based Model for Classifying Adversarial Behaviour in Multi-Agent Systems

Viliam Lisý, Michal Jakob, Jan Tožička, Michal Pěchouček
Gerstner Laboratory – Agent Technology Center
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University
Technická 2, 16627 Praha 6, Czech Republic
{viliam.lisy,michal.jakob,jan.tozicka,michal.pechoucek}@agents.felk.cvut.cz

***Abstract*—Interactions and social relationships among agents are an important aspect of multi-agent systems. In this paper, we explore how such relationships and their relation to agent's objectives influence agent's decision-making. Building on the framework of stochastic games, we propose a classification scheme, based on a formally defined concept of *interaction stance*, for categorizing agent's behaviour as self-interested, altruistic, competitive, cooperative, or adversarial with respect to other agents in the system. We show how the scheme can be employed in defining behavioural norms, capturing social aspects of agent's behaviour and/or in representing social configurations of multi-agent systems.**

I. Introduction

Recently, there has been a growing interest in studying complex systems, in which large numbers of agents pursue their goals while engaging in mutual interactions. Examples of such systems include real-world systems, such as diverse information and communication networks, as well as simulations of real-world systems, such as models of societies, economies and/or warfare. With the increasing diversity of these systems, there is a growing need to develop models which allow characterizing the behaviour of agents, and their interactions, in a compact form.

Fundamentally, the behaviour of an agent is driven by its objectives. In the absence of other constraints and influences, the agent is expected to perform actions which – incrementally but not necessarily monotonously – lead towards its objectives. In some situations, however, the way the agent acts can be affected by additional factors, be it the surrounding environment, agent's decision making capability or *its relations to other agents*. Knowing and understanding such factors may help in reasoning about the agent and in obtaining better prediction of its behaviour, compared to when only agent's overall objectives are considered.

Whereas the impact of the first two factors have been studied in several fields related to intelligent agents, including game theory and planning, comparatively less work seems to exist on relating agent's social relations, agents' objectives and the behaviour they ultimately execute. The aim of this paper is therefore to investigate and formalize *inter-agent relations* as an important behaviour-modifying factor in communities of social agents. Specifically, building on the framework of stochastic games and extending our earlier work [1], we propose a utility-based model which categorizes actions (and consequently also relationships among agents) as self-interested, altruistic, competitive, cooperative and adversarial. The concept of interaction stance allows to define, classify, and/or regulate agent behaviours not only with respect to agent's own objectives but also with respect to objectives of other agents in the system.

The rest of the paper is organized as follows. In Section II, we identify factors which affects how an autonomous agent pursues its objectives. Section III exposes the major part of the contribution – the classification model of agent's behaviour. Section IV shows several examples illustrating the application of the model. Section V overviews the related work and Section VI concludes with a summary.

II. Agent's Decision Making

As already mentioned, the behaviour of an agent is primarily driven by its overall objectives (also referred to as *desires* in the BDI architecture [2]). Out of the factors which constraints/affect how the agent actually behaves, we can distinguish

- *social relations* to other agents – if an agent is cooperative with another agent, it may consider the other agent's objective in choosing the action it performs. If it has multiple alternatives how to fulfil its objectives, it can choose the one that would help the other agent or even follow a suboptimal course of action in order to help the other agent reaching its goals.
- *environment* can limit the available actions the agent is able to perform; additionally, the environment can make action outcomes non-deterministic due to its stochastic nature or interfering actions of other agents
- *decision-making capability* reflects the extent with which the agent is able to pursue its objectives. Some agents are purely reactive, others can use planning or predict changes of environment. The agent can be trying to reach a goal, but due to insufficient computational resources or some architectural limits end up choosing a suboptimal or even contra productive action. Decision making capability is closely related to the issue of *bounded rationality* [3].

Explicit consideration of different factors affecting agent's behaviour provides for a more detailed characterization of agent's decision making, in turn allowing to reason not only about agent's current behaviour but also about its behaviour in situations where some of these factors change.

For example, knowing than an agent A has a cooperative stance towards agent B allows to infer that, in the presence of agent B, the agent A will, if at all possible, perform actions that contribute towards achieving agent B's objectives if such actions are not necessarily optimum when agent A's objectives are considered alone (and which agent A would pursue if agent B was not present). Similarly, knowing that the environment prevents an agent from executing an action that contributes towards its objectives allows to infer that, should the environment state change favourably, the agent would execute the action (even if it has not performed it so far at all).

The rest of the paper is dedicated to formalizing the above given notions. Primarily, we focus on the effect of social relations, though the role of the environment is also considered to an extent.

III. Interaction-based Classification Model

This section presents an interaction-based model developed for multi-agent systems with asymmetric agent's utility functions. The model allows classifying actions and their sequences with respect to their effects on utilities of agents in the system. It is based on the formalism of partially observable stochastic games [4] generalized to infinite state, action, and observation spaces and omitting the initial state and reward functions, which are substituted by agent utility functions.

The choice of utility functions as a linear combination of some predefined characteristics of the world as we define it later, over possibly more expressive reward functions is motivated primarily by the compactness of representation allowed by the former. The description of agents using utilities also seems closer to how people tend to think about agents – it is more straightforward to specify what an agent is trying to achieve in terms of the desired state of the world than trying to assign a reward for each possible state.

Although the underlying model of stochastic games is general enough to describe incompleteness of agent's knowledge about the world, we do not consider this factor in the current version of the model. Likewise, the explicit consideration of agent's possibly limited decision-making capability is currently not considered.

A. Fundamental Definitions

Definition 1. The game model is a tuple $(\mathcal{I}, \mathcal{W}, \{\mathcal{A}_i\}, \{\mathcal{O}_i\}, \mathcal{P})$, where

- $\mathcal{I}$ is a finite set of agents (players) indexed by $1, \ldots, n$
- $\mathcal{W}$ is a possibly infinite set of all states of the world
- $\mathcal{A}_i$ is a set of actions available to agent $i \in \mathcal{I}$ and $\mathcal{A} = \mathsf{X}_{i \in I}\, \mathcal{A}_i$ is a set of joint actions where $a = \langle a_1, \ldots, a_n \rangle$ denotes a joint action
- $\mathcal{O}_i$ is a set of observations for agent i and $\mathcal{O} = \mathsf{X}_{i \in I}\, \mathcal{O}_i$ is the set of joint observations where $o = \langle o_1, \ldots, o_n \rangle$ denotes a joint observation
- $\mathcal{P}$ is a set of Markovian state transition and observation probabilities, where $\mathcal{P}(w', o \mid w, a)$[1] denotes the probability that taking a joint action a in a state w results in a transition to state w' and a joint observation o.

Another part of the formal model is the space of all possible world characteristics that could concern some agent. We refer to this space as the utility space and we assume that it is a subset of $\mathbb{R}^m$, where m is the number of characteristics considered. A point in the utility space is a vector of world-specific values of these characteristics. Let us consider an example game in which an agent considers building a highway. Different characteristics the agent can consider in such a scenario include the impact on the traffic situation, financial cost or harm to the surrounding nature. Different agents have a different view on the importance of individual characteristics.

Definition 2. The *utility space* $\mathcal{U} \subseteq \mathbb{R}^m$ of a system is a vector space generated by all components of the utility functions in the system. We assume there exists a global utility function $\vec{u} : \mathcal{W} \to \mathcal{U}$ that assigns a utility vector to each state of the world.

Each agent in the world values the components of the utility space differently. A local scout organization cares more about trees and the local labour union values more employed workers. The government should consider both characteristics. The preferences of an agent over the utility space components (i.e. the characteristics) are expressed by a vector of real weights. The overall utility that an agent ascribes to a state of the world is then a preference-weighted sum of the utility components.

Definition 3. If $\vec{u}(w) \in \mathcal{U}$ is a point in the utility space corresponding to a state $w \in \mathcal{W}$ of a system and $\vec{u_A} \in \mathbb{R}^m, |\vec{u_A}| = 1$ is the preference vector of agent $A \in \mathcal{I}$, then the utility of the state w for the agent A is

$$u_A(\vec{u}(w)) = \sum_{i=1}^{m} u_A^i u^i = \vec{u_A} \cdot \vec{u}(w)$$

where the dot operation represents the dot product in $\mathbb{R}^m$. For a group of agents $\mathbf{G} \subseteq \mathcal{I}$ we define the preference vector as

$$\vec{u_{\mathbf{G}}} = \sum_{i \in \mathbf{G}} \vec{u_i}$$

Using this preference vector, the utility of the state of the world w for the group is defined as

$$u_{\mathbf{G}}(\vec{u}(w)) = \vec{u_{\mathbf{G}}} \cdot \vec{u}(w)$$

[1] Since we do not focus on agents' observations, we will omit them in the following text, i.e. we will use this notation: $\mathcal{P}(w' \mid w, a)$.

B. Normalization

Note that we require the preference vectors to be normalized for individual agents. The main reason for that is that we want to be able to compare and sum utilities of different agents. If we have two agents and both of them want to maximize only the amount of their money, we do not consider reasonable to say that one of them wants to maximize it more than the other. The difference comes only with introducing another utility component, e.g. harming innocent people. After that, the agents can differ in how much they want to maximize their money considering how their actions harm innocent people.

However, the preference vector for a group of agents is not normalized anymore. It expresses not only the preference relations between the different components, but also how big and consistent in the preferences the group is. A big group with agents that have random preference vectors has the group preference vector close to zero whereas a group of agents with identical preferences has a large preference vector in the direction of the preferences of individual agents.

The above definition of group preference vector also ensures the property of *associativity of subgroups*. If we have a set of agents grouped into several groups and we join the groups to create a bigger group including all the agents, the preference vector of the resulting group only depends on the individuals in the group and not on the subgroups they were previously part of. This property would not hold if we normalized group preference vector and combine the resulting normalized vectors.

C. Action Utility

Although utility is usually defined for a state of the world, we also define it for an action in a state of world. The utility of an action is the difference between the utility of the state after the action is performed and the utility of the state before. However, such a definition does not take into account the potential non-determinism of action effects. Instead we define the expected utility of an action as the average utility of an action if it was performed an infinite number of times in the same state of the world.

Definition 4. If $w_0 \in \mathcal{W}$ and $a \in \mathcal{A}$ are a world state and a joint action, then we define the expected utility of the action a in the world state w_0 as

$$\vec{eu}(a, w_0) = \left(\int_{\mathcal{W}} \mathcal{P}(w \mid w_0, a) \vec{u}(w) dw \right) - \vec{u}(w_0)$$

Note that $\vec{eu}(a, w_0) \in \mathcal{U}$.

D. Taxonomy of Actions

There is nothing but joint actions in the real world. All agents are concurrently choosing from amongst a huge number of their individual actions and the world changes accordingly. Some actions can be easily attributed to a single agent, e.g. pressing a button, however many other actions may have multiple actors involved to a different degree, e.g. a car accident. The outcome of a joint action can also have multiple independent parts relevant to different agents.

Our main goal in this section is to classify joint actions as cooperative, self-interested, competitive and adversarial, with respect to different agents or groups thereof. Collectively, we refer to these classes as interaction stance. In order to do this, we need to separate the part of the action effect which is relevant for the group and to think about what could the group have done differently to change its influence to the part.

First of all, we need to get rid of the irrelevant components of the outcome. Consider two groups of agents with preference vectors $\vec{u_{\mathbf{G}}}$ and $\vec{u_{\mathbf{H}}}$. These two vectors generate a vector subspace of the utility space. If the directions of the preference vectors are the same or opposite, the subspace degenerates to a single dimension; otherwise it is a two-dimensional plane.

The model works also for one dimension, but let us assume a more general vector orientation. For any action, the important part for classification from the viewpoint of the groups $\mathbf{G}$ and $\mathbf{H}$ is *the projection of the action's expected utility vector to the subspace*. If we assume for a while, that a group of agents $\mathbf{G}$ is fully responsible for a joint action a considering the groups $\mathbf{G}$ and $\mathbf{H}$, we can draw the action classification scheme as in Figure 1. Below we examine the different interaction stances

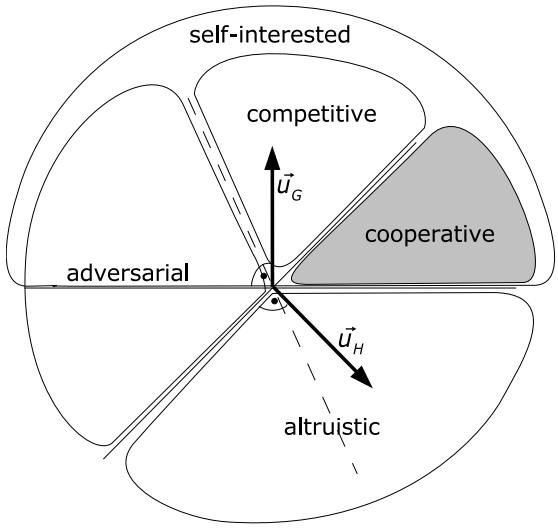

Fig. 1. Classification of a joint action.

in more detail. An action is considered self-interested from the group $\mathbf{G}$ if it increases the utility of the group.

Definition 5. We say that a joint action $a \in \mathcal{A}$ is *self-interested* for a group of agents $\mathbf{G} \subseteq \mathcal{I}$ in a state of the world $w \in \mathcal{W}$

$$\mathtt{si}_{\mathbf{G}}(a, w) \Leftrightarrow \vec{u_{\mathbf{G}}} \cdot \vec{eu}(a, w) > 0$$

We say that an action is cooperative with $\mathbf{H}$, if it changes the world in a way that increases the utility of both groups. Cooperative actions are symmetric. Actions that are cooperative from $\mathbf{G}$ towards $\mathbf{H}$ have exactly the same expected utility vectors as actions that are cooperative from $\mathbf{H}$ towards $\mathbf{G}$.

Definition 6. We say that a joint action $a \in \mathcal{A}$ is *cooperative* from a group of agents $\mathbf{G} \subseteq \mathcal{I}$ towards a group $\mathbf{H} \subseteq \mathcal{I}$ in a state of world $w \in \mathcal{W}$

$$\texttt{coop}_{\mathbf{G}\to\mathbf{H}}(a,w) \Leftrightarrow \vec{u_{\mathbf{G}}} \cdot \vec{eu}(a,w) > 0 \;\&\; \vec{u_{\mathbf{H}}} \cdot \vec{eu}(a,w) > 0$$

An action is competitive, if it does not decrease the utility of the group $\mathbf{H}$ more than it increases the utility of the group $\mathbf{G}$.

Definition 7. We say that a joint action $a \in \mathcal{A}$ is *competitive* from a group of agents $\mathbf{G} \subseteq \mathcal{I}$ towards a group $\mathbf{H} \subseteq \mathcal{I}$ in a state of the world $w \in \mathcal{W}$

$$\texttt{comp}_{\mathbf{G}\to\mathbf{H}}(a,w) \Leftrightarrow$$
$$\vec{u_{\mathbf{G}}} \cdot \vec{eu}(a,w) \geq |\vec{u_{\mathbf{H}}} \cdot \vec{eu}(a,w)| \;\&\; \vec{u_{\mathbf{H}}} \cdot \vec{eu}(a,w) < 0$$

An action is adversarial if it lowers the utility of the group $\mathbf{H}$ more then it increases the utility of the group $\mathbf{G}$ or even decreases the utilities of both the groups.

Definition 8. We say that a joint action $a \in \mathcal{A}$ is *adversarial* from a group of agents $\mathbf{G} \subseteq \mathcal{I}$ towards a group $\mathbf{H} \subseteq \mathcal{I}$ in a state of the world $w \in \mathcal{W}$

$$\texttt{adv}_{\mathbf{G}\to\mathbf{H}}(a,w) \Leftrightarrow$$
$$\vec{u_{\mathbf{G}}} \cdot \vec{eu}(a,w) < -\vec{u_{\mathbf{H}}} \cdot \vec{eu}(a,w) \;\&\; \vec{u_{\mathbf{H}}} \cdot \vec{eu}(a,w) < 0$$

The last interaction stance not yet defined is the altruistic action. We define it as an action which helps some other agent while lowering the utility of the performing agent.

Definition 9. We say that a joint action $a \in \mathcal{A}$ is *altruistic* from a group of agents $\mathbf{G} \subseteq \mathcal{I}$ towards a group $\mathbf{H} \subseteq \mathcal{I}$ in a state of the world $w \in \mathcal{W}$

$$\texttt{alt}_{\mathbf{G}\to\mathbf{H}}(a,w) \Leftrightarrow$$
$$\vec{u_{\mathbf{G}}} \cdot \vec{eu}(a,w) < 0 \;\&\; \vec{u_{\mathbf{H}}} \cdot \vec{eu}(a,w) > 0$$

At this point, we can define several classic game theoretic concepts in our framework. For example the utilities of a two player zero-sum game can be described by two vectors for which $\vec{u_A} = -\vec{u_B}$. Generally, we can define relationships between agents and group of agents based on the similarity of their weights of different utility components.

Definition 10. We say, that groups of agents $\mathbf{G}, \mathbf{H} \subseteq \mathcal{P}$ have *cooperative potential* if

$$\vec{u_{\mathbf{G}}}.\vec{u_{\mathbf{H}}} > 0$$

They have *adversarial potential*

$$\vec{u_{\mathbf{G}}}.\vec{u_{\mathbf{H}}} < 0$$

The definitions correspond to the correlation of agent payoffs used in game theory [5]. Two groups have cooperative potential if most of the self-interested actions of each group are cooperative with respect to the two groups. The adversarial potential occurs if most of self-interested actions are competitive or adversarial.

E. Intentionality of Adversarial Action

The above defined concept of the adversarial action considers the effects of an action with respect to agents' individual and collective utilities. Classifying an agent as an adversary is based on the purely external analysis of agents behaviour, without taking into account agent decision-making capabilities and the influence of the environment on the executability and outcomes of its actions. The model presented so far can only represent whether the actions of a particular agent or a group of agents are helping or harming someone else.

In the real world, however, cooperative agents are often forced to choose the smallest evil, e.g. to choose an action that will harm the others least from all the available actions. In such situations, classifying the least harmful action as adversarial, as done by the model, might not be appropriate, as it may be the case that the agent has no other choice than to harm the other agents. A similar problem arises with the classification of the least beneficial action as cooperative if an agent can only perform beneficial actions.

This problem has been in part addressed in [6], where the concept of intentional adversarial action has been introduced. This definition assumes intentional adversariality based on agents' knowledge of adversarial nature of the particular action and agents knowledge of the existence of an alternative action that can be performed with less harmful effect. The definition could be loosely rewritten in the presented formalism as follows. If $\mathcal{A}^w_{\mathbf{G}}$ is the set of actions available to the group of players $\mathbf{G}$ in the state of world $w \in \mathcal{W}$ and $a_0 \in \mathcal{A}^w_{\mathcal{I}\setminus\mathbf{G}}$ is a combination of actions of the players that are not in $\mathbf{G}$, then an action $a = (a_0, a_{\mathbf{G}})$ is intentionally adversarial if

$$\texttt{adv}_{\mathbf{G}\to\mathbf{H}}(a,w) \wedge \exists a' = (a_0, a'_{\mathbf{G}}) \in \mathcal{A} \text{ such that}$$
$$\vec{u_{\mathbf{H}}} \cdot \vec{eu}(a,w) + \vec{u_{\mathbf{G}}} \cdot \vec{eu}(a,w) <$$
$$\vec{u_{\mathbf{H}}} \cdot \vec{eu}(a',w) + \vec{u_{\mathbf{G}}} \cdot \vec{eu}(a',w)$$

This definition is quite strict. If a player with almost all its actions adversarial towards someone does not perform the single least harmful action, it is considered adversarial.

An alternative approach to the definition of an intentionally adversarial action and to effective separation of agent's intention and the interfering effect of the environment is based on the concept of *neutral behaviour*. Which action should an agent choose so that nobody could legitimately accuse it of acting self-interestedly or adversarially? The best solution in our opinion is to consider neutral an agent which chooses its action randomly with uniform distribution. With respect to this, the neutral outcome of an action is not the zero vector, but the average outcome of all the actions performable in a certain state of world. This corresponds to the centre of mass of all performable actions.

Definition 11. If $\mathcal{A}^w_{\mathbf{G}}$ is the set of actions available to the group of players $\mathbf{G}$ in the state of world $w \in \mathcal{W}$ and $a_0 \in \mathcal{A}^w_{\mathcal{I}\setminus\mathbf{G}}$ is a combination of actions of the players that are not in $\mathbf{G}$ then

the *neutral utility* is

$$\vec{c} = \frac{1}{|\mathcal{A}^w_{\mathbf{G}}|} \sum_{a_{\mathbf{G}} \in \mathcal{A}^w_{\mathbf{G}}} e\vec{u}((a_0, a_{\mathbf{G}}), w)$$

If we want to classify an action, we first have to subtract the neutral utility from its expected utility and classify the difference as described above. Based on the neutral utility, an intentionally adversarial action would be defined as follows:

Definition 12. We say that a joint action $a = (a_0, a_{\mathbf{G}}) \in \mathcal{A}$ is *intentionally adversarial* from a group of agents $\mathbf{G} \subseteq \mathcal{I}$ towards a group $\mathbf{H} \subseteq \mathcal{I}$ in a state of the world $w \in \mathcal{W}$

$$\texttt{adv\_int}_{\mathbf{G}\to\mathbf{H}}(a, w) \Leftrightarrow$$
$$u\vec{_{\mathbf{G}}} \cdot (e\vec{u}(a, w) - \vec{c}) < -u\vec{_{\mathbf{H}}} \cdot (e\vec{u}(a, w) - \vec{c})$$
$$\wedge\, u\vec{_{\mathbf{H}}} \cdot (e\vec{u}(a, w) - \vec{c}) < 0$$

If the agent has no alternative to the action that is a subject of our investigation, its expected utility would be the same as the expected utility of the neutral action, and thus $(e\vec{u}(a, w) - \vec{c}) = 0$. This fact does not allow classifying the action as intentionally adversarial. The least harmful action from all possible actions also cannot be classified as intentionally adversarial. The last statement is generalized in the following proposition.

Proposition 1. *If there is an action $a = (a_0, a_{\mathbf{G}}); a_{\mathbf{G}} \in \mathcal{A}^w_{\mathbf{G}}$ intentionally adversarial from group $\mathbf{G}$ towards a group of agents $\mathbf{H}$ then there exists a sub-action $a'_{\mathbf{G}} \in \mathcal{A}^w_{\mathbf{G}}$ that forms an action that is less harmful to $\mathbf{H}$.*

$$\texttt{adv\_int}_{\mathbf{G}\to\mathbf{H}}(a, w) \Rightarrow\ \exists a'_{\mathbf{G}} \in \mathcal{A}^w_{\mathbf{G}}$$
$$u\vec{_{\mathbf{H}}} \cdot (e\vec{u}((a_0, a'_{\mathbf{G}}), w)) > u\vec{_{\mathbf{H}}} \cdot (e\vec{u}((a_0, a_{\mathbf{G}}), w))$$

Proof: Assume for contradiction that

$$\forall a'_{\mathbf{G}} \in \mathcal{A}^w_{\mathbf{G}}\ u\vec{_{\mathbf{H}}} \cdot (e\vec{u}((a_0, a'_{\mathbf{G}}), w)) \leq u\vec{_{\mathbf{H}}} \cdot (e\vec{u}((a_0, a_{\mathbf{G}}), w))$$

then using the definition of the neutral utility

$$u\vec{_{\mathbf{H}}} \cdot \vec{c} = \frac{1}{|\mathcal{A}^w_{\mathbf{G}}|} \sum_{a'_{\mathbf{G}} \in \mathcal{A}^w_{\mathbf{G}}} u\vec{_{\mathbf{H}}} \cdot e\vec{u}((a_0, a'_{\mathbf{G}}), w)$$
$$\leq u\vec{_{\mathbf{H}}} \cdot e\vec{u}((a_0, a_{\mathbf{G}}), w) \frac{1}{|\mathcal{A}^w_{\mathbf{G}}|} \sum_{a'_{\mathbf{G}} \in \mathcal{A}^w_{\mathbf{G}}} 1$$
$$= u\vec{_{\mathbf{H}}} \cdot e\vec{u}((a_0, a_{\mathbf{G}}), w)$$

The inequality holds thanks to the assumption. Now look at the definition of intentional adversariality. The second condition in the definition is

$$u\vec{_{\mathbf{H}}} \cdot (e\vec{u}((a_0, a_{\mathbf{G}}), w) - \vec{c}) < 0$$

hence

$$u\vec{_{\mathbf{H}}} \cdot e\vec{u}((a_0, a_{\mathbf{G}}), w) < u\vec{_{\mathbf{H}}} \cdot \vec{c}$$

This contradicts the inequality above and thus concludes the proof. □ ■

The presented classification expects the agent to have full information about the world. This is usually not true, the information an agent has can be partial, incorrect, or the agent can even overestimate its capabilities. In this case the expected utility of an action changes to what agent believes is their expected utility and the perceived neutral utility becomes the average outcome of all the actions the agent believes it can perform.

F. Traces of Actions

We can easily generalize the classification of actions to classification of runs in a multi-agent system in a straightforward way. The expected effect of the actions in a run on the utilities is the sum of the expected effects of the actions included in the run. Other definitions would be possible, but this one manifests the intentions of an agent with full information about the world, and is therefore most suitable for the classification of the interaction stance.

Definition 13. *Run* in a multi-agent system is a sequence of world states and actions

$$\rho = (w_1, a_1, w_2, a_2, ..., w_{k+1});\ w_i \in \mathcal{W}, a_i \in \mathcal{A}$$

where the state of the world is transformed from w_i to w_{i+1} via a_i.

Definition 14. Let $\rho = (w_1, a_1, w_2, a_2, ..., w_{k+1});\ w_i \in \mathcal{W}, a_i \in \mathcal{A}$ be a run of a multi-agent interaction, we define the *real utility* of the run

$$\vec{u}(\rho) = \vec{u}(w_k) - \vec{u}(w_1)$$

The *expected utility* of the run is

$$e\vec{u}(\rho) = \sum_{i=1}^{k} e\vec{u}(a_i, w_i)$$

The expected utility of a run can be classified in the same way as the expected utility of an action. By comparison of the of the utility vectors $u(\vec{\rho})$ and $e\vec{u}(\rho)$, we can analyze how predictable or in a way intentional was the real outcome of a sequence of actions.

G. Illustrative Example

To illustrate the use of the classification model introduced in the previous section, we apply it to a simulation of a flood relief operation. We consider three agents in the operation: the local government, humanitarian non-governmental organization (NGO) and separatists. The overall goal of the government, helped by the NGO is to stabilize the situation whereas the separatists want to take advantage of the situation is gain control over the affected regions.

Specifically, applying the proposed model, we identify the following utility components in the scenario: (1) the number of people with sufficient food and shelter, (2) the number of villages that are under the control of the government and (3) the state of the infrastructure damaged by the flood. One of the basic rules of the scenario is that the government cannot control a village, where people do not have enough food and shelter because of the riots that arise. The objectives of the three agents involved in the scenario are as follow. The

government wants to maximize all three utility components. The NGO cares only about maximizing the number of people with food and shelter. Finally, the separatists also care about people's well-being, but prefer the government not to have control. Consequently, they do not want the infrastructure repaired, because it gives tactical advantage to the government.

If all three agents value all the utility components equally, then the normalized preference vectors, using the order in which the components were introduced above, are:

$$\vec{u}_{GOV} = (0.58, 0.58, 0.58)$$
$$\vec{u}_{NGO} = (1, 0, 0)$$
$$\vec{u}_{GNG} = (0.58, -0.58, -0.58)$$

Some of the relevant actions in the scenario are delivering food to a village, destroying food supplies in a village or rebuilding infrastructure in a village. Below, we classify these actions with respect to the government and separatist agent. If we project the expected utilities of the actions to the plane generated by their utility vectors, we get the situation depicted in Figure 2. A typical cooperative action of one of the players

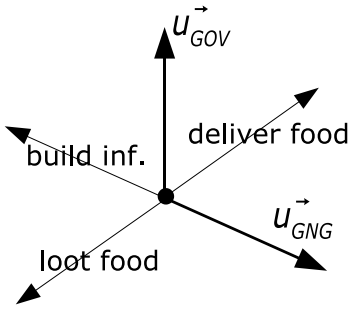

Fig. 2. Actions in the disaster relief scenario and their relation to the utilites of the Goverment and Separatists agent

towards the other is delivering food to starving people. The expected utility of the action is $(1, 0, 0)$ and it improves the utility of both players to the same extent. A competitive action of the government towards the separatists is e.g. rebuilding infrastructure, with the expected utility of $(0, 0, 1)$. The utility the government gains from this action is the same as the loss of the separatists. An adversarial action of the separatists is destroying food supplies in a village with the expected utility of $(0, -1, 0)$. It lowers both players' utilities equally, because the food cannot be delivered to the starving people. Still, the separatists may choose to perform such an action with the intention to harm the government and with the anticipation of riots and subsequent government's loss of control over the village, which would eventually increase separatists utility.

IV. Applications

In addition to providing formal grounding for a terminology used in categorising inter-agent relationships, the proposed scheme has several other applications.

A. Interaction Norms

Exploring how expected and/or allowed behaviours of agents in open multi-agent systems can be described, monitored and possibly enforced is an active research topic. Most of the developed formalisms, however, focus on defining norms as restrictions on actions an agent must or must not take in particular situations, without explicit reference to utilities of other agents.

The concept of interaction stance, as formally introduced in this paper, allows to specify norms which involve such a reference. It can be e.g. stipulated that every agent must be cooperative with respect to the head agent of the community, or that it must not be adversarial to any other member of the community[2]. Instead of prescribing literally which behaviours are allowed, such *interaction-based norms* allows prescribing the behaviour *relative* to other agents. The advantage of interaction-based norms in their flexibility – should the utility of the administrator agent change (e.g. due resource congestion arising in the system), individual member agents must adjust their behaviour accordingly (e.g. stop bandwith-intensive transfers).

An important property of the proposed classification scheme is that it is operationalizable. Whenever an agent performs an action in the system, it is categorized and depending on the resulting category, respective norms can be applied.

B. Agent Profiles

The interaction stance of an agent towards other agents (or classes thereof) can be made part of agent's profile. If combined with the knowledge of agent's base utility, a set of such profiles can provide for a compact representation of social relationships in agent community in way that allows reasoning about the behaviour of different (sub-)groups of agents in the system. Based on the profile of agents in a community and the profile of a potential new entrant, it can be determined whether the introduction of the new agent would benefit or harm the community.

C. Case-based Reasoning

The compact description using agent profiles, each combining agent's base utility and its interaction stance towards other agents, can be further used in a case-based reasoning system. When an agent is to operate in a community with a particular configuration of agents, it can search the case base for a situation in which the same or similar agents with the same or similar interaction stances were involved. For example, a humanitarian relief operation can proceed differently if farmers are cooperating than if they are adversarial.

D. Agent Design

Design of autonomous agents for open MAS is another application of the classification scheme. By taking into account the primary utilities of other agents in the system, the behaviour of the designed agent can be adjusted to implement

[2]Note that this is not possible for all combinations of agent's utlities.

the desired stance towards other agents in the system. Such an adjustment can be done off-line by a designer or on-line by the agent itself, provided agent's decision mechanism is flexible enough to realize such adjustment.

V. Related Work

The different factors affecting how agents choose actions they perform have been widely studied in the literature. The impact of limited decision-making capabilities has been explored within the topic of bounded rationality (see e.g. [3]). The role of the environment in affecting agent's ability to achieve its objectives has been long-studied in planning; the concept of joint action capturing the effect of other agents' actions on the desired outcome of the action performed by the agent has been long-known in game theory and multi-agent reinforcement learning.

The concept of adversariality has been studied from game theoretical perspective with applicability in economical theories and wargaming [7]. The repeated games approach with incomplete information and knowledge was used to model attackers and defenders in information warfare [8]. In robotics domains (especially robocup soccer) the adversarial actors are preventing the other actors from effective achieving their goals [9], [10]. An incentive-based modelling and inference of attacker intent, objectives, and strategies has been reported in e.g. [11]. Recently behaviour of adversarial agents in multi-agent domains has been defined via motivation to cause a drop of agents' social welfare, even at the cost of the adversarial agents individual utility [1]. However, so far there seems to be no computer-science literature attempting to formally ground the otherwise frequently used informal notions of adversariality or altruism and relating them to agent objectives defined as utility functions.

VI. Conclusion

In complex scenarios, agents do not always choose actions that lead optimally to the fulfilment of their objectives. The factors causing such a behaviour can be multiple, including agent's limited decision-making capability or the restrictive force of the environment. The paper focuses on the factor which has not yet been sufficiently addressed in this context – agent's social relationships to other agents. Agents may choose to perform actions not fully aligned with their objectives, if this helps or possibly prevents other agents from reaching *their* objectives.

In order to formalize such influence of social factors, we employ a formalism based on partially observable stochastic game to describe inter-agent interactions and define the concept of *interaction stance*, which categorizes agents' actions as self-interested, cooperative, competitive and adversarial. The classification is based on agents' preferences over different characteristics computed from the state of the system in consideration. We present two possible views of the classification. The first considers the effects of the actions of a group of agents towards other agents without taking into account actions available to the agents, whereas the other does take the real action options in account. The later approach allows differentiating between behaviours producing unintended negative effects and behaviours that are deliberatively adversarial. We show several applications of the model for the representation of norms, compact representation of social aspects of agent behaviour and representing social configurations of multi-agent systems.

In the future, we plan to include partial information about the world state, which can be represented via observations in the formalism, and extend the classification scheme to include agent beliefs about the world and the impact of their actions. Another important direction is the creation of methods for the reconstruction of agent utilities and their interaction stance from the observations of their actions in different social environments.

Acknowledgements

Effort sponsored by the Air Force Office of Scientific Research, USAF, under grant number FA8655-07-1-3083. The U.S. Government is authorized to reproduce and distribute reprints for Government purpose notwithstanding any copyright notation thereon.

References

[1] M. Pěchouček, J. Tožička, and M. Rehák, "Towards formal model of adversarial action in multi-agent systems," in *AAMAS '06: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. New York, NY, USA: ACM Press, 2006.

[2] M. E. Bratman, *Intentions, Plans, and Practical Reason*. Cambridge MA: Harvard University Press, 1987.

[3] A. Rubinstein, *Modeling Bounded Rationality*. The MIT Press, December 1997. [Online]. Available: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0262681005

[4] E. Hansen, D. Bernstein, and S. Zilberstein, "Dynamic programming for partially observable stochastic games," in *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, 2004, pp. 709–715.

[5] V. Könönen, "Multiagent reinforcement learning in markov games: Asymmetric and symmetric approaches." Ph.D. dissertation, Helsinki University of Technology, 2004.

[6] M. Rehák, M. Pěchouček, and J. Tožička, "Adversarial behavior in multi-agent systems," in *Multi-Agent Systems and Applications IV*, ser. Lecture Notes in Computer Science, M. Pěchouček, P. Petta, and L. Z. Varga, Eds., vol. 3690. Springer, 2005, pp. 470–479.

[7] P. Lehner, R. Vane, and K. Laskey, "Merging AI and game theory in multiagent planning," in *Intelligent Control, Proceedings of 5th IEEE International Symposium on*, 1990., pp. 853–857.

[8] D. Burke, "Towards a game theory model of information warfare," Master's thesis, Graduate School of Engineering and Management, Airforce Institute of Technology, Air University,, 1999.

[9] R. M. Jensen, M. M. Veloso, and M. H. Bowling, "Obdd-based optimistic and strong cyclic adversarial planning." in *In Proc. of ECP01*, 2001.

[10] T. F. Bersano-Begey, P. G. Kenny, and E. H. Durfee, "Multi-agent teamwork, adaptive learning and adversarial planning in robocup using a PRS architecture," in *IJCAI97*, 1997.

[11] P. Liu and W. Zang, "Incentive-based modeling and inference of attacker intent, objectives, and strategies," in *CCS '03: Proceedings of the 10th ACM conference on Computer and communications security*. New York, USA: ACM Press, 2003, pp. 179–189.