Towards Formal Model of Adversarial Action in Multi-Agent Systems*

Michal Pěchouček[1], Jan Tožička[1] and Martin Rehák[2]
Gerstner Laboratory[1] and Center for Applied Cybernetics[2]
Department of Cybernetics, Czech Technical University
Technická 2, Prague, 166 27, Czech Republic
pechouc@labe.felk.cvut.cz

ABSTRACT

Detecting and preventing the adversarial action of an agent with respect to the community of agents can be a serious issue in the design of open multi-agent systems. This task is severely domain dependent and it is hard to find a general solution. This contribution presents a utility based model of adversarial action and analyses few properties of adversarial behavior in multi-agent systems. Potential use and important drawbacks of this model are discussed in the paper.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

General Terms

Theory

Keywords

Adversarial action, Cooperation, Competitiveness, Formal model

1. INTRODUCTION

In this contribution, we use the conflict theory [4] and some fundamental principles from the economy and law [1] to consistently define the adversarial behavior in the multi-agent system. The agents in such system can have their own, sometimes completely antagonistic goals and they often compete for the shared resources or opportunities. Tackling a similar problem in conflict theory, James Fearon [4] analyzes the war between perfectly rational states, where the most important property of the war is the war's **ex-post inefficiency** – he argues that the states can reach the same result by negotiation, eliminating the cost of the adversarial actions. In [8], the aggression is defined as an action that is *socially undesirable and imposing net social cost*, consistently with [1], where the criminal activity is defined by imposing that *the gains of criminal are smaller than the social cost of act*.

AAMAS'06 May 8–12 2006, Hakodate, Hokkaido, Japan.

2. ADVERSARIAL ACTION DEFINITION

Consistently with previous work [2], we will define a simple abstract game model featuring agent set $Ag = \{A, B, C, ...\}$ with the agents playing a non-extensive (single round) game that is not strictly competitive – sum of all agents' utilities is not constant. Each agent X has a set of available actions denoted a^*_X, with actions $a^i_X \in a^*_X$ (whenever possible, we only write a_X). From this set, agent selects its action using its strategy. We suppose that all agents do the selection at one moment and therefore the selected actions are independent. The final state, *outcome* of the game $o(a_A, a_B, ...)$ is determined by strategies of the agents and determines both the individual agents' utilities $u_A(o), u_B(o), u_C(o), ...$ and the social choice function $u(o) = u_A(o)+u_B(o)+u_C(o)+...$, considered to represent the *social welfare* [3].

Utility is defined as *"a value which is associated with a state of the world, and which represents the value that the agent places on that state of the world"* by [7].

In the cooperative environment, all agents do share a single utility function.

DEFINITION 1. *We say that agent's A action a^{coop}_A is a* **cooperative action** *if it maximises social welfare:*

$$\texttt{coop}(a^{coop}_A) \Leftrightarrow u(a^{coop}_A, a_B, ...) = \max_{a^i_A \in a^*_A} u(a^i_A, a_B, ...)$$

The complete opposite is the self-interested environment, where the agents are trying to maximise their profit.

DEFINITION 2. *We say that agent's A action a^{si}_A is a* **self-interested action** *if it maximises agents' individual utility:*

$$\texttt{si}(a^{si}_A) \Leftrightarrow u_A(a^{si}_A, a_B, ...) = \max_{a^i_A \in a^*_A} u_A(a^i_A, a_B, ...)$$

In many contexts, the terms self-interestedness and competitiveness are considered to be synonymous. However, we consider the competitiveness to be more strict. In [2], self interestedness is defined as not taking the utility of the others into the consideration while maximizing own utility, while [5] requires the trust between competitors, allowing them to avoid globally undesirable outcomes. In the systems with carefully programmed mechanisms, the results are equivalent in both cases. However, in many real-world cases the total utility may decrease, even if each agent optimizes locally (a prison dilemma is an example of this situation).

In the competitive environment, agents select actions to maximize their own private utility, but they restrict their choice to the actions that at least conserve the social welfare.

DEFINITION 3. *We say that agent's A action a_A^{comp} is a* **competitive action** *provided that it maximizes individual profit while does not allow drop of the social welfare:*

$$\texttt{comp}(a_A^{comp}) \Leftrightarrow u_A(a_A^{comp}, a_B, ...) = \max_{a_A^i \in a_A^{**}} u_A(a_A^i, a_B, ...),$$

where $\forall a_A^i \in a_X^{**} : u(a_A^i, a_B, ...) \geq u(a_B, ...)$.

The expression $u(a_B, ...)$ represents hypothetical outcome of the community in the case when the agent A performs no action or is not in the community at all. This situation is illustrated later on Figure 1 by $\sim$ symbol.

Finally, let us try to define the concept of the adversarial action. The most intuitive definition would be that the adversarial action is such an action that is deliberatively preferred to another action that is equally achievable but has got higher social welfare utilities.

DEFINITION 4. *We say that agent's A action a_A^{adv} is an* **adversarial action** *if:* $\texttt{adv}(a_A^{adv}) \Leftrightarrow \exists a_A^i \in a_A^* :$

1. $u(a_A^{adv}, a_B, ...) \ll u(a_A^i, a_B, ...)$ *and*
2. $u(a_A^i, a_B, ...) - u(a_A^{adv}, a_B, ...) \gg u_A(a_A^{adv}, a_B, ...) - u_A(a_A^i, a_B, ...)$.

The definition 4 above states that the adversarial action a_A^{adv} selected by A from the set a_A^* hurts the social welfare without strong incentive. To make the formalism simpler, we have assumed that there is only single action a_A^{adv} of agent A that hurts the social welfare. There are several interesting points to consider in the general definition.

The first point is the non-emptiness of the set $a_A^* \setminus \{a_A^{adv}\}$ - we don't consider the behaviour with no alternative as adversarial.

Motivation and justification of the adversarial action is closely related to two relational operators used in the definition: $\ll$ and $\lesssim$. The first inequality $\ll$ signifies that the agent shall not cause significant harm to the common welfare, while the inequality $\lesssim$ means that the agent remains self-interested and it will not lose a significant part of its welfare to save the utility of other agents. The concept is illustrated by Fig. 1. In this context, it is important not to take our simplification of the game formalism literally and to consider only immediate payoff as the utility – in most systems, agents expect to encounter their partners again in the future and we suppose that the attitudes of their partners towards them and expected future profits are included in the utility u_X. Formally, we may pose:

DEFINITION 5. *We say that* ***action*** *a_A^j of agent A is* ***rationally adversarial*** *if it is both self-interested and adversarial. In the action is not self-interested and is adversarial, it is* ***irrationally adversarial****.*

2.1 Properties of Adversarial Action

2.1.1 Pareto-Optimality

In this context, we may mention the relationship between adversariality and Pareto-Optimality[1].

An outcome of an adversarial action is not Pareto optimal. Rationally adversarial action is not Pareto optimal in the situations

[1]Following [6], we denote as o^* a set of all achievable outcomes and we define: Outcome o is considered to be **Pareto optimal** if : (i) it is achievable (i.e. $o \in o^*$) and (ii) not majored by any other outcome $o' \in o^* \setminus \{o\}$, where we define majoring as: $\forall_{X \in Ag} u_X(o') \geq u_X(o)$ and $\exists_{X \in Ag} u_X(o') > u_X(o)$.

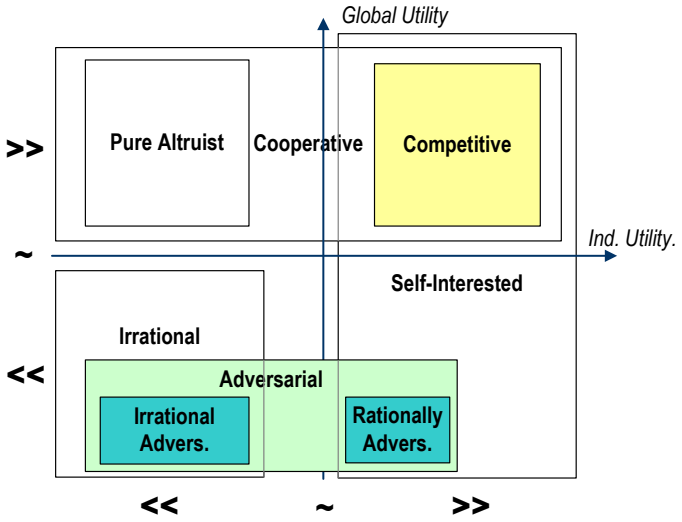

Figure 1: Classification of action with respect to global utility (social welfare) and individual utility of acting agent. The $\sim$ symbol corresponds to the situation where the agent A performs some neutral action or is not in the community at all, i.e. $u(a_B, ...)$.

where the agents may negotiate and transfer the utility - in such situations, the agents may always transfer enough utility to motivate the adversarial agent to behave cooperatively, therefore achieving socially acceptable outcome. When the utility is not transferable (e.g. indivisibility as defined in [4]), the set o^* is severely restricted and even an action that causes the overall social loss may be considered non-adversarial due to the lack of alternative. In the irrationally adversarial case, Pareto optimality does not hold neither, as the utility is lost both by adversarial agent and the society as a whole.

As stated above, an outcome of an adversarial action is not Pareto optimal in the situations:

- where the agents may negotiate and transfer the utility, and
- the action is irrationally adversarial – $u_A(a_A^{adv}) < 0$.

This property can be used mainly in the situation where the community consists of a coalition of mutually trusted actors and one agent whose adversariality is subject of investigation.

On the other hand, Pareto optimality as such doesn't preserve social welfare (due to the indivisibility), it only ensures that all agents behave rationally given the knowledge about the action of the others.

2.1.2 Implication of Adversarial Behaviour

We are interested in the impact of the adversarial action on the global social welfare of the community Ag. We can say that:

- decrease of social welfare *does not imply* existence of an adversarial behavior in the community,
- unnecessary decrease of social welfare *implies* existence of an adversarial behavior (intentional or unintentional) in the community, while
- existence of an adversarial behavior in the community *does not imply* decrease of social welfare.

3. EXAMPLE: UAV DECONFLICTION

In this example, we will illustrate rather abstract definitions of adversarial action provided above with the real example from domain of UAV (Unmanned Aerial Vehicle) flight-plan deconfliction. Let us operate two UAVs A and B. We have situation where the two UAVs are facing a collision and they individually deliberate about the actions $d(A)$ – the UAV A making the deconfliction manoeuver, the action $d(B)$ – the UAV B making the deconfliction manoeuver and $d(A, B)$ – the both UAVs making the deconfliction manoeuver. As the UAVs are different, they loss of their individual utility associated with the manoeuver is different. Let us assume:

- $mu_{\mathtt{d}(A)}(A) > mu_{\mathtt{d}(B)}(B) > mu_{\mathtt{d}(A,B)}(B) + mu_{\mathtt{d}(A,B)}(A)$
- $mu_{\mathtt{d}(A,B)}(A) > mu_{\mathtt{d}(A)}(A)$,
- $mu_{\mathtt{d}(A,B)}(B) > mu_{\mathtt{d}(B)}(B)$

The A UAV is smaller than the B UAV, which implies that it is cheaper for A to make the manoeuver than it is for B. However B making the manoeuver is even cheaper than sum of costs for both UAVs making the manoeuver. However, for each UAV it is cheaper to participate in a collective manoeuver than doing the manoeuver individually.

In the *cooperative environment* the UAVs with conflicting plans do minimize overall disruption and fuel consumption while solving deconfliction problems[2]:

- $A, B : \mathtt{d}(A) \succ \mathtt{d}(B) \succ \mathtt{d}(A, B) \succ \neg\mathtt{d}(A, B)$

Both the cooperative UAVs have the same strategy, that is the smaller UAV deconfliction is preferred to the bigger UAV deconfliction that is preferred to both plans deconfliction and which is preferred to a confliction.

In the *competitive environment* each UAV minimizes its own plan disruption and fuel consumption, but conserves overall welfare:

- $A : \mathtt{d}(B) \succ \mathtt{d}(A) \succ \mathtt{d}(A, B) \succ \neg\mathtt{d}(A, B)$
- $B : \mathtt{d}(A) \succ \mathtt{d}(B) \succ \mathtt{d}(A, B) \succ \neg\mathtt{d}(A, B)$

More specifically, if two UAVs would collide and only one evasive maneuver is necessary, it will try to make the other UAV divert from its course, but without compromising the security.

In the *self-interested environment* each UAV minimizes its own plan disruption and fuel consumption, regardless of the others.

- $A : \mathtt{d}(B) \succ \mathtt{d}(A, B) \succ \mathtt{d}(A) \succ \neg\mathtt{d}(A, B)$
- $B : \mathtt{d}(A) \succ \mathtt{d}(A, B) \succ \mathtt{d}(B) \succ \neg\mathtt{d}(A, B)$

In case of conflict, it only diverts from its course to protect its own safety.

In the adversarial environment the adversarial UAV causes a significant disruption of the other's plans, or even endangers them. Let us assume $mu_{\mathtt{d}(A)}(A) \ll mu_{\mathtt{d}(B)}(B)$ (e.g. A is pushing B away).

- $A : \mathtt{d}(B) \succ \neg\mathtt{d}(A, B) \succ \{\mathtt{d}(A, B), \mathtt{d}(A)\}$
- $B : \mathtt{d}(A) \succ \mathtt{d}(A, B) \succ \mathtt{d}(B) \succ \neg\mathtt{d}(A, B)$

If the B flying over an enemy area does not want to be pushed, the situation is as follows:

- $A : \mathtt{d}(B) \succ \neg\mathtt{d}(A, B) \succ \{\mathtt{d}(A, B), \mathtt{d}(A)\}$
- $B : \mathtt{d}(A) \succ \mathtt{d}(A, B) \succ \neg\mathtt{d}(A, B) \succ \mathtt{d}(B)$

[2]The symbol $\succ$ stands for collective choice preference

4. CONCLUSION

The problem of adversariality in the multi-agent systems is real. While the irrationally adversarial agents may be easy to identify, it may be much more difficult to identify the rationally adversarial behavior, especially if all the agents in the system are self-interested. In this context, the question of *bounded rationality* of agent's reasoning is crucial. To better illustrate the concept, we will cite several accepted causes for the emergence of the conflict between the rational actors. It is easy to realize that most of these causes can plausibly exist in the multi-agent system and shall be considered while designing autonomous agents: **Private information** of each agent is not available to the others, providing one of the causes of **miscalculation** about **capabilities or attitudes** of the other party. Such miscalculation may cause an adversarial behavior, as the agents will not be able to correctly estimate the utility function of the partners. Agents are often willing to **misrepresent** the reality about themselves, in order to obtain better payoff or negotiation position in the future.

5. ACKNOWLEDGEMENT

Effort sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA8655-04-1-3044 and FA8655-04-1-3044-P00001. The U.S. Government is authorized to reproduce and distribute reprints for Government purpose notwithstanding any copyright notation thereon[3]. We also gratefully acknowledge the support of the presented research by ARL project N62558-03-0819.

6. REFERENCES

[1] Gary S. Becker. Crime and punishment: An economic approach. *The Journal of Political Economy*, 76(2):169–217, 1968.

[2] Sviatoslav Brainov. The role and the impact of preferences on multiagent interaction. In *ATAL '99: 6th International Workshop on Intelligent Agents VI, Agent Theories, Architectures, and Languages (ATAL),*, pages 349–363. Springer-Verlag, 2000.

[3] Rajdeep K. Dash, Nicholas R. Jennings, and David C. Parkes. Computational-mechanism design: A call to arms. *IEEE Intelligent Systems*, 18(6):40–47, 2003.

[4] James D. Fearon. Rationalist explanations for war. *International Organization*, 49(3):379–414, 1995.

[5] D. Gambetta, editor. *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell, 1990.

[6] Milan Mares. Fuzzy coalition structures. *Fuzzy Sets Syst.*, 114(1):23–33, 2000.

[7] Simon Parsons and Michael Wooldridge. Game theory and decision theory in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 5(3):243–254, 2002.

[8] Eric A. Posner and Alan O'Neil Sykes. Optimal war and jus ad bellum, 2004.

[3]The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.