Adversarial Behavior in Multi-Agent Systems

Martin Rehák, Michal Pěchouček, Jan Tožička

Department of Cybernetics, Czech Technical University in Prague
Technická 2, Prague 6, 166 27 Czech Republic
{rehakm1|pechouc|tozicka}@labe.felk.cvut.cz

Abstract. Adversariality of the agents with respect to the multi-agent system can be a serious issue in the design of open multi-agent systems. Until now, many incoherent definitions of such behavior were used, preventing the consolidation of the knowledge about the domain. By basing ourselves on the valid and accepted results from economics, law and conflict theory, we propose a consistent definition of adversariality in the multi-agent systems and discuss the characteristics of the behavior that falls into this definition.

1 Introduction

The current trend in the multi-agent systems field is to emphasize the openness of systems, their ad-hoc integration capability and to capitalize on their syntactic and semantic interoperability. In open environments, we can no longer assume that the agents are cooperative. The agents in these system can have their own, sometimes partially or completely antagonistic goals and they often compete for the shared resources or opportunities.

In such environments, we must ensure that the system as a whole will autonomously maintain its sustainability and efficiency, that self-interested agents will be able to agree at least on some goals and that their cooperation will leverage their capabilities. To do so, agent researchers frequently introduce the concepts from microeconomics and game theory, most notably mechanism design [1]. Mechanism design is used to design interaction patterns in the system to promote globally desirable behavior and reduce incentive for undesirable behavior. However, despite the fact that it will provide the basis of the algorithms and protocols of such systems, it still suffers from some serious limitations. Mechanism design techniques have achieved some spectacular results, but their applicability is in general restricted to static environments, where the fine-tuned mechanisms perform well. However, the problems like bounded rationality of the agents, their possible polyvalence, strategic behavior and willingness to keep some of their knowledge private can not be completely addressed by the current mechanisms [2].

Alternatively, similar results can be achieved achieved using norms [3], enforcing flexible social commitments [4], adjustable policies [5] or trust and reputation [6, 7]. But in general, these approaches rely on the fact that the agents are able to distinguish the undesirable behavior in all possible contexts. Therefore, as

the system adapts to its environment, the norms, policies and trust mechanisms must be adapted as well to avoid becoming an obstacle of system efficiency, rather than to support it.

In this contribution, we will look at the problem from somewhat different perspective – after the brief analysis of existing approaches in the multi-agent field, we will use the conflict theory and some fundamental principles from the economy and law (Section 2) to consistently define the adversarial behavior in the multi agent system (Section 3) and provide a specific example that instantiates the definition in Section 4.

Currently, adversariality in the multi-agent systems is a concept that has been defined in many different contexts. Most of the current definitions are mutually exclusive, but they provide a valuable guidance in our attempt to formalize the definition using their overlaps.

In the field of multi-agent systems, **adversarial planning** was introduced [8]to analyze the behavior of two opponents. However, even if the approach remains interesting due to the analysis of planning in conflicting environment, it is of limited importance for the definition of adversarial behavior. The definition proposed by the authors, where they define adversariality by "opposite goals" doesn't fit our needs, as the agents in the general system we consider (i) are not always adversarial and at least some of their goals are common, (ii) communicate by other means than pure actions, (iii) have asymmetric and partial knowledge and (iv) are deliberative, therefore possibly adversarial within the limited scope of time or issues.

In the mechanism-design field, [2] defines adversarial entities as the entities who's goals can not be described by a utility function and assumes these actors to be irrational. This definition well captures the fact of bounded rationality of agent perceptions - some agents can have goals that are impossible to capture and understand during normal system operations and that are justified by large scale (time or space) behavior of their owners.

2 Conflict Theory and Economics of Conflict

We shall use the conclusions from the field of the conflict theory to (i) determine the defining properties of adversariality as they are currently understood.

In his contribution [9], James Fearon analyzes the war between two or more perfectly rational states. For Fearon, the most important distinguishing property of the war from the rationalist point of view 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: *"...ex-post inefficiency of war opens up an ex-ante bargaining range..."*([9], page 390). This is clearly visible from the simple conflict specification proposed by author.

In the work of **Posner** and **Sykes** [10], approaching the problem of optimal war from the legal perspective, the aggression (unilateral beginning of the war) is defined as an action that is *socially undesirable and imposing net social cost*, while the authors assume that the aggression is motivated by the expected profit

of the aggressor, either as a result of war or the threat. They argue that this definition of aggression is consistent with the studies on the economics of crime [11], where the *gains of criminal are smaller than the social cost of act.*

In his breakthrough article, **Gary Becker** [11] analyzes the economics of crime, incentives of criminals, their economic motivation and dissuasive effect of punishments and functional justice system. Besides the definition of criminal activity stated above, the notion of indirect costs is also important. Costs of crime are not only direct, but we must consider the cost of law enforcement as inseparable from the direct crime costs. In a multi-agent system, the well designed mechanisms and trust maintenance models come with a cost that may harm the system efficiency through their computational requirements and other associated requirements. This doesn't mean a refusal of the principle of trust maintenance and mechanism design, but it means that the mechanism must be efficient and well adapted to the current environment.

3 Adversarial Behavior Definition

This section is devoted to the formal definition and characterization of adversarial behavior in the multi-agent systems. We will depart from the conflict theory premise stated above that conflict is an *ex-post inefficient* method of resolving competitive issues that imposes a net cost on the society, and we will base our formal definition on these notion. Similar classification was done in [12], but focused on interaction between different types of agents rather than on definition of types of behavior and didn't use the conflict theory. However, some preliminary technical definitions are necessary.

In the following, we will use capitals to denote agents.

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 [13].

To simply state our problems, 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 with 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. The final state, *outcome* of the game[1] $o(a_A, a_B, ...)$ 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*[1].

In this simplistic game, we can define cooperative, competitive and adversarial behavior in accordance with the principles from section 2. Simplified graphical form of the definitions is presented in Fig. 1.

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

[1] The exact form of the outcome is irrelevant, if we are able to obtain the utility values. To simplify the notation, we will also write $u(a_A, a_B, ...)$ instead of technically more correct $u(o(a_A, a_B, ...))$.

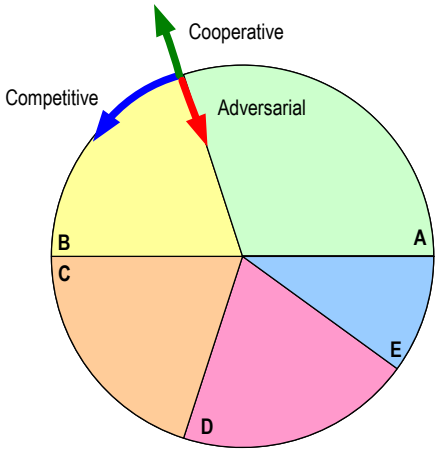

Fig. 1. Cooperative, Competitive and Adversarial actions. Pie represents the total utility u and individual utilities $u_A, u_B, ...$ We can see that purely cooperative action increases social welfare, purely competitive action doesn't modify the social welfare, but only changes its distribution among agents, while the purely adversarial action reduces the social welfare without any benefit for the agent. In practice, real actions are rarely pure and are a combination of the above types.

Definition 1 *We say that agent's A action* a_A^{coop} *is a* ***cooperative action*** *provided that* $a_A^{coop} = \arg\max_{a_X^*} u(a_A^{coop}, a_B, ...)$.

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 2 *We say that agent's A action* a_A^{comp} *is a* ***competitive action*** *provided that* $a_A^{comp} = \arg\max_{a_X^{**}} u_A(a_A^{comp}, a_B, ...)$*, where the set* a_X^{**} *contains the actions* $a_A^i \in a_X^*$ *that conserve or increase the social welfare* $u(a_A^i, a_B, ...)$.

In many contexts, the above terms self-interestedness and competitiveness are considered to be synonymous. However, we consider the competitiveness to be more strict - in [12], self interestedness is defined as not taking the utility of the others into the consideration while maximizing their own utility, while [14] 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 (see [15] for a nice analogy).

Definition 3 *We say that agent's A action* a_A^{si} *is a* ***self-interested action*** *provided that* $a_A^{si} = \arg\max_{a_X^*} u_A(a_A^{si}, a_B, ...)$.

And finally, the adversarial action is defined as an action that significantly decreases the social welfare while it causes loss or provides only small profit to the actor of the action.

Definition 4 *We say that agent's A action* a_A^{adv} *is an* ***adversarial action*** *if* $\exists a_A^i \in a_A^* : i \neq adv$ *such that* $u(a_A^{adv}, a_B, ...) \ll u(a_A^i, a_B, ...)$ *and* $u_A(a_A^{adv}, a_B, ...) \lesssim u_A(a_A^i, a_B, ...)$.

The definition 4 above states that the adversarial action a_A^{adv} selected by A from the set a_A^* of 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 $a_A^* \setminus \{a_A^a dv\}$ - we don't consider the behavior 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 $\lesssim$[2] 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. 2. 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[3]. 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****.*

In this context, we may mention the relationship between adversariality and Pareto-Optimality: [4]

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

[2] We actually mean that the agent has no, or very little motivation to make an adversarial move. In Def. 5, we treat the special case when we fall into the $\sim$ case.

[3] In this point, we are consistent with the utility definition given above. We have omitted the explicit future gains member in the definitions to simplify the notation by using this broader definition of utility.

[4] Following [16], 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 $\exists_{X \in Ag} u_X(o') > u_X(o)$.

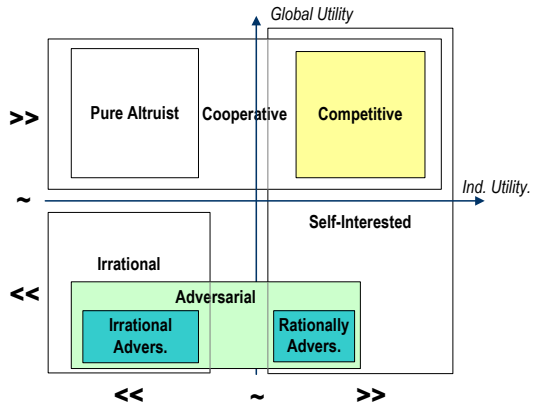

Fig. 2. Classification of action with respect to global utility (social welfare) and individual utility of acting agent.

transferable (e.g. indivisibility as defined in [9]), 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.

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.

Another point to address is the predictability of the outcome. The uncertainty of o arises from the simultaneity of all players' moves, while the uncertainty of values $u_X(o)$ and $u(o)$ exists due to the privacy of functions u_X. This seems to make the definition useless – but social knowledge and norms can provide solutions. In most situations, the individuals are able to estimate the actions of others (denoted a_X^{exp}) and the effects of different outcomes on their utility.

Therefore, without considering norms, we pose:

Definition 6 *We say that* ***action*** a_A^{ia} *of agent* A *is* ***intentionally adversarial*** *if the action is adversarial and the agent* A ***knows that*** $\exists a_A^i \in a_A^* : i \neq ia$ *such that* $u(a_A^{ia}, a_B^{exp}, ...) \ll u(a_A^i, a_B^{exp}, ...)$ *and* $u_A(a_A^{ia}, a_B^{exp}, ...) \lesssim u_A(a_A^i, a_B^{exp}, ...)$. *Otherwise, the action is* ***unintentionally adversarial****.*

More specifically, the lack of norms or conventions is a possible cause of unintentional adversariality – the adversarial outcome may arise due to the limited computational power or knowledge of agents, private knowledge or the environmental noise. Important question of attribution must be solved by each agent – we can not expect that all agents will agree on the cause of the common loss.

Existence of shared normative system reduces the uncertainty regarding the expected actions of other agents (a_X^{exp}). In our future work, we will use this system in the adapted definition of adversarial action. On the other hand, definition

4 remains valid, as it provides feedback for update of the normative system in changing environment.

So far, we have defined adversarial action, rational adversarial action and intentional adversarial action. However, we still have to define *adversarial agent*.

Definition 7 *We say that* ***agent*** *A is* ***adversarial*** *(or there exists* ***adversarial behavior*** *performed by the agent A) if the agent A performed at least one adversarial action in the past – $adv(A) \Leftrightarrow \exists a_A^{adv}$: so that $adv(a_A^{adv}) \wedge \mathsf{P}(\texttt{Perform}\ A\ a_A^{adv})$.*

In the definition, we assume that the predicate *adv* classifying the actions is defined according to the property 4, the P operator to be a temporal logic operator representing validity of a formula in the past and the operator `Perform` linking an agent and the action performed by the agent.

There are clear extensions of this definition of adversarial behavior that define adversariality in a time window, or agent's adversarial behavior with relation to a specific agent community. In the definition 7 we assume by default the whole of the community as a target of agent's adversariality and the whole past as the relevant time window.

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

- decrease of social welfare implies existence of an adversarial behavior in the community, while
- existence of an adversarial behavior in the community does not imply decrease of social welfare.

For the proof of these statements, let us consider only types of actions according to the definitions 1, 2 and 4. No combination of cooperative and competitive actions may cause an overall decrease of the social welfare, thus an existence of at least one adversarial action is inevitable. In contrary, for a combination of adversarial actions there may exist a compensating combination of cooperative or competitive actions that can be carried out by any member of the community in the finite time t so that in t the social welfare does not decrease.

The definition 7 does not classify performance of an action that has got a direct inevitability (or possibly an option) of an adversarial action as its effect as adversarial behavior.

4 Example: Adversariality in Coalition Formation

In this example, we will illustrate rather abstract definitions provided above with the real example, the coalition formation, approaching the problem from the utility side. We will start by introducing the necessary notation. In this section, we consider the coalition to be short-lived and therefore the terms adversarial action of agent A and adversarial agent A will be used interchangeably.

Using the concept of the marginal utility[5], we may now define cooperative and competitive behavior in our example.

We say that agent A is **collaborative** provided that: if an agent A makes an attempt to join the coalition C then always $mu_{A \mapsto C}(C) > 0$. We shall note that even if all agents are collaborative, the optimum result is not guaranteed. A typical case can be described as follows: $mu_{B \mapsto C}(C \cup B) > mu_{A \mapsto C}(C \cup A) \geq 0$ and $mu_{B \mapsto (C \cup A)}(B) < 0$. If A joins the coalition first, it blocks the entry of B and only local optimum is reached.

We say that agent A is **competitive** provided that: if an agent A makes an attempt to join the coalition C then always $mu_{A \mapsto C}(A) > 0$ and $mu_{A \mapsto C}(C) \geq 0$. Similarly, we say that agent A is **self-interested** provided that: if an agent A makes an attempt to join the coalition C then always $mu_{A \mapsto C}(A) > 0$.

As we have already stated before, self-interested agent considers only its own profit while it takes coalition entry decision. Competitive agent is both self interested and collaborative, as it maximizes its own profit, but it at least maintains the social welfare that is represented by the coalition utility. Therefore, in both competitive and cooperative behavior, the social welfare is maintained. This is not necessarily true in the self-interested or adversarial behavior.

In this example, we will use the marginal utility defined above to define adversarial behavior. We say that an agent is **adversarial** provided:

- $mu_{A \mapsto C}(A) \lesssim 0$
- $mu_{A \mapsto C}(C) \ll 0$
- agent A makes an attempt to join the coalition C

Informally, an agent is adversarial with respect to coalition C provided that the increase of his direct marginal utility is significantly smaller than the harm (decrease of the total payoff) caused to the coalition.

If the condition $mu_{A \mapsto C}(A) \geq 0$ holds, agent's action is rationally adversarial, otherwise it is irrationally adversarial, as defined in definition 5.

Main advantage of the above definition is that it provides a basis for the detection of adversarial agents, by defining the metrics measuring the adversariality.

Gathering and maintaining such experience is not trivial. However, we may reuse the existing work on trust, where one of the components of the trust[6] - intentional trust (willingness)- is a complement of intra-community adversariality defined above. Therefore, if we establish a reasonable value for trust (that may be actually lower, due to the capability trust), we may deduce an acceptable estimation of agent's adversariality.

[5] Agent's A *marginal utility* (mu) from joining the coalition C (an activity denoted as $A \mapsto C$) is a derivation of the agent's utility before and after it joins the coalition ($mu_{A \mapsto C}(A) = u_{A \in C}(A) - u_{A \notin C}(A)$, where $u_{A \in C}(A)$ is a utility the agent A (in parentheses) receives as a member of the coalition C (situation is described by subscript), while $u_{A \notin C}(A)$ denotes the utility agent A receives if it doesn't join the coalition C). The marginal utility of a coalition C in agent's A joining the coalition is defined as a derivation of the collective utility (such as social welfare) of the coalition before and after the agent joins the coalition ($mu_{A \mapsto C}(C) = u(C \cup A) - u(C)$).

5 Conclusion

The definition of the adversarial behavior that we present provides a useful complement of the current approaches to the open systems engineering. Even if the system is based on carefully designed mechanisms and/or norms, the changing system social structure and the environment or agent's strategic behavior may modify the system and make it inefficient or dysfunctional. To counter such danger, the agents in the system shall continuously monitor the behavior of the others and their own and detect potentially adversarial actions. As soon as these actions are identified, protocols or normative systems can be altered to counter the undesirable tendencies, or the adversarial agents can be completely cut-away from the system. Such detection can be done on peer-to-peer basis, but can be also entrusted to dedicated agents that would implement not the norm enforcement, but norm creation and maintenance.

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. For example, some agents may be willing to leave the local optimum to bring the system into the globally optimal (or simply better) state. However, if the other agents in the system lack this insight, they may consider this behavior as adversarial because they fail to see the long-term benefits. 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. However, if such behavior becomes widespread in the system (It can be often prevented by careful mechanism design.), agents are unable to communicate efficiently. In the more sophisticated extension of this behavior, agents can behave strategically and harm the others to gain higher relative power in the long term. In some situations, the system may even become purely competitive – agents or their groups have nothing to gain from cooperation, for example when the payoff is indivisible.

Acknowledgement

Effort sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA8655-04-1-3044. The U.S. Government is authorized to reproduce and distribute reprints for Government purpose

notwithstanding any copyright notation thereon[6]. We also gratefully acknowledge the support of the presented research by ARL project N62558-03-0819.

References

1. Dash, R.K., Jennings, N.R., Parkes, D.C.: Computational-mechanism design: A call to arms. IEEE Intelligent Systems **18** (2003) 40–47
2. Feigenbaum, J., Shenker, S.: Distributed algorithmic mechanism design: Recent results and future directions. In: Proceedings of the 6th International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications, ACM Press, New York (2002) 1–13
3. Conte, R., Castelfranchi, C.: From conventions to prescriptions - towards an integrated view of norms . Artif. Intell. Law **7** (1999) 323–340
4. Pasquier, P., Flores, R., Chaib-draa, B.: Modeling flexible social commitments and their enforcement. In Gleizes, M.P., Omicini, A., Zambonelli, F., eds.: Proceedings of Engineering Societies in the Agents World V, Toulouse, October 2004. Number 3451 in LNAI, Springer-Verlag, Heidelberg (2005) 153–165
5. Suri, N., Carvalho, M.M., Bradshaw, J.M., Breedy, M.R., Cowin, T.B., Groth, P.T., Saavedra, R., Uszok, A.: Enforcement of communications policies in software agent systems through mobile code. In: POLICY. (2003) 247–250
6. Castelfranchi, C., Falcone, R.: Principles of trust for mas: Cognitive anatomy, social importance, and quantification. In: Proceedings of the 3rd International Conference on Multi Agent Systems, IEEE Computer Society (1998) 72
7. Ramchurn, S., Huynh, D., Jennings, N.R.: Trust in multiagent systems. The Knowledge Engineering Review **19** (2004)
8. Willmott, S., Bundy, A., Levine, J., , Richardson, J.: An adversarial planning approach to go. In: Proceedings of the Firstrst International Conference on Computers and Games, Springer-Verlag, LNCS 1558 (1998) 93–112
9. Fearon, J.D.: Rationalist explanations for war. International Organization **49** (1995) 379–414
10. Posner, E.A., Sykes, A.O.: Optimal war and jus ad bellum (2004)
11. Becker, G.S.: Crime and punishment: An economic approach. The Journal of Political Economy **76** (1968) 169–217
12. Brainov, S.: 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),, Springer-Verlag (2000) 349–363
13. Parsons, S., Wooldridge, M.: Game theory and decision theory in multi-agent systems. Autonomous Agents and Multi-Agent Systems **5** (2002) 243–254
14. Gambetta, D., ed.: Trust: Making and Breaking Cooperative Relations. Basil Blackwell (1990)
15. Goldratt, E.M.: The Theory of Constraints. N.Y.: North River Press, Croton-on-Hudson, N.Y. (1990)
16. Mares, M.: Fuzzy coalition structures. Fuzzy Sets Syst. **114** (2000) 23–33

[6] 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.