No Agent Is an Island: A Framework for the Study of Inter-Agent Behavior

T.J.M. Bench-Capon
Department of Computer Science
The University of Liverpool
Liverpool
UK
044-151-3697
tbc@csc/liv.ac.uk

P.E. Dunne
Department of Computer Science
The University of Liverpool
Liverpool
UK
044-151-3521
ped@csc/liv.ac.uk

ABSTRACT

We describe a framework for the study of inter-agent behavior. Our starting point is the notion that choosing to perform an action will constrain the capacity to choose other actions, both for the agent concerned and for the other agents with which it interacts. We represent these conflicts between choices in an abstract fashion using an "option framework". In an option framework we are not concerned with the nature of the actions that can be chosen, only the ways in which they conflict with the other choices in the framework.

We can then partition the options in the framework according to the agent which can select them, and associate utilities, with respect to all the agents in the system, with them. Agents will then select options according to the constraints imposed by conflicts between actions. Where the choices of two agents conflict, the conflict is resolved according to which agent controls the conflict, resulting in the realization of some subset of the choices. The agent can then evaluate the realized actions according to a function which may take into account the utility produced for itself and other agents. The task of the agent is to select the set of actions which produces the subjectively most favored realization.

Having formally presented the framework we show how it can be used to explore the inter-agent behavior of systems of agents according to a number of factors which will determine how they go about their task. We also show how some other approaches to the investigation of the inter-agent behavior can be modeled in our framework.

The framework is sufficiently abstract to provide the means to explore all aspects of inter-agent behavior by both empirical and analytic means. We give examples of some hypotheses that may be investigated in this framework.

1. INTRODUCTION

What can we do? We have many options: I *can* go to a Chinese restaurant or to a Spanish restaurant for dinner tonight. I *can* go to Ireland or to France for my holiday next year. But, of course, I cannot do *all* these things. If I dine in the Chinese restaurant, I cannot dine in the Spanish restaurant: if I travel to Ireland, I cannot holiday in France. When choosing what to do, I need to recognize that what I choose will constrain my options, and I may need to ignore an option, desirable in itself, in order to keep other options open. So, when making my choices, I need to take account of the larger picture, taking account not only of the action itself, but its effects on my other actions.

Neither can I consider myself alone. John Donne wrote that "No man is an island, entire of itself". Jean-Paul Sartre wrote that "L'enfer est les autres", that hell is other people. The Rolling Stones sang "You can't always get what you want to". As writers, philosophers and musicians through the centuries have recognized, what I choose also constrains the choices of others, and what others choose, constrain my choices.

So we might replace the question we started with by the question "what can we choose?", so as to recognize that our options are restricted to those that can be done when constrained by the need for our choices to be compatible, and to be compatible with the choices of others. Or we could address the question "what *should* we choose", to recognize both the different value to us of the various co-realizable possibilities, and the fact that our choices will constrain the choices of others.

In the light of this interconnectedness of human activity, a number of social organizations and moral codes have developed, to guide and regulate inter-personal behavior so as to promote peaceful co-existence and even mutual benefit. If we are to take the notion of societies of agents seriously, we need to recognize that this interconnectedness will hold for agents also. In this paper we develop a framework for reasoning about agent choices and the regulation of inter-agent behavior.

Section 2 provides an abstract formal model to express the choices available to a group of agents, the internal, external and inter-agent constraints between them, and the benefits that they have for the various agents. This will enable us to define the problem faced by the agents of selecting from their options some subset of these options which they will attempt to perform.

Section 3 will discuss a variety of different constraints that can be added to the formal framework, including the notions of how agents compare different choices, strategies for making the choice, and possibilities for regulating inter-agent behavior.

Section 4 shows how the framework can model the kinds of situations explored in other work attempting to explore inter-agent behavior. Section 5 illustrates the use of the framework by considering the special cases of a single agent, and interaction between two agents.

Section 6 discusses future work, involving the extension to cases with arbitrarily many agents, and puts some hypotheses which can be tested using the framework. Section 7 offers some concluding remarks.

2. A FRAMEWORK FOR THE INVESTIGATION OF INTER-AGENT BEHAVIOR

We begin by introducing the notion of an *option.* For our purposes we want to keep this notion as abstract as possible. Thus we are not interested in the details of the action which can be opted for, nor in any of the mechanics of how it might be carried out. An option is simply something which an agent can do, which can have some utility for one or more agents and which can, if performed, block the selection of other options. The options which are available to the agents under consideration, and the conflicts between them are denoted by an *Option Framework*.

Definition 1. An *option framework* is a pair

OF = <OP,conflicts>

where *OP* is a set of options, and *conflicts* is a *symmetric non-reflexive* binary relation on OP, i.e. *conflicts Í OP ´ OP.*

For two options op_1 and op_2, the meaning of *conflicts(op_1, op_2)* is that op_1 and op_2 can never both be chosen together. This may be so because the state of affairs realized by op_1 is incompatible with that realized by op_2, or because performing op_1 violates some precondition for performing op_2, or any other source of incompatibility. The relation must be symmetric. Suppose that performing op_1 violates a precondition for performing op_2, then if op_2 is chosen, op_1 cannot be performed, since performing it would render it impossible to perform the chosen option. Thus the notion of conflict embraces both the notion of one action physically preventing another, and the notion of the choice of an action meaning that another cannot be consistently chosen.

OP is intended to describe the totality of options available to a set of agents *A*. Each agent *a Î A* will be capable of choosing some subset of *OP.* This means that agents do not have options in common: if an option selected by an agent involves the performance of some action, that is seen as a different state of affairs from that in which another agent chooses to perform that action, and there may well be different consequences in terms of the conflict relations the option enters into, and in the benefits its performance affords to various agents. We therefore define the *option set of an agent* as in Definition 2.

Definition 2 The option set, OP_a, of an agent *a* is a subset of *OP*. For a set of agents *A*, *A = {a_1, a_2,…, a_k}*, the option sets *{OP_i}* associated with each agent induce a partition of *OP* (i.e for distinct *agents a_i, a_j, OP_i Ç OP_j = Æ* and for all *x Î OP, x* is in the option set of some agent.)

We next need to provide some means of evaluating the various options, if there is to be a possibility of rational choice. We do this by introducing the notion of the *utility* of an *option for an agent.* Selection and exercise of an option may confer benefits *not only on the agent that chooses the option, but on other agents as well*. Equally it is possible that exercising an option may reduce the welfare of the agent concerned, or of other agents.

Definition 3. The utility of an option for an agent. For all *a Î A*, *op Î OP* there is a relation *utility(a,op,z),* where *z* is an integer.

We read *utility(a,op,z)* as *op has utility z for a.*

The task of an agent is to select a subset of its options, S_a. To be a legitimate selection this set must be consistent: it cannot contain any options that would conflict in the absence of other agents. This means that the selection of an agent cannot contain any option that conflicts with one of its own options. It can, however, include options that conflict with the options of *another* agent, in the hope that either that other agent will not choose to exercise the option, or that the conflict will be resolved favorably.

Definition 4 Selection of an Agent. A subset S_a of OP_a is *selectable* by an agent *a*, if *"x "y Î OP_a, Ø conflicts(x,y).*

If a conflict occurs between options in the option set of a given agent, then the agent is free to choose whichever option it wishes, although it cannot choose both. If, however, the conflicting options are in the selections of distinct agents, one agent must be dominant with respect to the particular option. We say that the dominant agent *controls* the conflict.

Definition 5. Control of a conflict. For all conflicts between a pair of options op_1 and op_2, *conflicts(op_1, op_2),* such that *op_1 Î OP_a* and *op_2 Î OP_b* for distinct agents *a* and *b*, either *controls(conflicts(op_2, op_2),a)* or *controls(conflict(op_1, op_2),b).* Note that if *controls(conflicts(op_1,op_2),a),* then also *controls(conflicts(op_2, op_1),a).* We can add that for all *op_1 Î OP_a* and *op_2 Î OP_a , controls(conflicts(op_1, op_2),a).*

In the case where two agents select options which conflict, the option selected by the agent which controls the conflict will be realized and the conflicting option will not. We call the options selected by an agent which are realized, the *realization* of the agent.

Definition 6. Realization of an agent. The realization of an agent *a Î A*, is a set R_a such that *R_a Í S_a* and no *r_1 Î R_a* is such that there exists an *r_2 Î S_b* for some agent *b Î A, b* distinct from *a,* such that *controls(conflicts(r_1, r_2),b).*

The total utility of a given agent will depend not only on the options in its own realization, but those in the realizations of its fellow agents. It is therefore also convenient to talk of the realization of an option framework, *R*, which is the union of all the individual realizations of the agents in *A*. We call this R_A. We can say that the utility of an agent U_a is the sum of the utilities for *a* of the elements of R_A. It is also convenient to talk about the utility of a group of agents. We will write this as U_B, where *B* Í *A*, and is the sum of the individual utilities U_a for all *a* Î *B*.

The final notion we need to introduce here is the notion of the evaluation of an agent of a realization. This is a function of the total realization R_A, $eval_a(R_A)$, and is intended to be some measure of how content with the overall realization the agent is. This function, $eval_a$, may be defined in a number of different ways, some of which will be explored in later section. It is, however, critical, since it is this function that the agent will try to maximize when determining its selection S_a, in so far as it is in its power to do so.

Definition 7. Task of an Agent. The task of an agent *a* in the framework is to construct the selection S_a which is expected to maximize the value of $eval_a(R_A)$.

Not everything that happens does so as the result of the action of an individual agent. Some things happen as the result of nature, and other states of affairs are the product of the actions of more than one agent. In order to represent this we distinguish a partition element OP_N, ("nature"). Since Nature is not an agent, each option (or, rather *event*, since no agent chooses it) is selected with some probability. This means that for all events in OP_N there is a given probability that they will be realized unless prevented by some other option. Similarly an event may prevent an agent from realizing an action. The probability of an event may be independent or conditionally dependent on the probabilities of other events.

Finally we should say that much of the previous work exploring inter-agent behavior, such as Axelrod (1986) and Shoham and Tennenholtz (1997) has been concerned with how behavior evolves over time. There is no reason to see the option framework as a one-off event. It is equally possible to have a sequences of frameworks, each representing one of a sequence of interactions.

3. USING THE FRAMEWORK

The framework given in section 2 is, and is intended to be, very abstract. In order to represent a particular situation we must determine a number of factors. The purpose of this section is to describe some of the factors which can be varied, and to offer suggestions as to how they might be varied.

3.1 Information Available To An Agent

An agent might have different degrees of information about *OF* and its associated relations. At one extreme it would be aware of all the options in *OP*, all the agents in *A*, how the options were partitioned between agents, the extension of the *conflicts* relation, the extension of the *controls* relation, and the extension of the utilities relation. At another extreme, the agent might be aware only of its own partition OP_a, and be unaware of any conflicts with options outside of OP_a. Obviously a number of intermediate positions with respect to all of these elements are possible: the agent might for example be aware only of agents whose actions were in conflict with its own, of the *controls* relation to the extent that it governed conflicts with options in OP_a, and of the utilities only in respect of the agents it was aware of. The amount of information available to the agent will have an effect on the most rational selection it can construct. Similarly it will evaluate a realization in terms of the utilities of only those agents of which it is aware.

Note that, in a multi-agent framework, different agents may have different amounts of information available.

3.2 Strategy of an Agent

Even if the agent has complete knowledge of the framework, for options that conflict with the options of other agents which are controlled by the other agent, the agent cannot tell whether a selected option will be realized. The agent may choose an option even if it is not in control of the conflict in the hope that the other agent will not select the conflicting option.

This means that an agent will need to choose whether or not to gamble on the selections of other agents being favorable. An adventurous agent will attempt to maximize its welfare on the assumption that its selection will be realized, whereas a cautious agent will attempt to maximize its minimum return. Alternatively agents may use game theory to inform their selection. If the framework is part of the sequence, agents may adopt a strategy in the light of past successes and failures.

The choice of strategy may also be affected by what happens once conflicts are resolved. The non-realization of some options in the original selection may mean that the options with which those options can be conflicted can still be included in the selection. If the agent is allowed this option to modify its original selection in the light of the revealed selections of others, there may be merit in being less cautious with the original selection.

3.3 Communication Between Agents

Another factor will be the degree of communication between agents. We may allow agents to communicate with some or all other agents of which they are aware both to extend their information about *OP*, and to negotiate as to which actions they will select. It may well be advantageous for agents controlling particular conflicts to announce that they will refrain from exercising their option in these conflicts, or to undertake to so refrain if the other agent refrains from exercising some of its options; or to undertake to perform a particular action if the other agents makes some similar concession. In this way the agent may make a selection with greater assurance that the selected options will be realized. It should, however, be remembered that this assurance cannot be complete: the other agent may renege on its commitments. This suggests another consideration: the degree of trust an agent places in other agents.

Of course, if there are several agents, different agents may employ different strategies, have different capabilities for communication, and have different degrees of trust and trustworthiness.

3.4 How Agents Evaluate Realizations

A very important factor in the selection of an agent is how the agent evaluates the various realizations, the nature of the $eval_a$ function. There are a number of possibilities, such as:

- the agent may consider only its own utility
- the agent may consider the utilities of other agents. The extent of its concern may be a single other agent ("partnership"); a small group of agents ("family"); a medium sized group of agents ("tribe"); an large group of agents ("nation"); or even all other agents ("agentkind").
- if the utilities of other agents are considered, it is not necessary that they be treated uniformly: different weights might be attached to the utilities of different agents, or groups of agents.
- agents have the possibility of exhibiting enmity as well as benevolence. There may be agents or groups of agents that the agent wishes to harm as well as those it wishes to help.
- $eval_a$ may not be a simple sum of utilities. The agent may attempt to equalize utilities, ensure that all agents have a certain level of utility, or use some other principle.

Again it is obvious that such differences in what the agent counts as a good realization will have a considerable influence on the option it chooses to select, and that it possible for different agents to evaluate the realizations according to different criteria.

3.5 Determining the Controls Relation

It could be the case that the decision as to which agent controls a conflict is simply arbitrary. Alternatively it may be determined in some systematic way:

- There could be a total ordering on agents, so that the more powerful agent controls all conflicts with less powerful agents.
- The ordering could be on groups of agents, with inter-group conflicts determined arbitrarily.
- The ordering could be within groups of agents, with extra-group conflicts being determined arbitrarily.
- Conflicts could be determined by some external regulation, discussed in section 3.6.

Obviously other possibilities are available, for example involving partial orderings.

3.6 External Regulation

Most human societies have laws and regulations to govern conflicts among their members. We can also use analogues of such regulation to influence the behavior of agents within our framework. We could envisage regulations to:

- prohibit agents from selecting certain options: for example those which harm other agents too much. Such restrictions might apply to all other agents, or only to agents in a certain grouping.
- construct the controls relation. For example we might say that the agent whose option had the greater overall utility should control the conflict
- to require that the selection of an agent included options which promoted the utility of some other agent or agents to a certain extent.

Many other systems of regulation would be possible. Again different groups of agent might be subject to different regulatory regimes.

4. MODELING SPECIFIC SITUATIONS

In this section we will show how the framework can be used to model some specific situations. We will give three examples. First we illustrate our framework by giving a full description of an example with two agents, each capable of three actions. We then model the famous Prisoner's Dilemma (e.g. Axelrod 1984), much used in the discussion of inter agent behavior, e.g. (Danielson 1992, Philipps 1993). Thirdly we will model the norm game used in Axelrod (1987).

4.1 A Two Agent Example

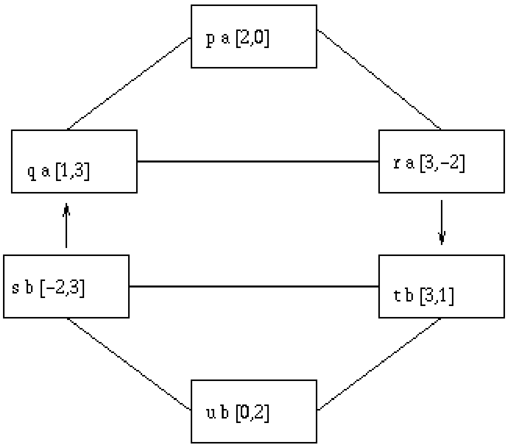

Figure 1: Example Option Framework with two agents

This example is intended simply as an example of an option framework, which I shall call EOF.

EOF = <EOP,Econflicts)

EOP = {p,q,r,s,t,u}

Econflicts ={conflicts(p,q),conflicts(q,p),conflicts(p,r),conflicts(r,p),

conflicts(q,r),conflicts(r,q),conflicts(q,s),conflicts(s,q),

conflicts(r,t),conflicts(t,r), conflicts(s,t),conflicts(t,s),

conflicts(s,u),conflicts(u,s),conflicts(t,u),c onflicts(u,t)}

A = {a,b}

OP_a = {p,q,r} OP_b = {s,t,u)

controls(conflicts(s,q),b). controls(conflicts(r,t),a).

utility(a,p,1). utility(a,q,1). utility(a,r,2).

utility(a,s,-2). utility(a,2,3). utility(a,u,0).

utility(b,p,0). utility(b,q,3). utility(b,r,-2).

utility(b,s,2). utility(b,t,1). utility(b,u,1).

The framework can conveniently be depicted as a graph as in figure 1. The control of a conflict between the options of distinct agents is indicated by a directed edge. Vertices are labeled with their name and their utilities, written as [utility for a, utility for b]

Each agent can select only one of its three available options, since its options mutually conflict. The possible selections, realizations and resulting utilities are shown in Table 1. An identifier is given to each realization for ease of later reference.

ID	S_a	S_b	R_A	U_a	U_b	U_A
R1	p	s	p,s	0	3	3
R2	p	t	p,t	5	1	6
R3	p	u	p,u	2	2	4
R4	q	s	s	-2	3	1
R5	q	t	q,t	4	4	8
R6	q	u	q,u	1	5	6
R7	r	s	r,s	1	1	2
R8	r	t	r	3	-2	1
R9	r	u	r,u	3	0	3

Table 1: Possible realizations in example 1.

As can be seen from Table 1, even a very simple framework such as this can give a variety of outcomes. This example will be discussed further in section 4.2.

4.2 The Prisoner's Dilemma

In the Prisoner's Dilemma, each agent can choose either to co-operate or defect. The utility of cooperation and defection depends on the choice of the other agent. If both co-operate, each receives two loaves, in both defect, both receive one loaf. If one defects and the other co-operates, the defector receives three loaves and the other none. The dilemma is whether to co-operate (which is collectively most beneficial, but could mean going hungry), or to defect. To model this we need to introduce our special agent, Nature.

PD = <PDOP,PDconflicts)

PDOP = {C_a, C_b, D_a, D_b, C_aC_b, C_aD_b, D_aC_b, D_aD_b }

PDconflicts = {conflicts(C_a, D_a), conflicts(C_b, D_b),

conflicts(C_a, D_aC_b), conflicts(C_a, D_aD_b),

conflicts(D_a, C_aC_b), conflicts(D_a, C_aD_b),

conflicts(C_b, C_aD_b), conflicts(C_b, D_aD_b),

conflicts(D_b, D_aC_b), conflicts(D_b, C_aC_b),}

A = {a,b,N}

OP_a = { C_a, C_b } OP_b = { D_a, D_b)

OP_N = { C_aC_b, C_aD_b, D_aC_b, D_aD_b }

The control of conflicts is determined by the fact that N cannot control a conflict.

All options in OP_a and OP_b have zero utility.

utility(a,C_aC_b,2). utility(b,C_aC_b,2).

utility(a,C_aD_b,0). utility(b,C_aD_b,3).

utility(a,D_aC_b,3). utility(b,D_aC_b,0).

utility(a,D_aD_b,1). utility(b,D_aD_b,1).

Again we can represent this as a graph, shown in Figure 2. Here no probability is assigned to nature's options, since which occurs is determined by the active agents. For clarity, the conflicts between these events have been omitted.

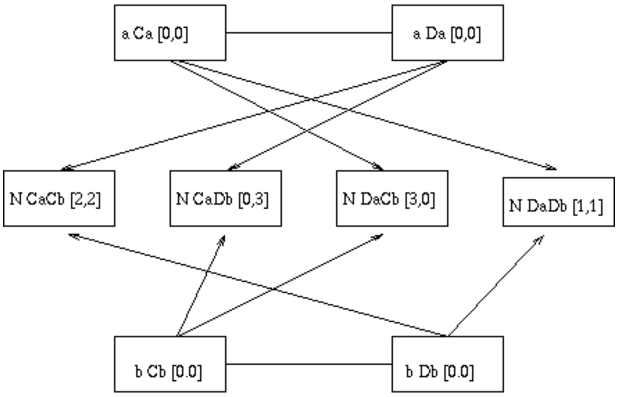

Figure 2; Prisoner's Dilemma

4.3 Axelrod's Norm Game

The third example will show the framework for the "norm game" discussed in Axelrod (1986). In this game players may conform or defect. If the person defects, if seen by another, the other may choose to punish them or not. Defection has utility 3 for the defector, and -1 for the other agents. Punishing has a utility of -9 for the punished, and a -2 enforcement cost for the punishing agent. An agent seeing a defection has a fixed probability.

For brevity we will represent this game only as a graph, shown in Figure 3.

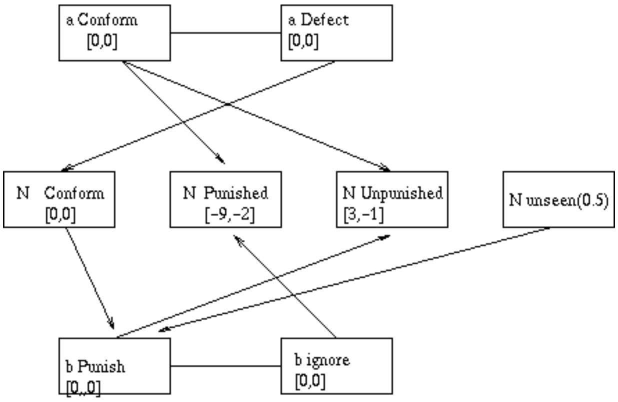

Figure 3: Norm Game

Thus *a* chooses to conform or defect. If *a* defects, it is liable to punishment. The other agent, *b*, may choose to punish or ignore. If it chooses to punish, it can be prevented either by there being nothing to punish, or by failing to see the defection, which here has a probability of 0.5.

4. EXAMPLES

In this section we will give some examples to show how the framework can be applied to some specific cases, and what can be learned from this. The examples we shall use will be based on assuming that the agents can have a range of attitudes to one another, expressed through different *eval* functions. The case is interesting because the previous work cited above (in common with most accepted economic theory) always assumes that the agent wishes to maximize its own utility, and is indifferent to the utility of the other agent. This does not seem to us a realistic assumption, and so it will be interesting to see what happens when it is relaxed. Before looking at this in two of examples given above, we will briefly discuss the case where there is only a single agent.

4.1 A Single Agent

Our first example will be of a framework containing only a single agent.

This is a simple case in that there will be no conflicts outside of the control of the agent, so that the agent's selection will always be fully realized. Moreover, here the agent can only consider its own utility, and so the evaluation can be represented as equivalent to its total utility, on the assumption that a rational independent agent will wish to maximize its utility.

Let us now make a further assumption, that all options available to the agent are of positive and equal utility to the agent. In this case the agent's task reduces to finding the maximum independent set of the graph in which options are the vertices and conflicts between options the edges. This is a well known problem in graph theory and, although it is NP-hard, there are reasonable algorithms to approximate it (e.g. Garey and Johnson 1979). Equally if we relax the assumption with respect to utilities, allowing different and negative utilities, we get a variant on this problem which will also be amenable to a graph theoretic approach.

The one agent case is, of course, particularly straightforward, and removes many of the considerations identified above. Let us therefore move to the two agent case, which re-introduces many of the interesting complications.

4.2 Two Agents

In all cases discussed below we will assume that the agents both have perfect information about the framework. We will, however, give them a range of attitudes. This is effected by giving them different *eval* functions. The five attitudes to the other agent we shall consider are, (with the evaluation of agent *a* in brackets):

- indifference(I) the agent considers only its own utility (U_a)
- benevolence(B): the agent considers its utility, and the utility of the other equally ($U_a + U_b$)
- love(L): the agent considers only the utility of the other (U_b)
- dislike(D): the agent values its own utility and the disutility of the other ($U_a - U_b$)
- hatred(H): the agent values only the disutility of the other (-U_a)

For each of these attitudes there is a realization which the agent will regard as optimal. We can also sum these subjective valuations to get a measure of the combined subjective worth and determine the best realization on this measure. The results of this are shown in Table 1. The rows are labeled with the best realization for *a*, the columns with the best realization for *b*, and the cells contain the collectively best realization. In some cases all situations will have the same collective subjective utility. For a description of the options in each realization, see Table 1.

	I (R6)	B (R5)	L (R2)	D (R4)	H (R4)
I (R2)	R5	R5	R2	R6	-
B (R5)	R5	R5	R5	R6	R6
L (R6)	R6	R5	R5	R4	R4
D (R8)	R2	R2	R8	-	R8
H (R8)	-	R2	R8	R4	R4/8

Table 2: Preferred realizations for Example 1

From our Olympian viewpoint, we might well consider R5({q,t}) to be the most desirable outcome, since it maximizes total utility, with an equal distribution. This is indeed evaluated as the best situation on the criterion of maximum collective worth in some cases, but as can be seen from Table 2, in many cases the agents themselves would prefer something else.

If we have interactions between agents displaying these attitudes, what options will they select? This will depend not only on their attitude but their strategy. If the agent does not care about its own utility, there is no problem: if *a* loves *b* it can do no better than *q*, and if *a* hates *b* it can do no better than *r*. For the other agents, however, there is a dilemma. If *a* chooses anything other than *r*, then if *b* chooses *s*, the result is very unfavorable to *a*. On the other hand, if *b* chooses other than *s*, *a* would typically do better and never do worse, by not choosing *r*. If *a* decides on a strategy of maximizing its minimum return, it will choose *r*, thus sacrificing potential benefits for a safe, if not especially desirable, outcome. The strategy is plausible, but will lead to sub-optimal results in most cases.

In such circumstances, agents would benefit greatly from some negotiation. It is the threat of the other agent performing the harmful action that removes the freedom to choose whether to act selfishly or altruistically. So if they could agree to refrain from the harmful action, they would be free to act in their better interests. This is true unless one agent hates the other, where harming the other *is* its genuine interest. Thus negotiation would either free the agent to act as it would choose, or, if negotiation was refused, allow it to apply the strategy of maximizing its minimum return without losing potential benefit. If we go beyond negotiation, and require agents to refrain from the harmful action, we force disliking and hating agents to behave as if they were indifferent.

This is a single and specific example, but it does illustrate the losses of benefit that comes from uncertainty, and how they might be reduced through communication and negotiation between agents.

5.3 Prisoner's Dilemma with Attitudes

Let us consider the effect of attitudes on the Prisoner's Dilemma. Table N shows the realizations and utilities,

	Ua	Ub	Ua + Ub	Ua - Ub
CaCb	2	2	4	0
CaDb	0	3	3	-3
DaCb	3	0	3	3
DaDb	1	1	2	0

Table 3: Payoffs in the Prisoner's Dilemma

Most work on the prisoner's Dilemma has considered only two indifferent agents, and has attempted to show how, over a series of games, the rational strategy will lead to mutual cooperation. The best strategy that has been found, e.g. Axelrod (1984), is known as "Tit for Tat": the agent offers cooperation on the first round, and thereafter plays what its opponents played on the previous round. Does our framework bear out this finding for two indifferent agents, and will agents with a different attitude adopt a different strategy?

The five attitudes rank the various outcomes as follows (own action given first):

- Indifferent: DC, CC,DD,CD
- Benevolent: CC, CD/DC DD
- Loving: CD,CC,DD,DC
- Disliking: DC, DD/CC, CD
- Hating: DC,DD,CC,DD

From this we can see that only for the indifferent agent is there a dilemma at all. The ill disposed agents have nothing to gain through cooperation, and benevolent and loving agents have nothing to gain by playing D, unless it may in some way lead to cooperation. We may therefore conclude that the ill disposed will play D on the first round and the well disposed will play C. The benevolent agent, when faced with a D in response can know that the other agent is, at best, indifferent. Will responding with a D induce cooperation? Not in the case of an ill disposed agent, but the indifferent agent prefers cooperation to mutual defection, and so may be expected to play C on the third round, once it recognizes the preparedness to tolerate mutual defection if it refuses to cooperate. If so the benevolent agent will accept the olive branch and cooperate again. Essentially the benevolent agent is playing tit for tat.

What of the indifferent agents? Suppose they begin with two rounds of defection. A loving agent will answer this with two rounds of cooperation, and so reveals that it can be defected against without penalty. Ill-disposed agents respond with two rounds of defection, and a benevolent agent with a round of C and a round of D. But the problem is that they still don't know that the agent which defects twice is ill disposed rather than indifferent. If they are to try to avoid mutual defection, they must cooperate at some point, and so they will need to play C on the third round, and then determine their remaining choices in the light of the other's third round move. (Switching to C on the second round is unwise, since a benevolent agent can be expected to defect on the second in the face of the original defection.) If, on the other hand, they begin with C, they are rational to switch to on the second round, whether or not the other defects, to test whether their defection will be punished. If it is not, they may continue to defect. If it is, they should switch back to cooperation. Thus the choice seem between playing D,D,C, and C,D,C. The returns for these two from the first three rounds for each type of agent (with the two possibilities for the indifferent agents) are shown in Table 4. By this time the agent can know whether cooperation is fruitful, and so subsequent rounds will give the same return for either initial strategy. Each cell gives four numbers; the score of the column strategy and the row strategy after three rounds, and after 10 rounds.

	D,D,C		C,D,C	
	3 x	10 x	3 x	10 x
I (D,D,C)	4-4	18-18	3-6	17-20
I (C,D,C)	6-3	20-17	5-5	19-19
B (tit for tat)	4-4	18-18	5-5	19-19
L (C,C,C)	9-0	30-0	7-4	28-4
D (D,D,D)	2-5	9-12	1-7	8-14
H (D,D,D)	2-5	9-12	1-7	6-14

Table 4: Returns in Prisoner's Dilemma for 3 and 10 rounds

From this we can see that against conditional cooperators, C,D,C is better, but in all other cases D,D,C is better. (Both are better in the long run for the indifferent agent than D,D,D). It is therefore unclear whether it is better to start by cooperating or defecting: this will depend on the distribution of attitudes in the population. If, however, minimizing *relative* loss is important, then D,D,C seems better, since it offers fewer chances to be defected from.

Only the benevolent agents follow the strategy of tit for tat, since we now have agents who will not defect, and so there can be some advantage in defecting. The essential idea of tit for tat, however, offering cooperation for a single round, and then defecting only in the face of defection, does apply, it is just that the initial rounds are non-cooperative as the agent attempts to be less vulnerable until it has information about the other agent's attitude. Notice also that the chosen strategy does not score better *than* tit for tat, and scores less well than initial cooperation *against* tit for tat. The strategy we have for our indifferent agent now resembles that of Danielson's reciprocal cooperator (Danielson (1992)), and also emerges from the discussion in Philipps (1993).

FUTURE WORK

Of course, no conclusions can be drawn from the analysis of a these small examples. Indeed concentration on specific games as in much previous work, in our opinion, has tended to make the conclusions drawn resistant to generalization. See, for example, the differences made by an adjustment in the relationships between the payoffs in the Prisoner's Dilemma in Philipps (1993). But the purpose of the previous section is not to draw conclusions: rather it is an attempt to show that the framework offered in this paper can provide a setting for the analysis of a wide range of situations in which the activities of agents interact and conflict. It is to this wider, more flexible range of situations that we must look if we want to draw conclusions that will not be vitiated by specific features of particular situations. We can see that we can use the framework to explore the effects of factors such as:

- differences in attitude of agents to one another
- differences in strategy for the selection of options by agents
- differences in negotiation strategies
- differences in external regulation of agents.

The intention is to use the framework to perform a systematic empirical study of how these factors affect inter-agent behavior. Using a large number of automatically generated option frameworks, and implemented agents to make the selections in accordance with different sets of principles we can evaluate such questions, as

- Under what circumstances do agents attain the realization that the agents collectively regard as subjectively preferred?
- Does forcing agents to select so as to avoid preventing of selections with higher utility always lead to a greater total utility?

Where a hypothesis does not hold universally, we can provide a characterization of the frameworks in which it does hold. For example in a framework in which an agent could perform no action beneficial to the other agent, different properties might hold.

We can then go on to use the framework to explore situations with many agents. As well as exploring whether the results from the two agent situation generalize, we can also explore hypotheses concerning groups of agents. For example in a framework comprising pairs of agents which are benevolent to one another and indifferent to all other agents, is the most effective partnership between agents with highly connected partitions? What effect would different amounts of information available to pairs of agents have? Or we could go beyond pairs and investigates groups to explore whether differently sized groups performed better, and whether groups with homogenous attitudes perform better than groups with different attitudes. These are just some of the issues: many more suggest themselves.

As well as looking at such general frameworks, we can also attempt to relate frameworks to particular models of social organization.

Hand in hand with this empirical work, should go analytic work. If a property is strongly suggested by the empirical work, it would be worth investigating whether it was actually provable. Similarly analytic work will generate hypotheses which can be explored empirically.

6. CONCLUDING REMARKS

The intended contribution of this paper is to identify a critical feature of inter-agent behavior: that selecting an option will preclude certain other actions both by the agent that selects it and other agents, and to provide a sufficiently abstract framework in which this problem can be explored. Following from this we have identified a number of factors which can be considered when exploring this problem, and illustrated their effects by reference to a limited example.

Investigation of these issues could be directed in a number of ways.

- modeling different social organizations
- consideration of how norms might emerge in agent societies
- evaluation of different regulatory regimes
- an notion of ethical behavior for agents

There are many and various possibilities: all this paper attempts to do is lay down a framework in which they can be investigated.

REFERENCES

Axelrod, R., (1984). The Evolution of Cooperation. Basic Books.: New York.

Axelrod, R., (1986). An Evolutionary Approach to Norms. *American Political Science Review*, Vol 80(4) pp. 1095-111.

Danielson P., (1992).Artificial Morality. Routledge: London.

Garey, M. R. and Johnson, D. S., (1979). *Computers and Intractability, a Guide to the Theory of NP-Completeness*. Freeman and Co.

Philipps, L., (1993).Artificial Morality and Artificial Law. *Artificial Intelligence and Law*, Vol 2 No 1. pp. 51-64.

Shoham, Y., and Tennenholtz, M., (1997).On the Emergence of Social Conventions, *Artificial Intelligence*, Vol 94 (1-2) pp. 139-166.