SUSANNE KALENKA, NICHOLAS R. JENNINGS

SOCIALLY RESPONSIBLE DECISION MAKING BY AUTONOMOUS AGENTS

Abstract. Most autonomous agents are situated in a social context and need to interact with other agents (both human and artificial) to complete their problem solving objectives. Such agents are usually capable of performing a wide range of actions and engaging in a variety of social interactions. Faced with this variety of options, an agent must decide what to do. There are many potential decision making functions which could be employed to make the choice. Each such function will have a different effect on the success of the individual agent and of the overall system in which it is situated. Therefore, this paper examines agents' decision making functions to ascertain their likely properties and attributes. A framework for characterising social decision making is presented and a socially responsible decision making principle is proposed which enables both the agent and the overall system to perform well. This principle is illustrated, and empirically evaluated, in a multi-agent system for unloading lorries at a warehouse.

1. INTRODUCTION

Intelligent agents are a new paradigm for developing software applications. More than this, agent-based computing has been hailed as "a new revolution in software" (Ovum 1994) and it has been predicted that "agents will be pervasive in every market by the year 2000" (Janca 1995). Consequently, agents are the focus of intense interest on the part of many sub-fields of computer science and artificial intelligence. Agents are being used in an increasingly wide variety of applications, ranging from comparatively small systems such as email filterers, to large, complex mission critical applications such as air traffic control. Despite this apparent diversity, in all cases the key abstraction used is that of an *agent*. Although there is much debate about exactly what an agent is (see Franklin and Graesser 1996 for a discussion), we consider the following to be necessary conditions for agenthood: (i) *autonomy*—can act without the direct intervention of others and has control over its own actions and internal state; (ii) *responsiveness*—can react in a timely fashion to environmental changes; (iii) *proactiveness*—can take the initiative where appropriate; and (iv) *social ability*—can interact to complete its problem solving and to assist others (Wooldridge and Jennings 1995).

The increased autonomy afforded by the agent paradigm means that an agent's decision making function is central to the success of any application. This function takes a subset of the agent's beliefs—including, for example, its current state, the

K. Korta et al. (eds.), Cognition, Agency and Rationality, 135-149.

state of the environment, and the state of other agents—and determines the course of action the agent should follow. Almost by definition, this function operates with a partial view of the world (Lesser and Corkill 1987) and, because of the inherent interdependencies between the agents (Davis and Smith 1983), the choices it makes affect not only itself but other agents in the environment. Given this situation, it is important to consider the following fundamental questions: what is a good decision making function? and what measure should be used to rate goodness? Present work in multi-agent systems can be divided into two broad camps (figure 1). The reductionist view (e.g. Erman and Lesser 1975, Lesser and Corkill 1983) is concerned with building effective overall systems using the notion of interacting agents, while the constructionist view (e.g. Ferber and Drogul 1992, Fischer *et al.* 1996, Overgaard *et al.* 1996, Steels 1989, Wavish and Graham 1996) considers interacting agents as a given and is concerned with determining what sort of overall system emerges from them.

Figure 1. Extant approaches to social decision making

	Reductionist View	**Constructionist View**
Motivation	Devise a system of interacting agents that work together to solve a common problem.	Devise agents which interact to further their own needs.
Measure of Goodness	Overall system performance.	Performance of individual agents.
Agents' Decision Making Function	Benevolence—accept all requests made.	Individual utility maximisation.
System Coherence	Carefully engineered by single design team.	Emerges out of interplay between agents.
Main Drawback	Fails to fully exploit concept of autonomous agents—too much system level design.	System behaviour defined through human-intensive refinement of individual agents.

Whilst both of these approaches have enabled useful applications to be developed, it has been predicted that agents will have the greatest impact in complex industrial and commercial applications such as process control, telecommunications management, business process management, air traffic control, manufacturing, and information management (Ovum 1994). In such systems, what is required is the ability to exploit the conceptual power of autonomous agents (as in the constructionist view), but to ensure the overall system performs in a coherent manner (as in the reductionist view). Given these contradictory demands, we believe the best means of building socially coherent multi-agent systems is to endow the individual autonomous agents with greater social awareness. This awareness enables the agents to explicitly consider the effects of their actions on the wider community. Given this information, a decision function can then be designed which enables agents to exploit interactions with others for their own gain, but which means that they are sometimes willing to do things for the greater good (to improve system coherence). Here the term *social responsibility* is used to denote such agents.

state of the environment, and the state of other agents—and determines the course of action the agent should follow. Almost by definition, this function operates with a partial view of the world (Lesser and Corkill 1987) and, because of the inherent interdependencies between the agents (Davis and Smith 1983), the choices it makes affect not only itself but other agents in the environment. Given this situation, it is important to consider the following fundamental questions: what is a good decision making function? and what measure should be used to rate goodness? Present work in multi-agent systems can be divided into two broad camps (figure 1). The reductionist view (e.g. Erman and Lesser 1975, Lesser and Corkill 1983) is concerned with building effective overall systems using the notion of interacting agents, while the constructionist view (e.g. Ferber and Drogul 1992, Fischer *et al.* 1996, Overgaard *et al.* 1996, Steels 1989, Wavish and Graham 1996) considers interacting agents as a given and is concerned with determining what sort of overall system emerges from them.

Figure 1. Extant approaches to social decision making

	Reductionist View	**Constructionist View**
Motivation	Devise a system of interacting agents that work together to solve a common problem.	Devise agents which interact to further their own needs.
Measure of Goodness	Overall system performance.	Performance of individual agents.
Agents' Decision Making Function	Benevolence—accept all requests made.	Individual utility maximisation.
System Coherence	Carefully engineered by single design team.	Emerges out of interplay between agents.
Main Drawback	Fails to fully exploit concept of autonomous agents—too much system level design.	System behaviour defined through human-intensive refinement of individual agents.

Whilst both of these approaches have enabled useful applications to be developed, it has been predicted that agents will have the greatest impact in complex industrial and commercial applications such as process control, telecommunications management, business process management, air traffic control, manufacturing, and information management (Ovum 1994). In such systems, what is required is the ability to exploit the conceptual power of autonomous agents (as in the constructionist view), but to ensure the overall system performs in a coherent manner (as in the reductionist view). Given these contradictory demands, we believe the best means of building socially coherent multi-agent systems is to endow the individual autonomous agents with greater social awareness. This awareness enables the agents to explicitly consider the effects of their actions on the wider community. Given this information, a decision function can then be designed which enables agents to exploit interactions with others for their own gain, but which means that they are sometimes willing to do things for the greater good (to improve system coherence). Here the term *social responsibility* is used to denote such agents.

The aim of this paper is, therefore, to elucidate the decision making principles of socially responsible agents. It is hypothesised that such agents are the best means of designing multi-agent systems in which a balance needs to be struck between the needs of the individual agents and the needs of the overall system (henceforth the society). To this end, section 2 presents an informal framework for characterising social decision making. Section 3 uses this framework to identify a principle of socially responsible decision making and identifies three specific socially responsible decision making functions (socially self-interested, helpful and cooperative). Section 4 illustrates the use of these functions in a multi-agent system for unloading lorries at a warehouse and presents an empirical evaluation of their effectiveness. Section 5 discusses related work and, finally, section 6 outlines some issues which require further investigation.

2. A FRAMEWORK FOR CHARACTERISING SOCIAL DECISION MAKING

Most work on decision making functions for autonomous agents concentrates on making individually rational choices (e.g. Doyle 1992, Russell and Wefald 1991, Wellman 1993). This work either uses a decision theoretic notion of utility maximisation (given the probability of being in a state after executing an action and a rating of the desirability of that state, choose the action which maximises the expected utility (Doyle 1992) or a more intuitive description such as Newell's Principle of Rationality (if an agent has knowledge that one of its actions will lead to one of its goals, then it will select that action (Newell 1982). In either case, however, the decision function is solipsistic—there is no consideration of the impact of actions on other members of the society and no notion of doing anything other than maximising the agent's own gain. It would be possible to manipulate the agent's utilities (or goals) so that they incorporate a measure of social awareness, but this would simply be hiding the underlying principles behind the numbers.

In order to be explicit about the underlying principles of social decision making, we note that both the agent who performs an action and the society in which that agent is situated can be affected by the execution. This affect can be beneficial (a positive utility value), detrimental (a negative utility value) or indifferent (a utility value of zero). Also individual and society benefit are orthogonal measures. In more detail:

> **Individual Benefit (a, S, α)**: the benefit[1] agent a, situated in society S, obtains for performing action α is a combination of the benefit attributed solely to the action executor (*agent sole benefit*) and the executor's share of the benefit the society obtains when one of its members executes a (*agent share benefit*). For example, if a team of agents is searching for a particular document on the world wide web, then the system may be organised such that the individual who actually finds the document receives 50% of the customer's payment (agent sole benefit) and the remaining 50% is split evenly between the team (in which case the finder also receives some portion of this 50% as its agent share benefit).

The aim of this paper is, therefore, to elucidate the decision making principles of socially responsible agents. It is hypothesised that such agents are the best means of designing multi-agent systems in which a balance needs to be struck between the needs of the individual agents and the needs of the overall system (henceforth the society). To this end, section 2 presents an informal framework for characterising social decision making. Section 3 uses this framework to identify a principle of socially responsible decision making and identifies three specific socially responsible decision making functions (socially self-interested, helpful and cooperative). Section 4 illustrates the use of these functions in a multi-agent system for unloading lorries at a warehouse and presents an empirical evaluation of their effectiveness. Section 5 discusses related work and, finally, section 6 outlines some issues which require further investigation.

2. A FRAMEWORK FOR CHARACTERISING SOCIAL DECISION MAKING

Most work on decision making functions for autonomous agents concentrates on making individually rational choices (e.g. Doyle 1992, Russell and Wefald 1991, Wellman 1993). This work either uses a decision theoretic notion of utility maximisation (given the probability of being in a state after executing an action and a rating of the desirability of that state, choose the action which maximises the expected utility (Doyle 1992) or a more intuitive description such as Newell's Principle of Rationality (if an agent has knowledge that one of its actions will lead to one of its goals, then it will select that action (Newell 1982). In either case, however, the decision function is solipsistic—there is no consideration of the impact of actions on other members of the society and no notion of doing anything other than maximising the agent's own gain. It would be possible to manipulate the agent's utilities (or goals) so that they incorporate a measure of social awareness, but this would simply be hiding the underlying principles behind the numbers.

In order to be explicit about the underlying principles of social decision making, we note that both the agent who performs an action and the society in which that agent is situated can be affected by the execution. This affect can be beneficial (a positive utility value), detrimental (a negative utility value) or indifferent (a utility value of zero). Also individual and society benefit are orthogonal measures. In more detail:

> **Individual Benefit (a, S, α)**: the benefit[1] agent a, situated in society S, obtains for performing action α is a combination of the benefit attributed solely to the action executor (*agent sole benefit*) and the executor's share of the benefit the society obtains when one of its members executes a (*agent share benefit*). For example, if a team of agents is searching for a particular document on the world wide web, then the system may be organised such that the individual who actually finds the document receives 50% of the customer's payment (agent sole benefit) and the remaining 50% is split evenly between the team (in which case the finder also receives some portion of this 50% as its agent share benefit).

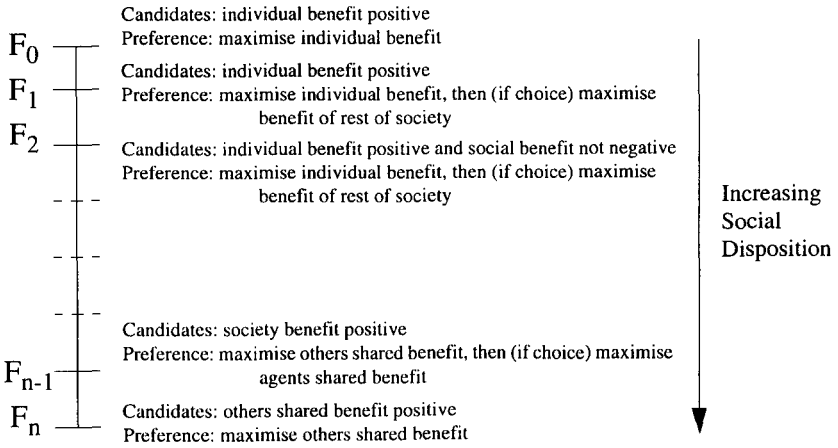
> **Society Benefit (a, S, α)**: the benefit society S obtains when one of its members (agent a) executes action α. This is composed of the executor's agent share benefit (as above) and the benefit accrued by those members of S who do not actually execute the action (*other share benefit*). For example, if there are four agents searching for a document in the above scenario and the benefit is distributed evenly amongst the group, then the agent which finds the document receives 25% of the team fund and the other three agents receive 75%.

In seeking to design socially responsible agents, neither individual, nor society benefit alone is adequate (the former may preclude doing actions for the good of the society and the latter may preclude the agent from furthering its own needs). Therefore a hybrid measure of benefit, termed *joint benefit* (Jennings and Campos 1997), is adopted. Joint benefit incorporates both the individual and the societal perspectives:

> **Joint Benefit (a, S, α)**: a combination of the individual benefit agent a obtains for executing action α and the benefit obtained by society S in which agent a is situated. Clearly the agent share benefit should not be counted twice, so joint benefit is a combination of the agent's sole benefit, the agent's share benefit, and the others' share benefit.

With this framework in place, it is now possible to return to the decision making function. All decision making functions perform two primary roles. Firstly, they identify the set of candidate actions that the agent may consider performing. For example, a self-oriented agent will only consider actions which have a positive individual benefit, an altruistic agent will only consider actions which have a positive value for other share benefit, and so on. Secondly, the function defines a preference ordering over the candidate actions. For example, a selfish self-oriented agent may rank actions in decreasing order of individual benefit, whereas a more socially aware self-oriented agent may rank them in decreasing order of individual benefit and where there is a choice in terms of decreasing society benefit.

Figure 2: Spectrum of Decision Making Functions



From this structure, a whole spectrum of potential decision making functions can be observed (figure 2)—ranging from the purely selfish (F_0) to the purely altruistic (F_n). Our concern, therefore, is with identifying the subset of this function space which leads to socially responsible behaviour. This activity is performed in section 3.

3. A SOCIALLY RESPONSIBLE DECISION MAKING PRINCIPLE

Rather than seeking to posit a single socially responsible decision making function, we feel it is more useful to identify the underlying principle(s) on which such functions are founded. This principle can then be instantiated in a variety of ways depending on the relative importance of the individual and the society in a given application. Thus our approach is in line with Newell's view of rationality (identifies a broad candidate set and no preference ordering), rather than the decision theoretic view of rationality (fixed, single-point candidate and preference structure). In seeking to strike a balance between the needs of the individual agents and those of the society, we use joint benefit (section 2) as our foundational basis. Thus, the guiding principle for socially responsible agents can be defined in the following way (Jennings and Campos 1997):

> **Principle of Social Rationality:** If a member of a responsible society can perform an action whose joint benefit is positive, then it may select that action.

In this work, we concentrate on three socially responsible decision making functions—namely, socially self-interested, helpful and cooperative (Kalenka and Jennings 1995). These functions represent a reasonable spread of the more individually-centred socially responsible alternatives and therefore maintain many of the individualistic aspects of traditional autonomous agents.

- *Socially self-interested:* represents the most individualistic type of responsible agent. Considers candidate actions which have a positive joint benefit (as per the Principle of Social Rationality) and which have a positive individual benefit. These agents primarily concentrate on their own actions, but ensure they are not overly detrimental to the society. They differ from purely self-interested functions in that they do not simply execute actions without any regard for their affects on others.

- *Helpful:* permits non-detrimental actions which are of no direct benefit to the agent, but which are beneficial for the society as a whole. These agents consider candidate actions which have a positive joint benefit (as above), but also consider actions which have no positive individual benefit. The latter stipulation enables agents to choose actions which benefit the society but which are not individually beneficial.

Both of the aforementioned decision making functions consider only isolated actions (i.e. without regard to other potentially related actions). However in many cases it is important to consider combinations of actions executed by groups of agents

(Ketchpel 1993, Shehory and Kraus 1995). Here we deal only with the simplest case in which there are two agents each executing a single (related) action. Consideration of such cooperative actions means a further extension to the agents' decision making functions—now they need to include both the action they are to execute and the action the other agent is to execute.

> • *Cooperative*: considers candidate actions which have a positive joint benefit and pairs of actions (one executed by the local agent and one by the other team member) which, when taken together, have a positive individual benefit. Thus, for example, one of the actions may have an individual benefit which is negative, but this loss is compensated for by the execution of the other agent's action. This means an agent can execute a personally detrimental action for the good of the society (joint benefit positive) as long as it receives sufficient individual benefit from the accompanying team action.

It is clear that the cooperative decision function requires some form of social commitment (Castelfranchi 1995, Jennings 1993) to be made between the two agents. Without such a commitment, agents will not consider individually detrimental actions since they cannot be guaranteed to recoup their loss by the execution of the subsequent action. The nature of the social commitment can vary in scope (from being valid for just the current pair of actions, up to an ongoing partnership to help one another out whenever necessary (Kalenka and Jennings 1995)) and the convention under which it can be terminated (from one agent opting out because it has a better offer, to one requiring mutual acceptance of the termination (Jennings 1993)). In this work, an ongoing commitment structure is adopted because it provides a safe, long-term basis for entering into cooperative problem solving. It works by the agents committing to support one another for periodically repeating time intervals. These intervals are called *support-duty* when one agent has to support the other and *support-right* when one agent can demand support from the other. Each agent alternates between support-right and support-duty intervals. During an agent's support-duty interval it supports the other agent if it is requested to do so. When an agent's support-right interval arises it can demand up to an equivalent amount of support from the other agent as gave in its last support-duty interval. Thus, the lower the amount of support requested, the lower the support that must be given. The social commitment terminates when one of the agents does not request any support in its support-right interval. This conceptualisation means there is no predetermined end point for a commitment and in some cases it may continue indefinitely. At each support-right interval a cooperative agent has to determine how much it should force the other agent to support it as this directly affects the time it may be unable to work on its own goals at the next support-duty interval.

4. A MULTI-AGENT SYSTEM FOR UNLOADING LORRIES IN A WAREHOUSE

This section describes how the socially self-interested, helpful and cooperative decision making functions can be used in a practical application context—in this

case for designing a multi-agent system for unloading lorries at a warehouse (section 4.1). The effectiveness of the decision making functions in this domain are then evaluated in a series of experiments (section 4.2).

4.1 Applying the socially responsible decision making functions

Lorries arrive randomly at a warehouse laden with goods which require unloading. The warehouse has a fixed number of unloading bays which each hold one lorry at a time. Upon arrival, lorries go to the nearest free bay. Lorries have an associated time by which they would like to be discharged (their desired time). Additionally, the warehouse tries to ensure that all lorries are processed by some maximum time (T_{max}). Since the aim of this work is to illustrate social problem solving behaviour, rather than to develop a real world solution, we make a number of simplifying assumptions: all lorries arrive with the same load; the time it takes one fork lift truck (agent) to unload a lorry is T_{max}; unloading time is directly proportional to the number of agents servicing a lorry (two agents will do it twice as fast as one); and there are as many fork lift truck agents unloading the lorries as there are bays in the warehouse.

In our system, each agent is responsible for dealing with a particular unloading bay. Each agent receives some sole benefit for ensuring the lorry at its assigned bay is processed by T_{max}. Moreover, if the lorry is processed in a desired time which is less than T_{max} the agent responsible for that bay receives further sole benefit. The society as a whole receives benefit proportional to the percentage of lorries that are processed by T_{max} and the percentage processed by their desired time (where this is less than T_{max}). This benefit is split between the agents which assist others with their unloading tasks and is in proportion to the amount of assistance provided.

In more detail, let L_i be the lorry to be discharged, A_i be the agent responsible for L_i, and T_i be L_i's discharge deadline. Four distinct cases need to be considered:

- $T_i > T_{max}$: A_i can discharge L_i on its own and it has some spare time ($T_{max} - T_i$) in which it could support other agents. (Lorry is *not-time-dependent*).
- $T_i = T_{max}$: A_i can discharge L_i on its own. (Lorry is *time-dependent-alone*).
- $T_i < T_{max}$: A_i needs support from other agents if it is to meet L_i's desired deadline. Without any support discharge will take T_{max}. (Lorry requires *time-dependent-support*).
- T_i unknown: No L_i has arrived at bay i, hence A_i is free to support other agents at least until T_i is defined.

If all agents used the socially self-interested decision function, then the multi-agent system could be guaranteed to meet its objectives as long as no lorry needed to be discharged in a time less than T_{max} (i.e. there are no time-dependent-support lorries). However, this situation can be improved by the agents adopting the helpful decision making function. In this case, agents could support one another when they have a not-time-dependent lorry or when they have no lorry at their bay. Helpfulness would

ensure the agents are more heavily utilised and that more deadlines less than T_{max} are met (simply because there is more agent problem solving power available in the system). The situation can be enhanced still further by the addition of the cooperative decision making function. For example, consider the case where two lorries arrive simultaneously. Lorry L_1 has the desired time of T_{max} (it is time-dependent-alone) and lorry L_2 the desired time of T_{max} / 2 (it requires time-dependent-support). If A_1 and A_2 cooperate, as defined in section 3, A_1 could assist A_2 for the first T_{max} / 2 units of time (meaning L_2 is discharged in time) so long as A_2 agreed to assist it for the second T_{max} / 2 units of time (meaning L_1 is also discharged in time). With only responsible or helpful agents, both lorries would be discharged in time T_{max} meaning L_2's desired departure time is not satisfied. In the cooperative case, the two agents make a social commitment to support one another while either of them are in danger of not fulfilling their responsible tasks. Thus if a lorry arrives at A_2's bay while it is assisting A_1 with L_1, then A_1 must commit itself to help A_2 clear the new lorry, and so on. In the worst case, new lorries, with deadlines less than or equal to T_{max} plus the time the responsible agent is unavailable to start work, may arrive continuously at A_1's and A_2's bays while they are supporting one another. Thus their ongoing commitment will continue until: (i) no lorries arrive before the agents fulfil their duties at the other's bay; (ii) one of the lorries has a sufficiently long discharge time; or (iii) either agent receives the necessary support from some other helpful agent.

4.2 Experimental results

These experiments provide an empirical assessment of the performance of the three different types of socially responsible decision making function in the warehouse unloading application. In the particular scenario considered here, there are ten agents and ten unloading bays and a 50% chance that a new lorry will arrive at an empty bay on a particular simulation cycle. New lorries arrive in the ratio of 1 not-time-dependent lorry to 1 time-dependent-alone lorry to 3 time-dependent-support lorries. In a given experiment, all the agents have the same decision making function —thus they are all socially self-interested, or all helpful, or all cooperative.

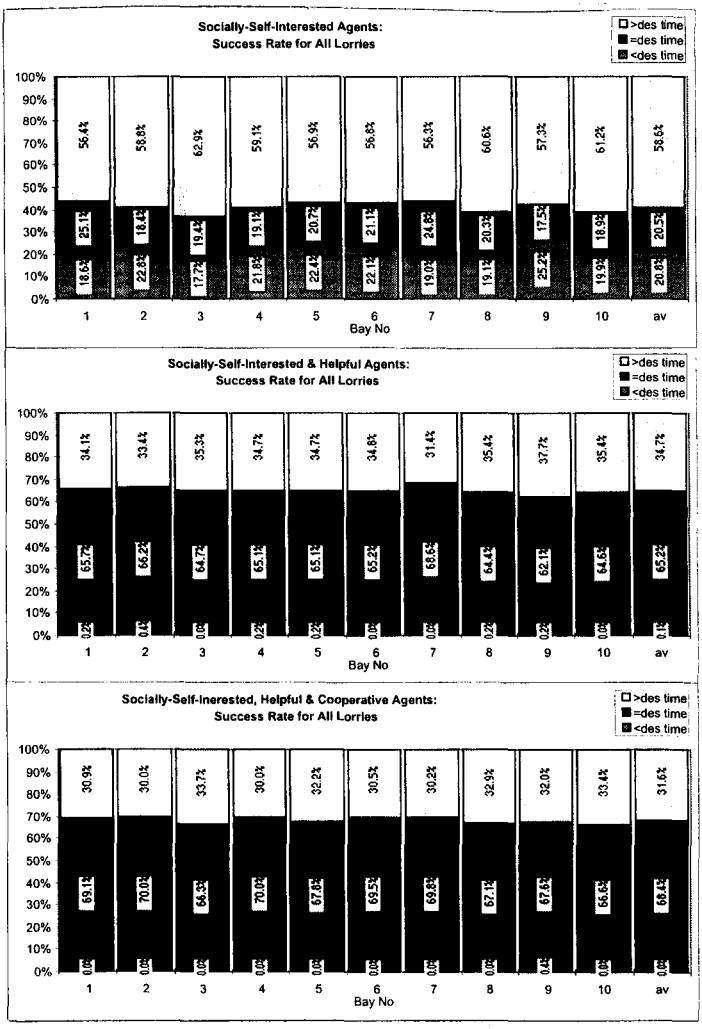
The basic socially responsible hypotheses being evaluated in these experiments can be stated in the following manner:

- as the agent's decision making function is made more socially aware (socially self-interested to helpful to cooperative), the performance of the society will improve.
- as the agent's decision making function is made more socially aware, the performance of the individual agents will not deteriorate significantly.

The first experiments consider the success rate of the different types of agents over the three types of lorry (subsequent experiments examine effectiveness by lorry type). Figure 3 (and all subsequent graphs) shows the success rate at each of the 10 bays separately and the average over all the bays. The average value can be considered as the performance of the overall society. Each bar shows the percentage of lorries at a particular bay which have been discharged in less, equal and greater than the desired

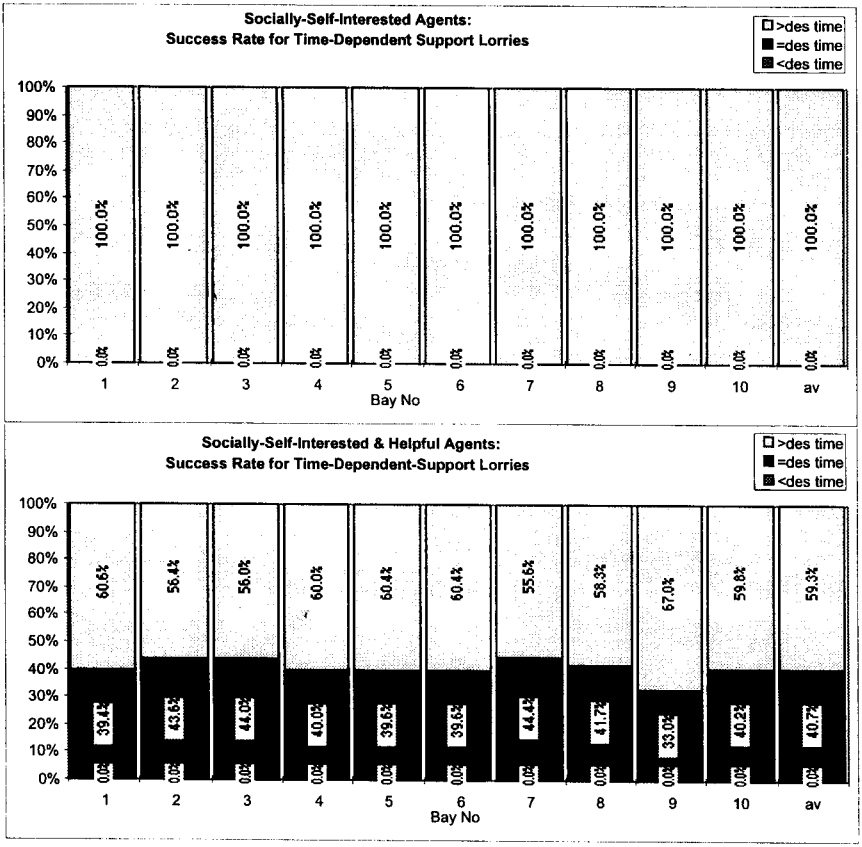
discharged time. Figure 3 demonstrates that the helpful decision function generates a significant improvement over the socially self interested function —the percentage of lorries which miss their desired time drops from 58.6% to 34.7%. This improvement is achieved by making better use of spare capacity in the system and ensuring that lorries are not needlessly processed before their desired time (for which no credit is given). The cooperative decision making function gives a still greater improvement (3.1%) over the helpful decision function although this was not as much as had been expected. The reason for the smaller than expected improvement is because the social commitment structure used in this scenario binds pairs of cooperating agents together for prolonged periods of time. This means there are comparatively few new opportunities to offer assistance.

Figure 3: Success Rate over all lorry types.



To provide a more detailed breakdown of these figures, an analysis of the success of the different functions for the different types of lorry was undertaken. For both the socially self-interested and the helpful cases, all of the lorries that were either not-time-dependent or which were time-dependent-alone were processed by their desired time (graphs not shown). The big improvement, as might be expected, occurs with respect to those lorries who require time-dependent-support. With the socially self-interested decision function no lorry meets its desired time, but with the helpful function 40.7% of the lorries meet their desired time (figure 4).

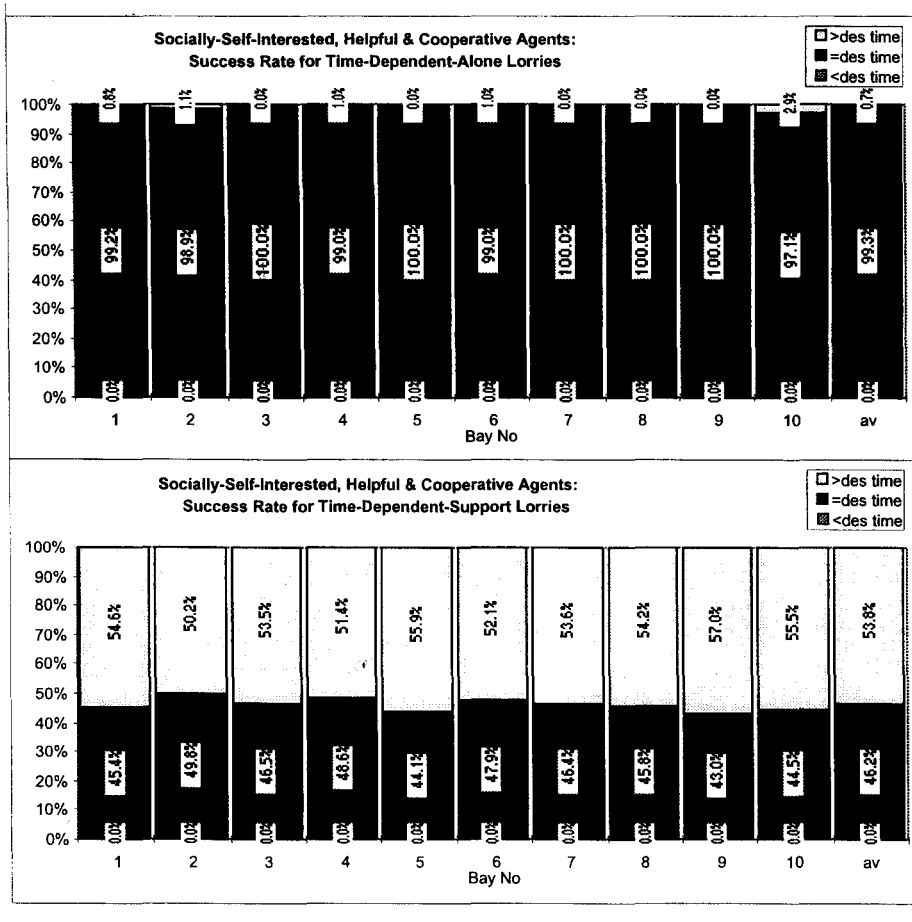
Figure 4: Success of Socially Self-Interested and Helpful Decision Functions for Time-Dependent-Support-Lorries



The cooperative decision making function reveals a further 5.5% improvement for time-dependent-support lorries over its helpful counterpart (figure 5, second graph). Thus indicating that the greatest benefit of more socially aware functions is likely to occur in cases in which there is a greater social dependence between the participating agents. However, it is interesting to note that not all of the time-dependent-alone lorries are processed by their desired time (figure 5, first graph) (cf. the socially self-interested and helpful decision functions). All of these lorries could have been

processed in time had the agent responsible for their bay taken a more self-oriented view. It is precisely because such agents enter into social commitments with one another that in some cases their individual performance suffers—they are doing actions for the greater good of the society even if they turn out to be individually detrimental some of the time.

Figure 5: Success of the Cooperative Decision Function for Time-Dependent-Alone and Time-Dependent-Support Lorries



5. RELATED WORK

A range of disciplines are concerned with the theoretical and experimental study of social interactions between autonomous agents. However a comprehensive analysis of this work is beyond the scope of this paper. Therefore, we concentrate on that research which is closest to our work as it is described in this paper.

Castelfranchi (1990) considers the motivation for social action in multi-agent systems from the perspective of social science. He identifies the notions of

individual and social power as important influences on the nature and type of interactions which occur. One agent has the power to influence the social problem solving behaviour of another when the latter is dependent on it. In this context, dependence means that an agent cannot complete one of its goals without assistance. With this view, agents assist one another (undertake helpful or cooperative actions) because of their interdependence, not because they wish to achieve greater levels of system coherence. This differing emphasis means that agents need to reason about dependency networks (Sichman *et.al.* 1994) rather than about societal utilities and also that agents are still essentially individual utility maximisers.

Cesta *et al.* (1996) explore the social problem solving behaviour of groups of agents with various interaction attitudes related to their degree of self-sufficiency and the degree to which they are willing to give help. In particular, they study the performance of social agents (which give and take support) in the presence of agents which try to exploit them. Their results show that, depending on the threshold at which agents are willing to be helpful, social agents can tolerate exploiter agents without a severe decrease in system performance. This is encouraging for our socially responsible societies as it indicates that they cannot be readily exploited by outside agents which may have different decision making principles.

Sen (1996) considers interactions between different types of self-interested agents. In particular, he considers the effects, on both individual and system performance, of including agents which receive help from others and which do, and do not, reciprocate. He concludes that reciprocal behaviour can improve the performance of the individuals and of the overall system in which they are situated. Moreover, agents which reciprocate helpful behaviour can both approach optimal global behaviour and resist exploitation by selfish agents. Again these results indicate the advantages of taking a socially responsible stance.

Finally, Marsh (1992) discusses trust as a computational concept which autonomous agents can use when deciding with whom to interact (in our context this equates to deciding who to help or who to cooperate with). Factoring such a concept into our social decision making would allow agents to use functions with an increased social disposition (towards F_n in figure 2) with agents for which they have a high degree of trust and more self-oriented functions (towards F_0 in figure 2) with those agents for which they have a low trust value.

6. CONCLUSIONS AND FUTURE WORK

This paper presented a framework for characterising social decision making in multi-agent systems. It provided a typology of the types of benefit associated with action execution and identified social commitment as a key supporting structure. The Principle of Social Rationality was put forward as a foundational basis for designing autonomous agents which strike a balance between their own problem solving objectives and the needs of the society. Three socially responsible decision functions which adhere to this principle were described and their success evaluated in a multi-agent system for unloading lorries in a warehouse. These experiments highlight the value, in terms of overall system performance, of making agents' decision making functions more socially aware. However they also indicate that, in some cases, individual performance may suffer when trying to enhance system coherence.

There are a number of aspects of this work which require further investigation. Firstly, it is assumed that all the agents are trustworthy and able to make accurate predictions about their level of commitment to one another. But what will happen to individual and system performance if some of the agents renege upon their commitments? (either deceitfully to try and exploit others or because they were simply unable to make accurate predictions about their resource availability). What sorts of mechanisms can agents put in place to protect themselves against such free-riding agents? Secondly, the present social context is represented at a coarse level of granularity—involving just the individual agent and the whole society. However in many cases, there are several social groupings within the society with various strengths of relationships between their members. Given this situation, how can a more differentiated framework for social decision making be defined? Finally, determining utilities for social actions consumes the agent's resources. In most cases these resources are limited and so it may not always be possible to determine all their values. Thus, the agent may need to take a resource-bounded view on social rationality (Hogg and Jennings 1997)—meaning the agent's decision making function is only an approximation to the ideal social rationality outlined here. In such cases, agents need to be designed which are able to manage their computations such that the more resources they have at their disposal the closer they perform to the ideal.

*Susanne Kalenka**
Nicholas R. Jennings
Queen Mary and Westfield College. University of London.
United Kingdom

NOTES

* This work has been supported by an EPSRC studentship and by a Drapers Company scholarship.

[1] In this context, the term "benefit" covers positive, negative and indifferent utility values.

7. REFERENCES

Castelfranchi, C., 1990: Social Power: A Point Missed in Multi-Agent DAI and HCI, in Y. Demazeau and J. P. Müller (eds.), *Decentralized AI* , Elsevier, pp. 49-62.

Castelfranchi, C., 1995: Commitments: From Individual Intentions to Groups and Organisations, Proc 1st Int. Conf. on Multi-Agent Systems, San Francisco, USA, 41-48.

Cesta, A., Miceli, M., and Rizzo, P., 1996: Help Under Risky Conditions: Robustness of the Social Attitude and System Performance, Proc 2nd Int Conf on Multi-Agent Systems, Kyoto, Japan, 18-25.

Davis, R., and Smith, R. G. 1983: Negotiation as a Metaphor for Distributed Problem Solving, *Artificial Intelligence **20***, 63-109.

Doyle, J., 1992: Rationality and its Roles in Reasoning, *Computational Intelligence **8*** (2), 376-409.

Erman, L. D., and Lesser, V. R., 1975: A multi-level organisation for problem solving using many diverse cooperating sources of knowledge, Proc. Int. Joint Conf. on AI, Stanford, CA., 483-490.

Ferber, J., and Drogul, A., 1992: Using Reactive Multi-Agent Systems in Simulation and Problem Solving in Distributed Artificial Intelligence: Theory and Praxis, in N. M. Avouris and L. Gasser (eds.), Kluwer Academic Publishers, pp. 53-80.

Fischer, K., Müller, J. P., and Pischel, M., 1996: Cooperative transportation scheduling: an application domain for DAI, *Int. Journal of Applied Artificial Intelligence* ***10*** (1), 1-33.

Franklin, S., and Graesser, A., 1996: Is it an Agent, or just a Program, Proceedings Third International Workshop on Agent Theories, Architectures and Languages, Budapest, Hungary, 193-206.

Hogg, L. M., and Jennings, N. R., 1997: Social Rational Agents— Preliminary Thoughts, Proc. of Second Workshop on Practical Reasoning and Rationality, Manchester, UK.

Janca, P. C., 1995: Pragmatic Application of Information Agents, BIS Strategic Report.

Jennings, N. R., 1993: Commitments and Conventions: The Foundation of Coordination in Multi-Agent Systems, *The Knowledge Engineering Review* ***8*** (3), 223-250.

Jennings, N. R., and Campos, J. R., 1997: Towards a Social Level Characterisation of Socially Responsible Agents, *IEE Proceedings on Software Engineering* ***144*** (1) 11-25.

Kalenka, S., and Jennings, N. R., 1995: On Social Attitudes: A Preliminary Report, Proc. First Int. Workshop on Decentralised Intelligent Multi-Agent Systems, Krakov, Poland, 233-240.

Ketchpel, S., 1993: Coalition Formation Amongst Autonomous Agents, Proc 5th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Neuchatel, Switzerland, August 25-27.

Lesser, V.R., and Corkill, D.D., 1987: Distributed Problem Solving, in S. C. Shapiro (ed.), *Encyclopedia of Artificial Intelligence*, John Wiley and Sons, pp. 245-251.

Lesser, V.R., and Corkill, D.D., 1983: The Distributed Vehicle Monitoring Testbed: A Tool for Investigating Distributed Problem Solving Networks, *AI Magazine*, Fall, 15-33.

Marsh, S., 1992: Trust and Reliance in Multi-Agent Systems, Proc 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Rome, Italy.

Newell, A., 1982: The Knowledge Level, *Artificial Intelligence* ***18***, 87-127.

Overgaard, L., Petersen, H. G., and Perram, J. W., 1996: Reactive motion planning: a multi-agent approach, *Int. Journal of Applied Artificial Intelligence* ***10*** (1), 35-51.

Ovum Report (1994): Intelligent agents: the new revolution in software.

Russell, S., and Wefald, E., 1991: *Do the right thing*, MIT Press, Cambridge Mass.

Sen, S., 1996: Reciprocity: a Foundational Principle for Promoting Cooperative Behaviour among Self Interested Agents, Proc 2nd Int. Conf. on Multi-Agent Systems, Kyoto, Japan, 322-329.

Shehory, A., and Kraus, S., 1995: Task Allocation via Coalition formation among autonomous agents, Proc. 14th International Joint Conference on Artificial Intelligence, Montreal, Canada.

Sichman, J. S., Conte, R., Demazeau, Y., and Castelfranchi, C., 1994: A Social Reasoning Mechanism Based on Dependence Networks, Proc 11th European Conf on AI, Amsterdam, The Netherlands, 188-192.

Steels, L., 1989: Cooperation between distributed agents through self organisation, *Journal of Robotics and Autonomous Systems.*

Wavish, P., and Graham, M., 1996: A situated action approach to implementing characters in computer games, *Int. Journal of Applied Artificial Intelligence* ***10*** (1), 53-73.

Wellman, M. P., 1993: A Market-Oriented Programming Environment and its Application to Distributed Multi-commodity Flow Problems, *Journal of Artificial Intelligence Research* 1, 1-23.

Wooldridge, M. J., and Jennings, N. R., 1995: Intelligent Agents: Theory and Practice, *The Knowledge Engineering Review* ***10*** (2), 115-152.