Towards a social level characterisation of socially responsible agents

N.R.Jennings
J.R.Campos

Indexing terms: Computer level, Multi-agent systems, Socially responsible agents

Abstract: A high-level framework for analysing and designing intelligent agents is presented. The framework's key abstraction mechanism is a new computer level called the *social level*. The social level sits immediately above the knowledge level, as defined by Allen Newell, and is concerned with the inherently social aspects of multiple-agent systems. To illustrate the working of this framework, an important new class of agent is identified and then specified. *Socially responsible* agents retain their local autonomy but still draw from, and supply resources to, the larger community. Through empirical evaluation, it is shown that such agents produce both good system-wide performance and good individual performance.

1 Introduction

The number of multi-agent systems being designed and built is rapidly increasing as software agents gain acceptance as a powerful and useful technology for solving complex problems [1–3]. As applications become more complex, so there is a concomitant increase in the sophistication of the agent systems. In particular, there is a trend towards exploiting the metaphor of an *autonomous problem solving agent* much more fully (see [4] for a review of work in this area). Autonomy, in this context, means the ability of an agent to decide for itself which goals it should adopt, how these goals should be pursued (including deciding when interactions with other agents are appropriate) and which actions should be performed at which time.

In the early applications, such as the distributed vehicle monitoring testbed [5] and HEARSAY II [6], the individual components were of secondary importance compared with the needs of the overall system [Note 1]. The dominant design philosophy was reductionism: a design team had a particular problem to solve and they decomposed it into a number of sub-problems that they then assigned to one or more agents. The ensuing set of agents were then carefully engineered, at the level of the overall system, so that through their individual efforts and their interactions with one another the problem was solved. For example, the design team endeavoured to ensure that the system was constructed in such a way that conflicts, deadlocks, livelocks and resource starvation did not occur. The agents themselves had limited autonomy: their role in solving the overall problem was usually pre-ordained by the system designer, although they typically had the freedom to choose how they met their objectives. Also, because of the dominance of the holistic view, the majority of systems adopted the benevolent agent assumption [7, 8]: agents willingly performed all tasks requested of them by their acquaintances and volunteered their services to others if they had sufficient free capacity. In such cases, the agents' predisposition to be helpful to others, even to the detriment of their own problem solving in some cases, was another serious erosion of their autonomy. There was no worthwhile notion of them making a decision about which goals to adopt as all feasible requests were accepted.

Although this reductionist approach enabled many useful systems to be developed, it fails to exploit the full potential of the software agent paradigm, because its notion of agenthood is weak, and its emphasis on system level issues is restricting [9]. Recently, a significant number of developers have adopted a more constructionist approach to multi-agent system design, in which individual autonomous agents, rather than the overall system, are the central unit of analysis (see [10] for a discussion of the reasons behind this shift). This approach has the advantage of being better suited to open and distributed domains [9]. Also, by reducing the amount of design at the level of the overall system, larger problems can be tackled more easily, because the scope of the individual components is narrower [11]. Relevant exemplars include: game playing agents [12], robot manipulation agents [13], social simulation agents [14], transportation scheduling agents [15] and exploration agents [16]. In such cases, the overall system and its properties emerge out of the interplay between the agents.

The nature of this interplay is determined by two main factors. First, there are the type and degree of interdependence between the agents. For example, agents may have the potential to help, hinder, facilitate or negate [17–20] the actions of others. These depend-

© IEE, 1997

IEE Proceedings online no. 19971021

Paper received 22nd October 1996

N.R. Jennings is with the Department of Electronic Engineering, Queen Mary and Westfield College, University of Longon, Mile End Road, London E1 4NS, UK; n.r.jennings@qmw.ac.uk

J.R. Campos is with the Information Technology Department, LABEIN, Parque Tecnológico, Ed. 101, 48016 Zamudio, Bizkaia, Spain; campos@labein.es

Note 1: The term 'overall system' is used throughout this paper to refer to the set of agents in the multi-agent system.

encies are an inherent property of the environment in which the agents are situated and of how the environment's resources are divided between, and controlled by, the various agents. The second determining factor is the problem solving strategy adopted by the various agents. In the majority of cases, agents are designed to maximise their individual benefit and so they adopt a self-interested stance [21, 22] or they attempt to maximise their individual utility [23, 24]. Such strategies predominate because, in this agent-centric view of the world, success of the individual is the sole metric for performance evaluation.

The downside of empowering the individual agents is that it is considerably more difficult to predict how the overall system will behave/perform and what its properties will be. When agents are viewed as a tool for social simulation [14, 25], this is precisely what is desired. The aim of building the system is to harness the power of the computer to discover and observe the patterns of behaviour that emerge out of the interplay between the individuals. However arguably the role in which agent systems will have their greatest impact [25] is being used as a solution technology for complex industrial and commercial applications, e.g. industrial control [27], manufacturing [28], telecommunications [29] and air traffic control [30]. In such applications, a balance needs to be struck between the needs of the individual agents and those of the overall system. For example, if one agent has sole control over the production of a resource that several others need, then from a system perspective it is important that the provider is willing to produce this resource for the consumers in some situations (otherwise the system may deadlock). Similarly, if one agent can perform an action that negates the efforts of many others, then the majority should have some means of influencing the destructive agent so that it does not always perform its undesirable action (otherwise the system may be unproductive). This individual-community balance is required to ensure that the overall system, as well as the individual agents, is able to function in an effective manner, and it is necessary because of the dependencies that exist between the agents [Note 2].

In the previously described reductionist work, the individual-community balance was obtained through the holistic design stance and/or by making individual agents inherently helpful. In the constructionist work, when particular system-wide patterns were required the designer had to expend significant effort tuning the values and behaviour of the individual agents [35]. However neither of these approaches is very satisfactory: the former erodes the agent's autonomy, and the latter is something of a black art. Therefore this work takes the view that the best way of attaining such a balance is to develop agents that retain their local autonomy, are able to benefit from interactions with other agents, but that are nevertheless willing to provide some resources for the benefit of the overall system some of the time. Here, the term 'socially responsible' will be used to describe the problem solving behaviour of such agents.

Note 2: Inter-agent interdependencies are inevitable in such applications because complex problems cannot be decomposed into self-contained units operating solely on local resources [31]. In applications in which there is relatively little interdependence between the agents (e.g. information agents that scour the world wide web to find relevant data [32], softbots that provide an Internet power tool [33] and personal digital assistants that filter email [34]) there is no real need to perform such a balancing act.

Attaining socially responsible behaviour is the motivation behind a significant proportion of current research in the multi-agent system arena, indeed Gasser [36] and Hewitt [9] identify it as perhaps the key problem in this area. To date, a diverse range of solutions have been adopted, ranging from organisational structures [3, 4] and meta-level information exchange [41, 42] to multi-agent planning [7, 43–45]. Although all of these mechanisms offer some insight into how the balancing problem can be solved in a more or less narrow context, the solutions contain a vast array of application and implementation-specific details. This means the underlying foundational principles [Note 3] of responsible agents and of the overall system's structure are often obscured. Ideas akin to the responsible stance have also received backing from some social scientists [48–50] who argue that it is the most effective way of organising social structures. However, this work also lacks a clear view of the structures and mechanisms that must be in place to allow such behaviour to occur.

The purpose and contribution of this paper are therefore twofold. First, to identify the foundational principles and structures that enable coherent and effective overall systems to be constructed out of socially responsible agents. Secondly, to define an abstract framework in which the key features of a multi-agent system can be analysed and specified. The chosen means of tackling this latter problem is to follow and adapt an approach that proved highly successful in elucidating the foundational principles and structures of individual asocial agents. Newell's [46] knowledge level (KL) analysis provided the seminal characterisation of intelligent agents: stripping away implementation and application-specific details to reveal the core of asocial problem solvers. As this work seeks to do the same for social agents, Newell's basic approach is aped.

Section 2 shows that a KL analysis cannot adequately describe the behaviour of responsible agents. In particular, its key tenet of individual rationality fails to deal with the notion of being helpful or sometimes doing things for the greater good. This Section also highlights the limitations of a number of extant multi-agent system models in describing socially responsible agents. The conclusion of this analysis is that a new computer level needs to be defined. This level is called the social level [Note 4] (SL): it sits immediately above the KL and provides firm social principles and foundations for responsible agents. The primary benefit of an SL description is that it enables the overall system behaviour to be studied without having to delve into the implementation details of the individual agents; thus prediction of the behaviour of the social agents and of the overall system can be made more easily. Section 4 investigates empirically how the performance

Note 3: Foundational, in this context, means abstract and broadly applicable guidelines that define how the overall system should be structured and how the constituent agents should behave during their problem solving. For example, the notion of rationality (if an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action [16] is a foundational principle for individual agents. Similarly, Assimov's [47] laws (a robot may not injure a human being or, through inaction, allow harm to come to a human being; a robot must obey the orders given to it by human beings, except where such orders would conflict with the first law; a robot must protect its own existence as long as such protection does not conflict with the first or second laws) are foundational principles for physical agents.

Note 4: In earlier work, we used the term 'co-operation knowledge level' to identify this level [51]. However, we believe that such a level can describe multifarious social activities, not all of which are co-operative. Hence, we changed to a more encompassing terminology.

of socially responsible agents compares with that of other types of agent. In particular, it highlights how responsible agents strike a good balance between the needs of the individuals and those of the overall system. Section 5 compares the SL concept with related work in the fields of distributed artificial intelligence, decision theory, economics and game theory. Finally, some recommendations for future work in extending and utilising the SL are given.

2 Knowledge level representations

The fundamental insight and power of Newell's [46] KL characterisation is that there are a few key concepts that underpin the behaviour of a large class of problem solving agents. By explicitly drawing out these concepts, it is possible to understand and predict agent behaviour without having to delve into the operational model of processing that is actually being employed. Analysing an agent at the KL involves treating it as having some knowledge and some goals and believing it will do whatever is within its power to attain its goals, insofar as its knowledge allows. Thus for an observer at the KL, if two agents have the same knowledge and the same goals, they are indistinguishable even if they have radically different physical structures (e.g. knowledge representation schemes, functional decompositions, problem-solving paradigms etc.). Agent problem-solving behaviour can then be characterised through the principle of rationality:

> 'if an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action' [46]

Given the success of the KL approach [52], it seemed natural to see whether it could be used directly to describe problem solving in multi-agent contexts. Aitken *et al.* [53] attempt to do precisely this and present a KL analysis of a number of multi-agent systems (focusing in particular on the problem of global coherence). However, their work has three major shortcomings: first, it focuses on a limited type of agent (knowledge-based systems) and a limited range of social phenomenon (co-ordination); secondly, it describes agents at the level below the KL (the symbol level), which means that many of the key facets of their decision-making and problem-solving behaviour, as they relate to social problem solving, are lost; and thirdly, it assumes the structures and mechanisms used for individual problem solving, are appropriate for describing social phenomena (see the following scenarios for counter examples). In addition to the specific criticisms of their work, the basic approach is fundamentally flawed. First, multi-agent systems are best described using models that treat the existence and interactions of multiple actors as a fundamental category [36]; the KL has no concept of distribution and, hence, cannot represent multiple entities nor interactions. Secondly, the principle of rationality is insufficient for describing a range of desirable social behaviour. For example, consider the case in which a number of agents decide to work together to search for a lost item (adapted from [54]). Assuming the item is discovered, one of the agents will have satisfied its original goal. However, it is part of the intuitive notion of working together that the successful agent should inform its fellow searchers that the target has been found, so that they can abandon their search [55–57]. However, such additional behaviour is not warranted by the principle of rationality, because this informing action does not satisfy one of the finder's goals. In particular for rational agents, the notion of performing actions for the greater good is not permitted: helpful actions not connected to one of the agent's goals contradict the principle of rationality (helpful actions that further the agent's goals are naturally permissible).

An alternative means of using the KL formulation is to change the unit of analysis from the individual agent to the entire multi-agent system. This would be consistent with the symbolic interactionist school of social psychology that postulates that the conduct of individual entities should be explained in terms of the organised conduct of the social group [58] (rather than accounting for the organised conduct of the social group in terms of the separate individuals belonging to it). In this view, the principle of rationality would refer to the overall system rather than to its individual components (an approach that is broadly in line with the reductionist design philosophy). Thus a society should only carry out actions that lead to one of its goals. Although this approach may suffice for collections of relatively straightforward entities, e.g. bees [59] or neurons [60], it is inappropriate for societies of autonomous agents, as the goals of the individuals, which may differ from those of the society, are not represented. There are also many theoretical problems associated with attributing knowledge to a collective [61, 62].

Given the shortcomings of using the KL directly, the next logical step is to see whether it can be extended for use in social contexts. Here, the obvious place to start is with Newell's own work on human problem solving. Newell's [63] unified theory of cognition identifies a separate level above the rational level (the human equivalent of the KL) for dealing with social contexts. He terms this level the social band. He argues that a separate level is needed because social systems cease to act, even approximately, as a single rational agent. This divergence occurs because in a social group both the knowledge and the goals are distributed and the group cannot assemble all the information relevant to a given goal (because of the relatively small communication bandwidth available in proportion to the knowledge of the agents). The social band interacts with the rational band to define an individual's behaviour in a social context. Although Newell acknowledges that his social band is not clearly defined, he states that it should only contain knowledge that is specifically social in nature (e.g. norms, values, morals, myths etc.).

In subsequent work, Carley and Newell [64] construct a model social agent to identify more clearly the range of behaviour for which the social band must account. They describe the social agent along two dimensions: processing capabilities and differentiated knowledge of the self, task, domain and environment. In terms of their processing dimension, our work is primarily concerned with social rational agents; thus our agents are not omniscient, we do not consider emotional agents [Note 5], nor do we consider the problem of bounded social rationality. Thus, just as rationality is the ideal for individual agents and bounded rationality allows more tractable agents to be constructed [66, 67], so we first need to uncover the ideal for responsible agents before we can start building more tractable responsible agents. In terms of the knowledge dimension, our work is primarily concerned with identifying

Note 5: Refer to [65] for a good example of work in this area.

the social structures and social goals that need to be in place for responsible agents to operate effectively. Although their taxonomy helps to identify the types of structure and mechanism that need to be described, and it offers a preliminary analysis for humans, it fails to provide any substantial details that can be used to describe our artificial socially responsible agents

Based on this broad method of approach, identification of the foundational principles of socially rational agents requires three steps to be completed. First, the social band concept needs to be mapped from the domain of human cognition into the domain of artificial systems. Secondly, the social band for artificial systems needs to be structured in such a way that it is useful for describing computer-based social agents. Finally, the general social framework needs to be used to specify a particular class of agents, namely socially responsible agents. The following Section addresses each of these issues in its characterisation of the SL and its subsequent usage in specifying socially responsible agents.

3 Detailing the social level

The structure of this section closely follows that used by Newell to introduce the KL in his original paper on the subject. It starts with a statement of existence:

> Social level hypothesis: there exists a computer level immediately above the KL, called the social level, which is concerned with the inherently social aspects of multiple agent systems.

The SL provides an abstract characterisation of those aspects of multi-agent system behaviour that are inherently social in nature. Thus it is concerned with representing phenomena such as co-operation, co-ordination, conflicts and competition. For such phenomena, it describes the structures and mechanisms which must be present inside the agents and inside the overall system to produce the desired type of behaviour. To be considered a distinct level, the SL must be defined in a manner that is independent of the KL. However, as the SL sits immediately above the KL, there must be a clear mapping between the two. The SL can be described using Newell's general nomenclature: the system (the entity to be described at that level), the components (the primitive elements from which the system is built up), the composition laws (the rules that define how the components are assembled to form the system), the behaviour laws (the principles that determine how the system's behaviour depends upon its composition and on its components' behaviour) and the medium (the elements the system processes to obtain the behaviour it was designed to achieve). These five attributes cover the main viewpoints, structural, behavioural and functional [68], from which a system can be described, and therefore they offer a complete characterisation. The structural viewpoint equates to the system's components and the composition laws, the behavioural viewpoint is covered by the behavioural law, and the functional viewpoint is strictly related to the medium.

In this paper, the particular phenomenon that will be described at the SL is social responsibility. Producing such a description entails defining what the system, the components, the compositional laws, the behavioural laws and the medium are for responsible agents (see Table 1). However, the abstract characterisation offered by the SL structure is also applicable to other types of multi-agent system. For example, communities of benevolent or autonomous agents could also be described using the SL nomenclature. In such cases, some of the details will obviously vary (e.g. the behavioural laws and the medium), but the basic structure remains.

Table 1: Defining attributes of the knowledge and social levels

Attributes	Knowledge level	Social level
System	Agent	Responsible society
Components	Goals and actions	Members, environment, interaction means and goals
Compositional laws	None	Variable
Behaviour laws	Principle of rationality	Principle of social rationality
Medium	Knowledge	Acquaintance, influence, rights and duties

3.1 System

As discussed in the preceeding Sections, there are many different types of multiple agent system. However to define clearly the scope of this work, this analysis is specifically targeted at systems in which

- the agent's are autonomous
- the agents balance their individual needs with those of the overall system.

The first part of the definition refers to agents that choose for themselves when, how and with whom they interact. The second part requires the agents to be capable of assessing the impact of their choices on their own problem solving and on that of the overall system. For the purposes of this work, let multi-agent systems in which both of these conditions are satisfied be called responsible societies. Other types of multi-agent can also be described using SL characterisations, however, here we limit ourselves solely to responsible societies.

3.2 Components

A responsible society is composed of the following components: members (entities doing the problem solving); environment (where members are situated); interaction means (ways members interact); and goals (motivational force for members' problem solving). Together these components completely characterise the system: all the society's elements are contained in the members or in the environment; interaction means describe the way members interact with one another and with their environment; and goals explain why members carry out problem solving.

3.2.1 Members: These are the society's problem solving components. The most basic member is a socially responsible agent. However, in general, members can be composed of many such agents or even other members. This recursive definition of the problem solving entity has two main advantages. First, arbitrarily complex structuring of the society can be described in a uniform manner [27]. Secondly, it means that whatever can be said about the society can also be said about its members [70]. Thus, a member is a soci-

ety in its own right and can be represented at the SL. The distinction between a responsible agent and a basic member is the level at which it is represented. If it is represented as a member, then it is viewed using SL constructs, whereas, if it is represented as an agent, it is viewed using KL constructs. In either case, it will have the same overall properties and will be the same entity. The benefits of this representational duality are that it allows a clear distinction to be drawn between the knowledge and social levels. It also means that the problem-solving entities can be described, viewed and analysed at different levels of granularity and detail, depending on the needs of the observer.

3.2.2 Environment: This is the context in which the members perform their problem solving, e.g. the Internet, a nuclear power plant or an office. In most agent systems, there is a bi-directional interplay between the agents and the environment. Thus, in this context, members can perform actions that alter their environment, and changes in the environment can alter the actions of members.

3.2.3 Interaction means: These are the ways in which members interact. They offer the sole means by which members can influence one another and manage their interdependencies, e.g. change the view a member has of itself, of other members, of the environment or of the actions it should perform. This component is purely social as it requires at least two members to be present. In the general case, there are three interaction means: first, directly through whatever communication channels exist (e.g. using an agent communication language [71] over the Internet or via a shared blackboard [72]); secondly, indirectly through the environment: when a member modifies the environment, it may also modify the perception of other members who can detect that change [16]; and finally, via an intermediator such as a mediator or a facilitator [71] (who must ultimately use one of the first two approaches).

3.2.4 Goals: Members carry out problem solving to achieve goals. These goals can be precise (e.g. rendezvous with member M1 at 12 pm) or they can be relatively vague (e.g. 'thrive' or 'survive') They can be encoded explicitly (e.g. in BDI architectures [30, 56]) or implicitly (e.g. in reactive architectures [73]) within the members. Goals can either be individual and belong to one member or they can be social and belong to a number of members (or even to the whole society). A member's goals can be satisfied in three ways: by actions it carries out itself, by actions carried out by other members or by a combination of the two.

3.3 Composition laws

Compositional laws govern the way the society's members are structured. This structure gives information about the types of inter-agent interdependency that exist and therefore is a crucial determinant of social behaviour. Both practical (e.g. [5, 74]) and theoretical (e.g. [75, 76]) analysis has shown that multi-agent societies can be organised in many different ways; e.g. in flat, hierarchical, matrix, circular or linear topologies [75]. The reason for this burgeoning variety is explained by the premises of contingency theory [75]

- there is no best way to organise
- all ways of organising are not equally effective.

Hence, the chosen compositional structure is important, but it will depend upon the characteristics of the domain under analysis: its function, its spatial distribution, its control relationships and so on. In some cases, the society may even change its composition at run time to cope better with its evolving problem-solving requirements [77].

3.4 Behaviour laws

Like their KL counterpart, the behavioural laws are perhaps the key component of an SL description. As Section 2 demonstrated, the principle of rationality is inadequate for describing the full range of behaviours that socially responsible agents are required to exhibit. Hence, a new law needs to be defined. Here, this law will be called the principle of social rationality. This principle defines how the society's individual members should behave [Note 6].

In contrast to Newell's characterisation of the principle of rationality, it was decided to adopt a more formal, utility-based characterisation of the behavioural law [Note 7]. The motivation for this is that the principle of social rationality is a more complex expression than its KL counterpart, and, hence, a more descriptive framework needs to be put in place to characterise it concisely and unambiguously. Let

- $\mathbf{S}$ be the responsible society under study
- $\mathbf{M}$ be a responsible member of $\mathbf{S}$
- $\mathbf{A}(\mathbf{M})$ be the set of actions $\mathbf{M}$ can perform
- $\mathbf{a}$ be an element of $\mathbf{A}(\mathbf{M})$.

As the system under study is a socially responsible society (Section 3.1), derivation of the behavioural law starts from a macro-level perspective. A reasonable and sustainable basic premise is that all actions $\mathbf{a}$ carried out by all members $\mathbf{M}$ must have a positive overall net benefit to society $\mathbf{S}$. Designating the net benefit for member $\mathbf{J}$ as $\mathbf{N_J(M,a)}$,

$$\sum_{\mathbf{J}\in\mathbf{S}} \mathbf{N_J}(\mathbf{M},\mathbf{a}) > \mathbf{0} \qquad (1)$$

To cover all possible benefits only once, it is assumed that all members have disjoint sets of agents $\forall \mathbf{J}, \mathbf{K} \in \mathbf{S}, \mathbf{J} \neq \mathbf{K}, \mathbf{J} \cap \mathbf{K} = \varnothing$, and that their union covers all the agents in thc society $\cup_{\mathbf{J}\in\mathbf{S}} \mathbf{J} \equiv \mathbf{S}$. In more detail, let

- $\Delta(\mathbf{M},\mathbf{a})$ be the benefit to $\mathbf{S}$ when $\mathbf{M}$ performs $\mathbf{a}$ (social benefit)
- $\Omega(\mathbf{M},\mathbf{a})$ be the loss to $\mathbf{S}$ when $\mathbf{M}$ performs $\mathbf{a}$ (social loss)
- $\delta(\mathbf{M},\mathbf{a})$ be the benefit to $\mathbf{M}$ when it performs $\mathbf{a}$ (member benefit)
- $\omega(\mathbf{M},\mathbf{a})$ be the loss to $\mathbf{M}$ when it performs $\mathbf{a}$ (member loss).

Without loss of generality, it is assumed that all of the above amounts are positive or zero and that they are measured on the same scale (or at least on a comparable one). 'Benefit' corresponds to all that the entity gains when it performs an action; 'loss' corresponds to the cost of performing the action. For instance, a

Note 6: In practice, the principle of social rationality can either be explicitly encoded in the members' data structures and processes or it can be used as a general guideline to which the designers of responsible members must adhere.

Note 7: Naturally we acknowledge the difficulties associated with precisely quantifying utility values. However, the purpose of this work is to identify foundational principles and so the practical difficulties of doing this are not considered here.

number of information retrieval agents may decide to work together to deal with complex user requests. When a request is received, a number of agents search for the target item, and the one who finds it receives 50% of the price paid by the user (member benefit). However, during their search, the agents incur some costs (member losses) (e.g. costs to access databases, transportation costs etc.). The remaining 50% of the user's payment is placed in a common pool that is divided up equally between the participating agents at the end of the week (social benefit). In one particular week, the agent pool may take on a large query for an important customer which means that it has to refuse a number of other queries (social loss). Thus, when performing a particular search, the benefits and losses are in some way distributed between the members who perform the actions and the society in which the members are situated. $\Delta(\mathbf{M},\mathbf{a})$ includes all the benefits the society accrues when **M** performs **a** [Note 8]; whereas $\delta(\mathbf{M},\mathbf{a})$ includes just the benefit that **M** receives for **a**. The same interpretation also applies to $\Omega(\mathbf{M},\mathbf{a})$ and $\omega(\mathbf{M},\mathbf{a})$. Note that this work is not concerned with the causes or nature of these benefits/losses or the way they may be quantified, as such computations are highly problem-specific. The interested reader is referred to any standard text on utility or decision theory [78–80] for an account of how to estimate these amounts, and to the work of Gmytrasiewicz and Durfee [23], in particular on how to quantify the different kinds of action.

With this structure in place, overall net benefit can now be defined as $[\Delta(\mathbf{M},\mathbf{a}) - \Omega(\mathbf{M},\mathbf{a})] + [\delta(\mathbf{M},\mathbf{a}) - \omega(\mathbf{M},\mathbf{a})]$. Expressing eqn. 1 in these terms allows the relation between the society and its members to be identified [Note 9]:

$$\forall \mathbf{M} \in \mathbf{S}, \forall \mathbf{a} \in \mathbf{A}(\mathbf{M})$$
$$\Delta(\mathbf{M},\mathbf{a}) + \delta(\mathbf{M},\mathbf{a}) > \Omega(\mathbf{M},\mathbf{a}) + \omega(\mathbf{M},\mathbf{a}) \qquad (2)$$

Defining joint benefit as $\Delta(\mathbf{M},\mathbf{a}) + \delta(\mathbf{M},\mathbf{a})$ and joint loss as $\Omega(\mathbf{M},\mathbf{a}) + \omega(\mathbf{M},\mathbf{a})$, eqn. 2 can be paraphrased as 'only perform actions whose joint benefit is greater than their joint loss'. From this, the principle of social rationality can be formulated

> Principle of social rationality: If a member of a responsible society can perform an action the joint benefit of which is greater than its joint loss, then it may select that action.

From this principle, a number of important observations can be made about the types of behaviour in which responsible agents can engage:

- Corollary 1: Members may perform actions for which their member benefit is less than their member loss. This may occur if the society gains more in total than it loses. This captures the responsible notion of doing actions for the benefit of the society [Note 10].

- Corollary 2: Members may perform actions which bring them personal benefit, but which are detrimental to the overall society. This may occur if the member benefit is greater than the societal loss. This captures the empowerment notion of responsibility and the intuitive notion that agents should derive individual benefit from interactions with others.

- Corollary 3: The notion of society cannot be separated from that of its members: social rationality implies considering both joint benefits and joint losses. Thus, in keeping with the essence of the individual-community balance problem, social responsibility cannot be explained just in terms of the benefit to the member, nor just in terms of the benefit to the society.

- Corollary 4: Social rationality is insufficient to determine the actual behaviour of the society in many situations (e.g. when more than one member can accomplish an action). Social rationality also says nothing about the relative importance of actions (e.g. it does not say that social actions should take precedence over individual ones, nor even that they should be equally balanced). Both of these situations, and many others besides, can only be covered by adding domain-dependent knowledge and problem-solving requirements, both at the knowledge and symbol levels of the individual members.

In more detail, attending to the relative values of $\Delta(\mathbf{M},\mathbf{a})$, $\delta(\mathbf{M},\mathbf{a})$, $\Omega(\mathbf{M},\mathbf{a})$ and $\omega(\mathbf{M},\mathbf{a})$, four classes of action can be identified (Table 2). These classes are disjoint, $\forall \mathbf{J},\mathbf{K} \in \{\mathbf{S}(\mathbf{M}),\mathbf{I}(\mathbf{M}),\mathbf{D}(\mathbf{M}),\mathbf{F}(\mathbf{M})\}$, $\mathbf{J} \neq \mathbf{K}$, $\mathbf{J} \cap \mathbf{K} = \varnothing$, and complete, $\mathbf{S}(\mathbf{M}) \cup \mathbf{I}(\mathbf{M}) \cup \mathbf{D}(\mathbf{M}) \cup \mathbf{F}(\mathbf{M}) \equiv \mathbf{A}(\mathbf{M})$. This structuring allows a comprehensive analysis to be undertaken of the relationships between the benefits/losses of the society and of the members. It should be noted that $\Delta(\mathbf{M},\mathbf{a})$ and $\delta(\mathbf{M},\mathbf{a})$ are independent, as a member's social benefit is separate from its individual benefit. Thus there can be situations in which $\Delta(\mathbf{M},\mathbf{a})$ is greater than $\delta(\mathbf{M},\mathbf{a})$ and situations in which the opposite is true. As an example of the former, member **M** may carry out action **a** to satisfy a goal of another member: $\delta(\mathbf{M},\mathbf{a})$ zero and $\Delta(\mathbf{M},\mathbf{a})$ positive. As an example of the latter, **M** may carry out **a** to satisfy a local goal: $\Delta(\mathbf{M},\mathbf{a})$ zero and $\delta(\mathbf{M},\mathbf{a})$ positive. The same considerations also hold for $\Omega(\mathbf{M},\mathbf{a})$ and $\omega(\mathbf{M},\mathbf{a})$.

Table 2: Action classes and relationships between their benefits and losses

Type	Society	Member
Social: **S(M)**	$\Delta(\mathbf{M},\mathbf{a}) > \Omega(\mathbf{M},\mathbf{a})$	$\delta(\mathbf{M},\mathbf{a}) \leq \omega(\mathbf{M},\mathbf{a})$
Individual: **I(M)**	$\Delta(\mathbf{M},\mathbf{a}) \leq \Omega(\mathbf{M},\mathbf{a})$	$\delta(\mathbf{M},\mathbf{a}) > \omega(\mathbf{M},\mathbf{a})$
Divided: **D(M)**	$\Delta(\mathbf{M},\mathbf{a}) > \Omega(\mathbf{M},\mathbf{a})$	$\delta(\mathbf{M},\mathbf{a}) > \omega(\mathbf{M},\mathbf{a})$
Futile: **F(M)**	$\Delta(\mathbf{M},\mathbf{a}) \leq \Omega(\mathbf{M},\mathbf{a})$	$\delta(\mathbf{M},\mathbf{a}) \leq \omega(\mathbf{M},\mathbf{a})$

3.4.1 Social actions: Social actions are those that have a net benefit to the society, but that fail to benefit the member that carries them out. An example is when a member undertakes a goal, which is not one of its own, for the good of the society. In this case, the member obtains no direct personal benefit (apart from its share of $\Delta(\mathbf{M},\mathbf{a})$) and it incurs some loss, as it spends resources and time performing **a**. Inside this class are pure social actions: those whose benefit to the member and loss to the society are zero:

Note 8: Included in $\Delta(\mathbf{M},\mathbf{a})$ is the proportion that **M** receives, as part of its role in the society. This amount is in addition to its individual member benefit $\delta(\mathbf{M},\mathbf{a})$.

Note 9: This form of the equation provides a basic characterisation of the problem. In certain contexts, it may be desirable to assign differing weights to the individual and the social components to reflect a particular emphasis in system design. In certain domains, there may even be better ways of combining the values. However, our work aims to identify the main ingredients at this stage, and so we stick with the simple form.

Note 10: This represents a fundamental departure from the notion of individually rational behaviour. Individually rational agents only perform actions that bring them personal benefit. Thus, using our framework, Newell's principle of rationality can be expressed as $\delta(\mathbf{M},\mathbf{a}) > \omega(\mathbf{M},\mathbf{a})$.

$$\mathbf{PS(M)} \equiv \{\mathbf{s} \in \mathbf{S(M)} : \delta(\mathbf{M},\mathbf{s}) = \Omega(\mathbf{M},\mathbf{s}) = \mathbf{0}\}$$

Substituting into the definition of social rationality

$$\forall \mathbf{M} \in \mathbf{S}, \forall \mathbf{s} \in \mathbf{PS(M)} \qquad \Delta(\mathbf{M},\mathbf{s}) > \omega(\mathbf{M},\mathbf{s}) \qquad (3)$$

where $\Delta(\mathbf{M},\mathbf{s})$ is positive and $\omega(\mathbf{M},\mathbf{a})$ is positive or zero. This delimits the situations in which members perform personally detrimental actions for the good of the society. It places a limit on the member's helpfulness: a member will not perform actions which bring a personal loss greater than the benefit accrued by the society. Note, by definition,

$$\forall \mathbf{M} \in \mathbf{S}, \forall \mathbf{s} \in \mathbf{PS(M)} \qquad \delta(\mathbf{M},\mathbf{s}) \leq \omega(\mathbf{M},\mathbf{s}) \qquad (4)$$

which means that this class of action violate Newell's principle of rationality.

3.4.2 Individual actions: Individual actions are those that have a net benefit to the member who carries them out, but which do not report a net benefit to the society. This occurs when a member performs an action to satisfy one of its goals, and that goal does not further the aims of the society as a whole. In this case, the society gains no benefit and it may even experience a loss if the member uses the society's resources to accomplish its goal. Inside this class are pure individual actions: those whose loss to the member and benefit to the society are zero

$$\mathbf{PI(M)} \equiv \{\mathbf{i} \in \mathbf{I(M)} : \Delta(\mathbf{M},\mathbf{i}) = \omega(\mathbf{M},\mathbf{i}) = \mathbf{0}\}$$

Again substituting into the definition of social rationality,

$$\forall \mathbf{M} \in \mathbf{S}, \forall \mathbf{i} \in \mathbf{PI(M)} \qquad \delta(\mathbf{M},\mathbf{i}) > \Omega(\mathbf{M},\mathbf{i}) \qquad (5)$$

where $\delta(\mathrm{M,s})$ is positive and $\Omega(\mathbf{M},\mathbf{a})$ is positive or zero. This quantifies the cases in which members may perform actions that bring them personal benefit and that are detrimental to the overall society. This does not mean that members are free to perform whatever actions they see fit without regard to the rest of the society (unlike individual utility maximising agents). Rather they can only perform those actions in which their individual benefit is greater than the society's loss. Note, by definition

$$\forall \mathbf{M} \in \mathbf{S}, \forall \mathbf{a} \in \mathbf{A(M)} \qquad \Delta(\mathbf{M},\mathbf{i}) \leq \Omega(\mathbf{M},\mathbf{i}) \qquad (6)$$

which shows that it is impossible simply to apply the principle of rationality to the society as a whole, as there are cases in which members perform actions the result of which is a net loss for the society.

3.4.3 Divided actions: Divided actions are those that result in a net benefit to the member that carries them out and to the society as a whole. These actions straightforwardly satisfy both the principle of rationality and the principle of social rationality. They are a trivial case and shed no light on the trade-offs involved in designing socially responsible agents.

3.4.4 Futile actions: Futile actions are those that bring no benefit to the member or the society. These actions satisfy neither the principle of rationality nor the principle of social rationality.

Eqns. 3 and 5 show the connection between a society and its members. They are both composed of crossed terms, giving further credence to the claim that the notion of the overall system cannot be separated from that of its members [36, 49]. Hence, no description of social responsibility, nor any solution to the individual-community balance problem, can be made solely in terms of members or solely in terms of the society (corollary 3). The equations constrain both the behaviour of the members and of the society as a whole. In terms of the former, they ensure that the society is protected from actions that would cause it significant harm. In terms of the latter, it ensures that members are not overly exploited for the good of the society. Eqns. 4 and 6 define the situations in which the members or the society are prepared to experience a loss so that the other can obtain a greater benefit. Thus, when participating in a society, members lose some privileges, but in compensation they gain some new ones (and the gain will be greater than the loss, see Section 4).

3.5 Medium

The medium is what the members process to obtain the behaviour specified by the principle of social rationality. For responsible societies, the minimum set of necessary concepts are: acquaintance: the notion that the society contains other members; influence: the notion that members can affect one another; and rights and duties: the notions that a member can expect certain things from others in the society and that certain things are expected of it by the society.

- Acquaintance: To be capable of balancing its needs against those of the society, a member must first be aware that there are other members who can derive some benefit or incur some loss as a consequence of its goals. This does not require a member to be aware of every other member in the society; indeed, the member may simply view everything else as either the society or the environment and not be aware of their internal workings or their structure. However, the notion that acquaintances exist is insufficient for responsible behaviour. What is also needed is for the society to be aware of the member; without this reciprocal notion a member could not exploit the society's resources for its personal benefit (i.e. pure individual actions could never occur).

- Influence: The second component required for responsible behaviour is that members can, under certain circumstances, influence the goals of other members (e.g. a member may request assistance) and similarly that the society can influence the goals of its members (i.e. the helpfulness aspect of responsibility). The actual way this influence is exerted, e.g. through the promise of prizes, threats or sanctions [81], is unimportant at this level of description; it is simply sufficient to say that it will be through the interaction means.

- Rights and duties: The final concepts required for responsible behaviour are related to rights (things the member can do as a result of being in the society) and duties (things expected of the member as a result of being in the society) [Note 11]. In human societies, for example, rights take the form of 'cans' (e.g. 'you can use public transport' or 'you can join this association'), and they indicate ways in which members are able to

Note 11: Social scientists often use the notions of rights and duties (also called obligations or norms) to explain two of the key questions about sociality: how does a self-interested agent do something for the group which may cost it more than it gains, and how does an agent cope with cheating and free-riding? Social norms and the obligation to reciprocate the benefits received are fundamental mechanisms that prevent cheating and foster group-wide concerns [46].

exploit the society's resources [82]. Duties are restrictions imposed upon the member by the fact that it is in a society; members accept such limitations because they are offset by the rights endowed by being a society member. In general, duties take the form of 'shoulds': examples from human society include 'you should help elderly people' and 'you should not kill' [82]. From these basic concepts, responsible behaviour requires the members to be aware that they receive certain rights for being in the society (e.g. the ability to exploit, to a certain extent, the resources of the community for the good of their own problem solving), but that the cost of receiving these rights is that they also have certain duties to perform (e.g. they should sometimes perform actions for the good of the others, even if they are personally detrimental). The SL is not concerned with any particular instantiation of rights or duties, only with the fact that such notions exist. In different scenarios, and even between different members in the same scenario, rights and duties may vary. The only restriction is that their instantiation must be consistent with the principle of social rationality (i.e. there can be no rights or duties that allow members or the society to transgress the basic principles).

All of the above concepts must be in place before responsible societies can be developed. The notion of acquaintance must be in place before interaction can occur, as no member can tell that its goals will influence those of others if it is unaware of the concept of others. Also, it is straightforward to claim that the concept of influence must be present if rights and duties are to exist: a member cannot realise that its actions can be enhanced or hindered if it does not realise that its actions can be influenced or modified by the rest of the society. The concepts of rights and duties are what underpin the various types of actions identified as permissible by the principle of social rationality. Eqn. 3 relates to member's rights, eqn. 4 relates to member's duties, eqn. 5 relates to the society's rights, and eqn. 6 relates to the society's duties. Thus, for example, without the notion of duties, there would be nothing preventing a member from transgressing eqn. 5 and overly exploiting the society for its individual gain. Similarly, without the notion of rights there would be nothing preventing the society from exploiting its members as expressed in eqn. 3 [Note 12].

4 Experimental evaluation

There are a number of important questions that need to be answered before it can be claimed that socially responsible agents provide a good solution to the individual-community balance problem:

- How do the responsible agents fare in terms of their individual benefit? How does the overall system fare in terms of the global benefit?
- How do socially responsible agents compare, on an individual and overall system perspective, with other types of multi-agent system?
- How do responsible communities behave when there is an imbalance in terms of the value of individual and/or social actions?

Note 12: When it is said that something is related to the member's rights, it could equally well be described as the society's duties (and the same for all the other cases). This is because the notion of right in the acting entity is inseparable from the notion of duty in the entity which is acted upon.

To provide a means of assessing the computational performance of responsible agents, a comparison with three other commonly occurring types of agent system was undertaken:

- Individually rational agents: agents that adhere to Newell's principle of rationality and only perform goals that bring them a positive net benefit; typical of the self-interested, utility-maximising agents that predominate in the constructionist view of multi-agent systems
- Benevolent agents: agents that accept all goals they are capable of performing, typical of the agents that predominate in the reductionist view of multi-agent systems
- Selfless agents: agents which only perform goals if they have a positive net benefit to the society. These agents are not concerned with their individual benefits or losses; they are the antithesis of self-interested agents.

To enable comparisons to take place, all four classes of agent were situated in identical social contexts. The agents and their interactions were made as simple as possible, so that the results depended solely on the type of the agent rather than on the vagaries of a particular co-ordination or co-operation algorithm. For this reason, each experiment involved precisely two members, the individual whose performance is being monitored and the rest of the society (an arbitrarily complex collection of members in some arbitrary social organisation).

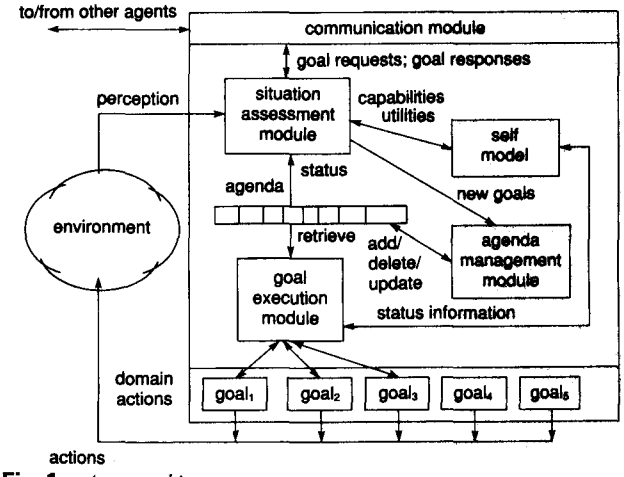

Fig. 1 *Agent architecture*

All agents had the same basic architecture (Fig. 1) which is loosely based on GRATE* [56]. At the core of this architecture is a situation assessment module that generates new local goals for the agent (based on environmental conditions) and decides whether to adopt incoming requests from other members in the community. This component is the one which embodies the agent's behavioural rules. The agenda management module ensures the agenda is kept up to date and changed to reflect appropriately the agent's current perception of its situation. The goal execution module is responsible for determining how the goals should be met and then invoking the necessary domain problem-solving activity. The self model contains domain-specific information about the agent's own capabilities.

The agent architecture is implemented in C++ and uses the message passing facilities provided by the parallel virtual machine (PVM) [83] distributed computing environment. During the course of the experiments, the

agent collects all new potential goals that have arisen since the last goal was started (some of which will be requests from other agents, and some of which will be the consequence of the agent's own problem-solving role). Each of these potential goals has an associated member and society benefit [Note 13] (stored in the self model). The agent then assesses the various benefit values and decides whether to adopt the goal (situation assessment module). If the goal is acceptable, it is passed to the agenda management module which places it in the appropriate position on the agent's agenda. The goal at the front of the agenda is then processed; this takes a varying amount of (real) time, depending on its complexity. The loop then restarts.

In terms of their SL characterisations, all communities have the same basic components. The systems are composed of the same number of members, who are situated in the same environment (which gives reproducible patterns of problem-solving behaviour), with the same interaction means (directed message passing only) and the same mix of goals (between joint and individual). The compositional law in each case is peer-to-peer. The behavioural law for each member type class obviously differs, although within a given class all members share the same instantiation. Members model all their acquaintances and they all know which members can assist with their goals. Members influence one another by sending their rating of the importance of the goal along with their goal requests. Members are assumed to have a common frame of reference and are assumed to be accurate and truthful about their values. Agents have no special rights or duties over and above those associated with their behavioural role. Individually rational agents have the right to refuse goals that violate the principle of rationality; selfless agents have the right to reject goals that do not bring benefit to the society; benevolent agents have the duty to accept all requests; and responsible agents have the right to exploit others and the duty to help others as specified in eqns. 3–6 in Section 3.

To answer the questions posed at the beginning of this Section, a number of experiments were undertaken. In these experiments, two main dimensions were varied: the load on the individual agents and the relative importance placed upon the individual and the societal goals. This led to four types of experiment:

- the agents have a heavy processing load (Section 4.1): new goals constantly arrive at the society and at the individual members; the agents always have some goals to process.
- the agents have a light processing load (Section 4.2): significant gaps exist between the arrival of new goals; agents are often idle as they have no appropriate goals to process.
- the goals of the society are dominant (Section 4.3): the first two experiments assume an approximately equal importance weighting for the individual and social benefits. These experiments examine what happens when the social benefits significantly outweigh the individual ones.
- the goals of the individual are dominant (Section 4.4): same as above, except that the individual benefits significantly outweigh the social ones.

Note 13: As a simplification, the notion of loss is incorporated as a negative benefit. Thus, if the member benefit to **M** for action **a** is 10 and the associated member loss for **a** is 20, then $\delta(\mathbf{M}, \mathbf{a})$ is –10. A similar interpretation is assumed for societal benefit and loss.

In the graphs in the following Sections, the unit of time displayed is seconds, and it equates to the real-time operation of the simulation; to avoid statistical anomalies associated with start-up and termination conditions, the results are taken from the simulation's middle period (between 100s and 500s). In all cases, the time taken to complete a goal varies between 1s and 4s. Three types of benefit are shown: joint benefit to the overall system ($[\Delta(\mathbf{M},\mathbf{a}) - \Omega(\mathbf{M},\mathbf{a})] + [\delta(\mathbf{M},\mathbf{a}) - \omega(\mathbf{M},\mathbf{a})]$); agent benefit for one of the members of the overall system ($\delta(\mathbf{M},\mathbf{a}) - \omega(\mathbf{M},\mathbf{a})$); society benefit for the overall system ($\Delta(\mathbf{M},\mathbf{a}) - \Omega(\mathbf{M},\mathbf{a})$). These values were chosen because they correspond to the three different types of benefit that could be maximised: joint benefit represents a balance between the individual and the community; agent benefit represents the individualistic stance; and society benefit represents the overall system perspective. The benefits shown are cumulative values. So that the different ranges of benefit in the different types of experiment can be compared, the y-axes are normalised [Note 14].

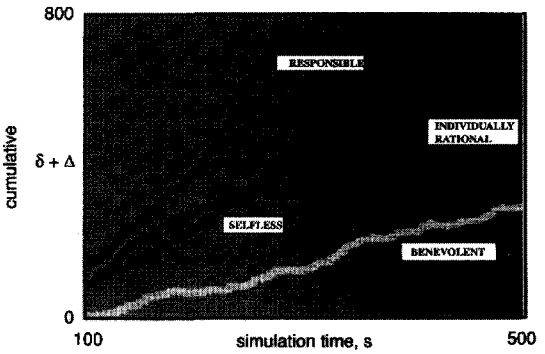

Fig. 2 *Joint benefit*

4.1 Heavy loading

In this set of experiments, the overall system and the agents under investigation were heavily loaded so that they always had goals to process. The member and society benefits were of a comparable scale (i.e. $m = n$). Fig. 2 shows how the joint benefit varies between the four different types of agent society. Responsible societies always obtain the highest joint benefit, indicating a good overall performance. Individually rational and selfless societies perform moderately well, and benevolent societies perform badly. This graph shows that responsible agents obtain the greatest joint benefit (hardly surprising, as this is the measure they are designed to maximise). However, what is perhaps of greater interest is seeing how much better responsible agents are than individually rational agents (one-third better, on average). The selfless and individually rational agents are approximately equal, as member and society benefit are approximately equal. Fig. 3 shows that individually rational agents always obtain the greatest amount of individual benefit; this is because they aim to maximise their member benefit and they are not concerned with the effects of their actions on the rest of the system. Selfless agents perform particularly badly as they are not concerned with their personal benefit, only with the functioning of the overall system. Selfless agents perform worse than benevolent ones because they rarely get the chance to execute

Note 14: The relationship between society and agent benefits can vary. The agent's benefit ranges between $\pm 10^{*}\mathbf{m}$, and the society's benefit ranges between $\pm 10^{*}\mathbf{n}$, where **m** and **n** are positive integers between 1 and 100. When displaying benefits, their limits are normalised by **n+m**.

locally generated goals, whereas our benevolent agents simply execute goals on a first-come, first-served basis, which means that they sometimes execute locally generated goals (which bring them individual benefit). Responsible agents perform reasonably well. The reason for this good showing is that they have the right to exploit the society's resources, up to a point, for their own self benefit. In contrast to Fig. 3, Fig. 4 shows how selfless agents produce high society benefits, whereas individually rational ones produce corresponding low values (the latter never perform any action for the benefit of the society unless it is also in their own interest). As hoped, responsible agents produce a creditable showing. This level of performance is attributable to the duty of responsible agents to assist other members with their problem solving where necessary.

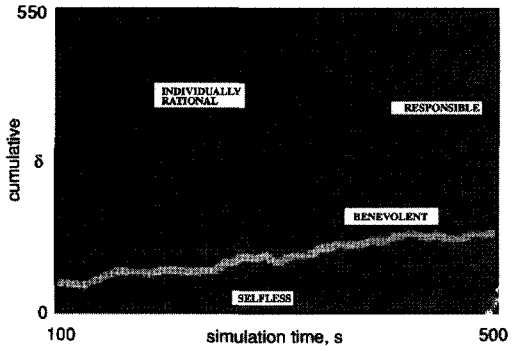

Fig. 3 *Agent benefit*

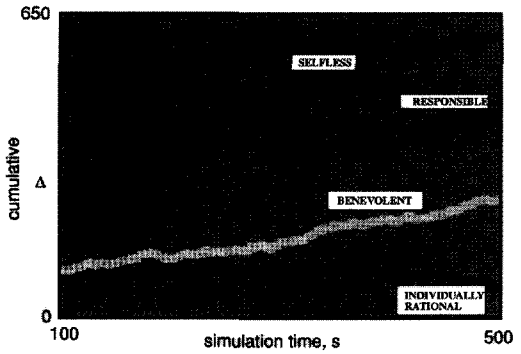

Fig. 4 *Society benefit*

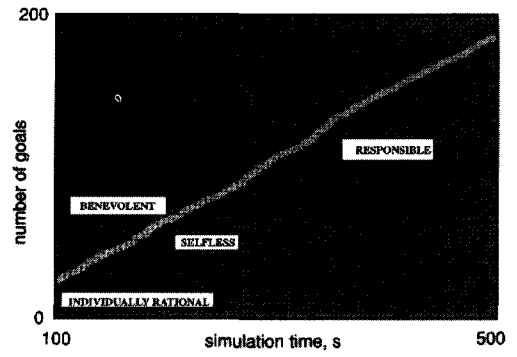

Fig. 5 *Goals completed*

Fig. 5 shows that all the societies complete virtually the same number of goals. This happens because the system is so highly loaded that the agents always have goals to process and are therefore never idle.

4.2 Light loading

In this set of experiments, the system was lightly loaded, and so the agents sometimes had no goals to process. The remaining parameters are the same as in the preceding Section. Fig. 6 again highlights the superior performance of the responsible agents in terms of the system's overall joint benefit. It differs from its Fig. 2 counterpart in three important ways. First, the plot is less smooth because, at times the agents had no goals to process (being idle they accrued no benefit). Secondly, the benevolent agents perform comparatively better (an average of 30% worse than the responsible agent, as opposed to 50% worse in the first set of experiments) as they process more goals (see Fig. 7) than the other types of agent. (Fig. 8 shows how the benevolent agents spend a greater amount of time processing goals than the other types of agent, and thus they need more time, and ultimately more resources, to achieve joint benefits comparable with that of the individually rational or selfless agents). Therefore, in domains with scarce resources, benevolent agents would be a bad choice, even though they can achieve joint benefits comparable to individually rational or selfless agents. Finally, the actual benefit accrued is substantially less (typically half) in magnitude, as far fewer goals are processed. The figures for agent and society benefits are affected in similar ways and so are not shown.

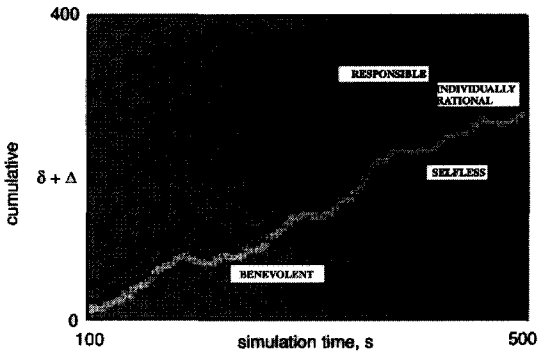

Fig. 6 *Joint benefit*

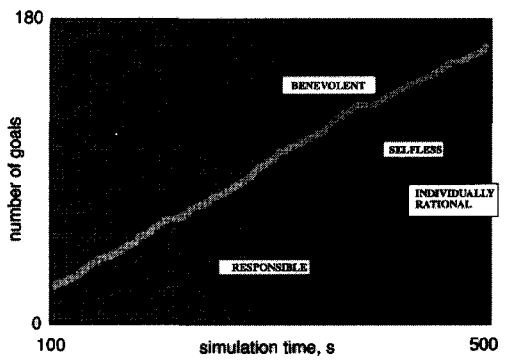

Fig. 7 *Number of goals completed*

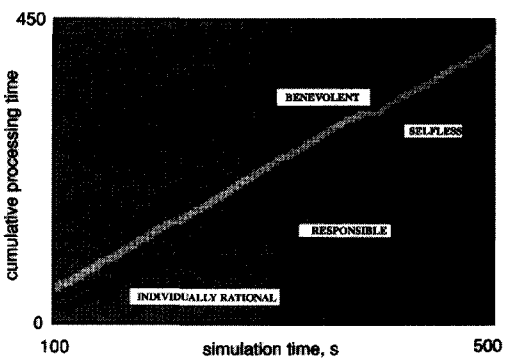

Fig. 8 *Processing time*

4.3 High societal benefit

Having seen the effect of different system and agent loadings, the final set of experiments explore the system's properties when the balance between the benefits to the individuals and to the overall system are varied. In this case, the agents and the system have a medium loading of new goal arrivals. In this Section, the effect of making the society's benefits two orders of magnitude greater than that of the individual is examined: thus society benefits range between ±1000, whereas member benefits range between ±10. These experiments aim to identify how behaviour of the different types of agent and the society vary, when the actions of the individuals are considerably less valuable than those of the society.

Fig. 9 shows how responsible societies still attain the highest joint benefit. However, selfless societies perform almost as well, because joint benefit is virtually equivalent to society benefit, as this amount overwhelms the agent benefit by a factor of 100. For this reason, the graph for society benefit is not shown. Consequently, it can be seen that, as the range of society benefit increases, the behaviour of selfless agents will converge with that of responsible ones. However, it can be shown [84] that selfless societies will never overtake responsible ones. What was suprising in these experiments was how badly the individually rational agents performed. They are so bad because their constituent agents sometimes carried out actions that had a high negative society benefit (up to –1000) because they were not concerned with the effect of their actions on the other agents. This scenario highlights the deficiencies of the individual utility maximising approach as a global problem-solving strategy.

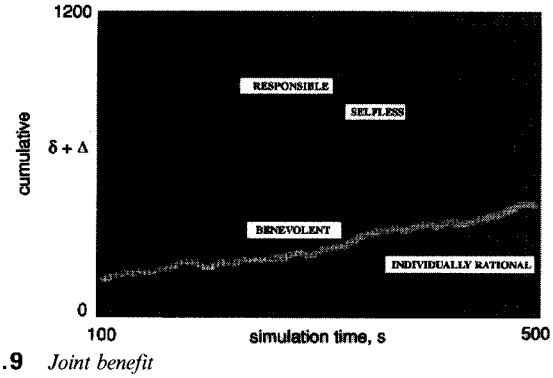

Fig. 9 *Joint benefit*

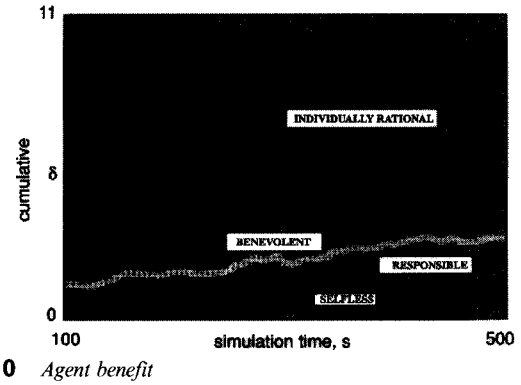

Fig. 10 *Agent benefit*

Fig. 10 clearly highlights how individually rational agents obtain by far the greatest individual benefit. The gap between its performance and that of the other types of agent is larger than in previous experiments. This is because, by allowing society benefit to be 100 times higher than member benefit, the actions of the society dominate those of the individual. All the other types of agent perform poorly: benevolent agents perform better than responsible or selfless ones, because they are less affected by the need to execute dominant societal goals that may have a low individual benefit. Nevertheless, this performance has to be interpreted as the minimum benefit each agent obtains, because its part of the society's benefit is not included in these figures.

4.4 High agent benefit

In this set of experiments, the actions of the individual agents were the dominant factor, ranging between ±1000, whereas the society's benefits ranged between ±10. Fig. 11 again shows the responsible society as the one with the greatest joint benefit. However, the individually rational society performs almost as well in this context, because joint benefit is virtually equivalent to agent benefit, as member benefit is two orders of magnitude greater than society benefit. (For this reason the agent benefit graph is not shown). If the range of agent benefit increases still further, the behaviour of individual rational agents will converge with that of responsible ones. However, it can never overtake the responsible one [84]. Selfless societies perform poorly as they sometimes undertake actions that bring a high negative benefit to their members (up to –1000).

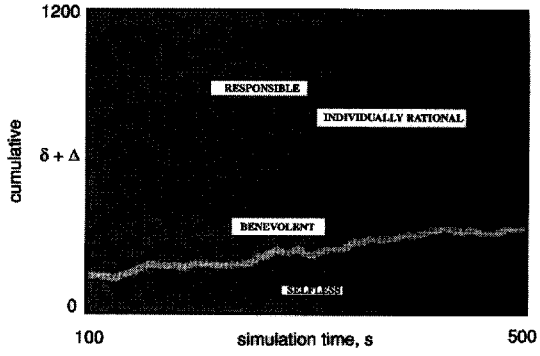

Fig. 11 *Joint benefit*

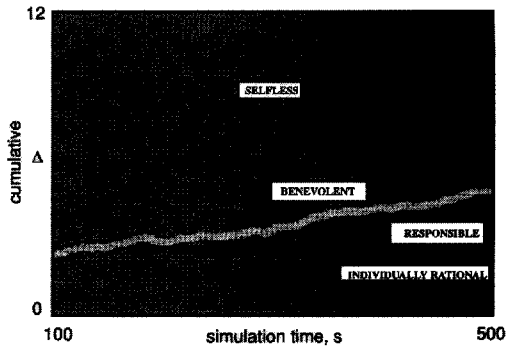

Fig. 12 *Society benefit*

Fig. 12 examines societal benefit and shows that selfless societies perform significantly better than the other types of society. Benevolent agents perform better than responsible or individually rational ones because they process goals on a first-come, first-served basis. This means benevolent agents sometimes execute social goals (with social benefit), whereas responsible and individually rational agents focus almost exclusively on individual goals (with individual benefit), as they are

more profitable. Again, the gap between this performance and that of the other types of society is much greater than in previous experiments, because agent benefit can be 100 times larger than society benefit. The explanations for the poor performances of the benevolent, responsible and individually rational agents are the same as those for the agent benefit in the preceding Section, except that the high-ranking actions relate to the society rather than to the individual member.

4.5 Summary

In all cases, the responsible agents obtained the highest joint benefit. This is true for different degrees of goal loading and for all possible relationships between the utility of individual and social actions. As joint benefit encompasses both individual and societal aspects, this result offers further evidence for the claim that responsible agents attain a good balance between servicing their individual problem solving needs and the needs of the overall system. In terms of the constituent components of joint benefit, responsible agents can be outperformed by individually rational and selfless agents with respect to maximising individual and societal benefit, respectively, although in both cases responsible agents perform creditably. The two situations in which responsible agents performed badly were for individual benefit when the value of the societal actions was high, and for social benefit when the value of individual actions was high. In both cases, the poor showings were due to the fact that the other types of action dominated the responsible agents' choices.

Perhaps the most interesting and surprising sets of results refer to the benevolent and individually rational agents. The former was shown to be the most stable: it was rarely the best or the worst. This is partly owing to the notion of benevolence itself, but also to the fact that it processed goals in a first-come, first-served manner. Hence, it never made any attempt to maximise particular values, but just aimed to give a fair spread. Of even greater interest is the poor performance of the individually rational agents. It has been strongly argued that individual self-interest and utility maximisation are the best selection criteria for software agents. Although this was shown to be true if the performance measure was individual utility maximisation, it is clearly not the case when there is an important and strong notion of society.

5 Related work

When considering related work for something as broad as the SL, we were faced with an overwhelming literature that could validly claim to offer insights on particular facets of this work. A complete analysis of this work would be infeasible in this context, and so this Section is necessarily partial in nature. Moreover, it is biased towards the discipline of computer science in which this work originates This is not to say that disciplines such as biology, game theory, economics, psychology, control theory, philosophy, anthropology and organisational science do not have a bearing on this work, because they surely do, but rather that they are more distantly related. None of these disciplines has as its primary focus the issues that are important for designing and building software systems.

The first place to start looking for related literature is in the field of multi-agent systems itself. The first observation is that there has been comparatively little foundational work that is broadly based. Specific social phenomena have been investigated in depth, e.g. negotiation [24, 85, 86], co-operation [8, 55], conflict resolution [87, 88], commitment [19, 89, 90] and coordination [18, 41, 91], but, in general, these descriptions are not integrated with one another. Hewitt's [9] work on developing open information systems semantics (OISS), which brings together methods from sociology and concurrent systems setence, is perhaps the best current example of a broadly based model. The OISS viewpoint has two primary components: an investigation of the deductive indecision problem [Note 15] and a characterisation of open systems and the nature of problem solving in them. A style of reasoning built from concepts such as trials of strength, commitments and negotiations is what binds these two together. OISS places at its core the fundamental trade-off between developing agents that take effective local actions (i.e. autonomy) and the need for agents to contribute to the performance of the overall aggregate. It exemplifies many of the key problems of building large-scale multi-agent systems, identifies the central modes of interaction and provides a mindset and terminology within which designs and implementations can be constructed. However in terms of identifying foundational principles, OISS has three major shortcomings. First, it fails to have a broad enough scope: not all multi-agent systems are open systems and some OISS assumptions apply to a relatively narrow class of agent [35]. Secondly, it fails to define which goals the agents and the society should pursue by means of trials of strength, commitments and negotiations (i.e. there is nothing akin to the principle of rationality). Finally, it fails to identify the structures that need to be in place for trials of strength to occur.

Within the broader discipline of artificial intelligence, Fox [92] makes a proposal for a new computer level to describe the competence of distributed systems. He terms this level the 'organisational level' and defines it as being above the KL. His reasons for suggesting such a level are very similar to ours and they relate predominantly to the deficiencies of the KL in describing distributed problem-solving systems. However, his characterisation differs from ours in a number of important areas. His system is an organisation (as is ours), his components are agents (this means his knowledge and organisational levels are not distinct as required by Newell), his medium is transactions, his compositional laws are contracts and his behavioural laws are many. He defines a transaction to be a quantisation of knowledge that can be studied independently or in relation to other transactions. By choosing such a broad definition, he fails to identify some of the more detailed components that are present in all social interactions. His compositional law is based on the notion

Note 15: The following explanation is taken from Gasser's [36] response to the OISS proposal. Deductive micro-theories are the primary competing foundation (with OISS) for multi-agent systems. Logical semantics are sufficient for reasoning in closed systems and, hence, can be used as a foundation for reasoning within deductive micro-theories. However, problem solving in open systems involves interacting proposals founded in different micro-theories. Different micro-theories are generated and modified asynchronously and involve different commitments among their participants. Thus logical and representational conflict is endemic to open systems. Logic is not well suited to reasoning in the presence of conflict and therefore is not suited to conflict resolution. Thus deductive micro-theories are insufficient as a foundation for large-scale, open multi-agent systems.

of entities making commitments to one another which is the same as the organisational structures we use. Finally, he gives no real hints about the behavioural law (which should be the key component of such a characterisation), whereas we provide a framework within which behavioural laws can be described and, for a particular class of agent, we define the optimal behavioural law.

Perhaps the largest area of literature relevant to this work is that of rational decision making; see [93] for an excellent overview of this area. The goal of this work is to design the agent's decision-making mechanism so that it does the right thing (i.e. it chooses actions to achieve its goals). In traditional decision theory (or rational choice theory), every rational agent must maximise something [94] and it is usually the expected utility of an action. Thus, decision theory defines a normative theory of belief and action and can be characterised in the following way [80]:

Decision Theory =	Probability Theory and	Utility Theory
what an agent should do	what an agent should do based on evidence	what an agent wants

This basic model underlies much of the work from this field that has found its way into the multi-agent arena, e.g. through negotiation based on game theory [24], through market-oriented programming [95] and through coordination based on recursive modelling [23]. However, from the perspective of multi-agent systems, all of these approaches are based on the idea of methodological solipsism [96]. That is, in the mind of an agent the existence of other agents plays no role. This view may be appropriate when there is no meaningful notion of society; however as we have shown, it does not always lead to good system-wide performance. Indeed Wellman [95] notes that 'combining individual rationality with laws of social interaction provides perhaps the most natural approach to generalising the KL analysis idea to distributed computations'.

A number of researchers have recognised the shortcomings of individual utility maximisation when agents are placed in a social context. Gauthier [97] argues that individual utility maximisation destroys any real possibility of society. He offers an alternative definition appropriate for societal situations: 'a person acts rationally if the expected outcome of his action affords each person with whom his action is inter-dependent a utility such that there is no combination of possible actions, one for each person acting independently, with an expected outcome which affords each person other than himself at least as great a utility, and himself a greater utility'. Although this definition avoids the solipsism of the previous approaches, it fails to capture the notion of doing things for the greater good. Thus it is still based on an agent-centric utility maximisation approach and cannot account for the full range of social actions identified as desirable in Section 3.4. Marschak and Radner [98] also attempt to avoid solipsism in their work on the economic theory of teams. They define a set of decision functions that combines the utilities of the individual team members to give the best set of actions for the team to take. If this combination function is appropriately specified then it will allow agents to perform actions for the greater good. However, their approach requires a centralised controller for the selection of team actions (our work distributes this function) and, hence, erodes the autonomy of the constituent agents.

6 Conclusions and future work

This paper has highlighted the need for a new, distinct computer level to describe social problem-solving behaviour in multi-agent contexts. We believe the SL provides a unifying framework for comparing and analysing all types of multi-agent system. Further evidence for this claim can be found in [15, 99, 100], where the authors used the philosophy of an SL to assist in the design of a wide variety of agent systems. Here, we used the SL framework to study a particular class of system, namely responsible societies. Such societies represent an important and relatively ubiquitous class of multi-agent system. The key facets of our analysis are the principle of social rationality which defines the agents' problem-solving behaviour and the notions of rights and duties, which, respectively, capture the notions of balancing the benefits obtained from interactions with being helpful towards others. When taken together, these concepts lead to agents that are good team players but that, nevertheless, retain their local autonomy.

To be clear about the extent of this work, we are not claiming that responsible behaviour is appropriate to all types of multi-agent system. Rather, we limit our scope to those systems in which both the individuals and the overall system need to perform well. In other cases, benevolent, self interested, or some other strategy may be a more appropriate characterisation of the desired problem solving behaviour.

We can identify three main benefits of an SL analysis. First, it provides an abstract framework in which existing multi-agent systems can be compared. It peels away the implementation- and application-specific details to reveal the central core and the key assumptions of the system under investigation. This is important because it presents the opportunity for the builders of agent-based systems to develop 'conventional wisdom' [10] about design options and trade-offs. Secondly, it provides a high-level model which can assist in the development of future multi-agent systems. In the same way that there are now many KL design methodologies and tools (e.g. [101–103]), it is hoped that the SL will act as a catalyst for similar developments in multi-agent systems. Such developments are essential if multi-agent systems are to move into the mainstream market of solution technologies for complex systems. Finally, by concentrating on core generic functionality, it is hoped that greater re-use of multi-agent software and problem-solving modules can be obtained; see [104–107] for examples of such re-use from KL models. Such re-use is important in that it will facilitate greater robustness and faster development times through the availability of shared libraries.

By its very nature, the SL is neutral about how it is realised in computational systems. Ultimately the approach chosen will depend upon the characteristics of the problem being addressed. Thus the principle of social rationality and the notions of rights and duties may be encoded (either implicitly or explicitly) within the agent architecture, they may be a design guideline that all developers must adhere to, or they may be something that has to be negotiated from scratch for each and every encounter. (Indeed, for rights and duties, they may even vary during an encounter.) At this point it should be re-emphasised that, to implement specific types of social agent, many details not

covered at the SL need to be filled in. This involves specialising the SL constructs for the application under development (e.g. rights, duties, interaction means, compositional laws etc.), defining the KL issues related to individual asocial behaviour (e.g. individual goals and preferences) and adding all the necessary domain-dependent features at the symbol level.

As this work represents a preliminary characterisation of the SL, many open issues still remain. Some of these issues relate specifically to the characterisation of social responsibility:

- What happens if agents fail to obey the principle of social rationality (either through bad design or malicious intent)?
- What types of right and duty are suitable for which types of agent and which types of application?
- How can specific instances of rights and duties be shown to be consistent with the principle of social rationality?
- How can the values of the benefits and losses be computed in an effective manner and in a reasonable time?

Others relate to the SL concept in general:

- What sorts of tool can be devised to support SL modelling?
- What sort of design methodology can be devised to exploit SL descnptions?
- To what degree must the SL constructs be explicitly visible within the multi-agent system?
- How can re-use of designs and analysis be maximised?

7 References

1 CHAIB-DRAA, B.: 'Industrial applications of distributed AI', *Comm. ACM,* 1995, **38**, (11), pp. 47–53
2 JENNINGS, N.R.: 'Cooperation in industrial multi-agent systems' (World Scientific Publishing, 1994)
3 Proc. First Int. Conf. Practical Applications of Intelligent Agents and Multi-Agent Systems (PAAM), London, 1996
4 WOOLDRIDGE, M.J., and JENNINGS, N.R.: 'Intelligent agents: Theory and practice', *Knowledge Engineering Review,* 1995, **10**, (2), pp. 115–152
5 LESSER, V.R., and CORKILL, D.D.: 'The distributed vehicle monitoring testbed: A tool for investigating distributed problem solving networks', *AI Magazine,* 1983, pp. 15–33
6 ERMAN, L.D., and LESSER, V.R.: 'A multi-level organisation for problem solving using many diverse cooperating sources of knowledge'. Proc. Int. Joint Conf. AI, Stanford, CA, 1975, pp. 483–490
7 ROSENSEHEIN, J.S., and GENESERETH, M.R.: 'Deals among rational agents'. Proc. Int. Joint Conf. AI, Stanford, CA, 1985, pp. 91–99
8 GALLIERS, J.R.: 'The positive role of conflict in cooperative multi-agent systems'. Proc. First European Workshop on Modelling an Autonomous Agent in a Multi-Agent World, Cambridge, 1989
9 HEWITT, C.E.: 'Open information systems semantics for distributed artificial intelligence', *Artificial Intelligence,* 1991, **47**, pp. 79–106
10 JENNINGS, N.R., and WOOLDRIDGE, M.J.: 'Applying agent technology', *Int. J. Applied Artificial Intelligence,* 1995, **9**, (4), pp. 351–369
11 SIMON, H.A.: 'Models of man' (Wiley, 1957)
12 WAVISH, P., and GRAHAM, M.: 'A situated action approach to implementing characters in computer games', *Int. J. Applied Artificial Intelligence,* 1996, **10**, (1), pp. 53–73
13 OVERGAARD, L., PETERSEN, H.C., and PERRAM, J.W.: 'Reactive motion planning: a multi-agent approach', *Int. J. Applied Artificial Intelligence,* 1996, **10**, (1), pp. 35–51
14 FERBER, J., and DROGUL, A.: 'Using reactive multi-agent systems in simulation and problem solving' *in* AVOURIS, N.M., and GASSER, L. (Eds.): 'Distributed artificial intellignce: theory and praxis' (Kluwer Academic Publishers, 1992), pp. 53–80
15 FISCHER, K., MÜLLER, J.P., and PISCHEL, M.: 'Cooperative transportation scheduling: an application domain for DAI', *Int. J. Applied Artificial Intelligence,* 1996, **10**, (1), pp. 1–33
16 STEELS, L.: 'Cooperation between distributed agents through self organisation', *J. Robotics and Autonomous Systems,* 1989
17 CASTELFRANCHI, C., MICELI, M., and CESTA, A.: 'Dependence relations among autonomous agents' *in* WERNER, E., and DEMAZEAU, Y. (Eds.): 'Decentralized AI 3' (Elsevier, 1992), pp. 215–227
18 DECKER, K.S.: 'TÆMS: a framework for environment centered analysis and design of coordination mechanisms' *in* O'HARE, G.M.P., and JENNINGS, N.R.(Eds.): 'Foundations of distributed artificial intelligence' (Wiley, 1996), pp. 429–447
19 JENNINGS, N.R.: 'Commitments and conventions: The foundation of coordination in multi-agent systems', *The Knowledge Engineering Review,* 1993, **8**, (3), pp. 223–250
20 LESSER, V.R.: 'A retrospective view of FA/C distributed problem solving', *IEEE Trans. Systems Man and Cybernetics,* 1991, **21**, pp. 1347–1363
21 AXELROD, R.: 'The evolution of cooperation' (Basic Books, 1984)
22 DAWKINS, R.: 'The selfish gene' (Oxford Unviersity Press, 1976)
23 GMYTRASIEWICZ, P.J., and DURFEE, E.H.: 'Elements of a utilitarian theory of knowledge and action'. Proc. Int. Joint Conf. AI, Chamberry, France, 1993, pp. 396–402
24 ROSENSCHEIN, J.S., and ZLOTKIN, G.: 'Rules of encounter' (MIT Press, 1994)
25 GILBERT, N., and DORAN, J.E. (Eds.): 'Simulating societies: the computer simulation of social phenomena' (UCL Press, 1994)
26 GUILFOYLE, C., and WARNER, E.: 'Intelligent agents: the new revolution in software' (Ovum Ltd, 1994)
27 JENNINGS, N.R., CORERA, J., LARESGOITI, I., MAMDANI, E.H., PERRIOLAT, F., SKAREK, P., and VARGA, L.Z.: 'Using ARCHON to develop real-word DAI applications for electricity transportation management and particle accelerator control', *IEEE Expert,* 1996, **2**, (6), pp. 64–70
28 PARUNAK, H.V.D.: 'Applications of distributed artificial intelligence in industry' *in* O'HARE, G.M.P., and JENNINGS, N.R. (Eds.): 'Foundations of distributed artificial intelligence' (Wiley, 1996), pp. 139–164
29 WEIHMAYER, R., and VELTHUIJSEN, H.: 'Applications of distributed AI and cooperative problem solving to telecommunications' *in* LIEBOWITZ, J., and PREREAU, D. (Eds.): 'AI approaches to telecommunications and network management' (IOS Press, 1994)
30 RAO, A.S., and GEORGEFF, M.P.: 'BDI agents: From theory to practice'. Proc. First Int. Conf. Multi-Agent Systems, San Francisco, CA, 1995, pp. 312–319
31 LESSER, V.R., and CORKILL, D.D.: 'Functionally accurate, cooperative distributed systems', *IEEE Trans. Systems Man and Cybernetics,* 1981, **21**, pp. 1347–1363
32 COEN, M.: 'SodaBot: A software agent environment and construction system'. Proc CIKIM Workshop on Intelligent Information Agents, Gaithersburg, USA, 1994, pp. 18–32
33 ETZIONI, O., and WELD, D.: 'A softbot based interface to the internet', *Comm. ACM,* 1994, **37**, (7), pp. 72–76
34 MAES, P.: 'Agents that reduce work and information overload', *Comms. ACM,* 1994, **37**, (7), pp. 31–40
35 WAVISH, P.: 'Exploiting emergent behaviour in multi-agent systems', *in* WERNER, E., and DEMAZEAU, Y. (Eds.): Decentralised AI, (North Holland, 1992), Vol. 3, pp. 297–310
36 GASSER, L.: 'Social conceptions of knowledge and action: DAI foundations and open system semantics', *Artificial Intelligence,* 1991, **47**, pp. 107–138
37 DURFEE, E.H., LESSER, V.R., and CORKILL, D.D.: 'Trends in cooperative distributed problem solving', *IEEE Trans. Knowledge and Data Engineering,* 1989, **1**, (1), pp. 63–83
38 GASSER, L., ROUQUETTE, N., HILL, R.W., and LIEB, J.: 'Representing and using organisational knowledge in DAI systems' *in* GASSER, L., and HUHNS, M.N. (Eds.): 'Distributed artificial intelligence, vol II' (Pitman Press, 1989), pp. 55–78
39 SHOHAM, Y., and TENNENHOLTZ, M.: 'On the synthesis of useful social laws for artificial agent societies'. Proc. 10th National Conf. AI, San Jose, USA, 1992, pp. 276–280
40 WERNER, E.: 'Cooperating agents: a unified theory of communication and social structure' *in* GASSER, L., and HUHNS, M.N. (Eds.): 'Distributed artificial intelligence, vol II' (Pitman Press, 1989), pp. 3–36
41 DURFEE, E.H.: 'Coordination of distributed problem solvers' (Kluwer Academic Publishers, 1988)
42 GASSER, L.: 'DAI approaches to coordination' *in* WERNER, E., and DEMAZEAU, Y. (Eds.): 'Decentralized AI 3' (Elsevier, 1992), pp. 31–51
43 CAMMARATA, S., MCARTHUR, D., and STEEB, R.: 'Strategies of cooperation in distributed problem solving'. Proc. Int. Joint Conf. AI, Karlsruhe, Germany, 1983, pp. 767–770
44 CORKILL, D.D.: 'Hierarchical planning in a distributed environment'. Proc. Sixth Int. Joint Conf. AI, Cambridge, MA, 1979, pp. 168–175
45 GEORGEFF, M.P.: 'Communication and action in multi-agent planning'. Proc. National Conf. AI, Washington, DC, 1983, pp. 125–129
46 NEWELL, A.: 'The knowledge level', *Artificial Intelligence,* 1982, **18**, pp. 87–127
47 ASIMOV, I.: 'The rest of the robots' (Grafton Books, 1968)

48 CASTELFRANCHI, C., and CONTE, R.: 'Distributed artificial intelligence and social science: critical issues' *in* O'HARE, G.M.P., and JENNINGS, N.R.(Eds.): 'Foundations of distributed artificial intelligence' (Wiley, 1996), pp. 527–542
49 GERSON, E.M.: 'On quality of life', *American Sociological Review,* 1976, **41**, pp. 793–806
50 PRUITT, D.G., and CARNEVALE, P.J.: 'Negotiation in social conflict' (Open University Press, 1993)
51 JENNINGS, N.R.: 'Towards a cooperation knowledge level for collaborative problem solving'. Proc. 10th European Conf. AI, Vienna, Austria, 1992, pp. 224–228
52 NEWELL, A.: 'Reflections on the knowledge level', *Artificial Intelligence,* 1993, **59**, pp. 31–38
53 AITKEN, J.S., SCHMALHOFER, F., and SHADBOLT, N.: 'A knowledge level characterisation of multi-agent systems' *in* WOOLDRIDGE, M.J., and JENNINGS, N.R. (Eds.): 'Intelligent agents' (Springer–Verlag, 1995), pp. 179–190
54 SEARLE, J.R.: 'Collective intentions and actions' *in* COHEN, P.R., MORGAN, J., and POLLACK, M.E.(Eds.): 'Intentions in communications' (MIT Press, 1990), pp. 401–416
55 BRATMAN, M.E.: 'Shared cooperative activity', *Philos. Rev.,* 1992, **101**, (2), pp. 327–341
56 JENNINGS, N.R.: 'Controlling cooperative problem solving in industrial multi-agent systems using joint intentions', *Artificial Intelligence,* 1995, **75**, (2), pp. 195–240
57 COHEN, P.R., and LEVESQUE, H.J.: 'Teamwork', *Noûs,* 1991, **25**, (4), pp. 487–512
58 MEAD, G.H.: 'Mind, self and society' (University of Chicago Press, 1934)
59 SEELEY, T.D.: 'The honey bee colony as a superorganism', *American Scientist,* 1989, **77**, pp. 546–553
60 MINSKY, M.: 'The society of mind' (Simon and Schuster, 1985)
61 GASSER, L.: 'Boundaries, identity and aggregation: plurality issues in multi-agent systems' *in* WERNER, E., and DEMAZEAU, Y. (Eds.): 'Decentralized AI 3' (Elsevier, 1992), pp. 199–214
62 HALPERN, J.Y., and MOSES, Y.O.: 'Knowledge and common knowledge in a distributed environment'. Proc. 3rd ACM Conf. principles of Distributed Computing, 1984, pp. 50–61
63 NEWELL, A.: 'Unified theories of cognition' (Harvard Unviersty Press, 1990)
64 CARLEY, K., and NEWELL, A.: 'The nature of the social agent'. Technical report, Department of Social and Decision Sciences, Carnegie Mellon University, 1992
65 BATES, J.: 'The role of emotion in believable agents', *Comm. ACM,* 1994, **37**, (7), pp. 122–125
66 RUSSELL, S., and WEFALD, E.: 'Do the right thing' (MIT Press, 1991)
67 ZILBERSTEIN, S.: 'Models of bounded rationality'. Proc AAAI-95 Fall Symposium Series Rational Agency: Concepts, Theories, Models and Applications, 1995
68 RUMBAUGH, J., BLAHA, M., PREMERLANI, W., EDDY, F., and LORESNSEN, W.: 'Object-oriented modelling and designing' (Prentice–Hall, 1991)
69 JENNINGS, N.R., FARATIN, P., JOHNSON, M.J., NORMAN, T.J., O'BRIEN, P., and WIEGAND, M.E.: 'Agent-based business process management', *Int. J. Cooperative Info. Syst.,* 1996, **5**, (2&3), pp. 105–130
70 RAO, A.S., GEORGEFF, M.P., and SONENBERG, E.A.: 'Social plans: a preliminary report' *in* WERNER, E., and DEMAZEAU, Y. (Eds.): 'Decentralized AI 3' (Elsevier, 1992), pp. 57–76
71 GENESERETH, M.R., and KETCHPEL, S.P.: 'Software agents', *Comm. ACM,* 1994, **37**, (7), pp. 48–53
72 ENGLEMORE, R., and MORGAN, T. (Eds.): 'Blackboard systems' (Addison–Wesley, 1988)
73 BROOKS, R.: 'Intelligence without representation', *Artificial Intelligence,* 1991, **47**, pp. 139–159
74 GLANCE, S.N., and HUBERMAN, B.A.: 'Organizational fluidity and sustainable cooperation'. Technical report 94304, Xerox Palo Alto, Research Center, Palo Alto, CA, 1993
75 GALBRAITH, J.: 'Designing complex organizations' (Addison–Wesley, 1973)
76 MALONE, T.W.: 'Modelling coordination in organizations and markets', *Management Science,* 1987, **33**, pp. 1317–1332
77 ISHIDA, T., YOKOO, M., and GASSER, L.: 'An organisational approach to adaptive production systems'. Proc. 8th National Conf. AI, Boston, USA, 1990, pp. 52–58
78 HARSANYI, J.: 'Bayesian decision theory and utilitarian ethics', *American Economy Review,* 1978, **68**, pp. 223–228
79 JEFFREY, R.: 'The logic of decision' (University of Chicago Press, 1983)
80 RUSSELL, S., and NORVIG, P.: 'Artificial intelligence: a modern approach' (Prentice–Hall, 1995)
81 CASTELFRANCHI, C.: 'Social power: a point missed in multi-agent, DAI and HCI' *in* DEMAZEAU, Y., and MÜLLER, J.P. (Eds.): 'Decentralized AI' (Elsevier, 1990), pp. 49–62
82 GILBERT, M.: 'Walking together: A paradigmatic social phenomenon', *Midwest Studies Philosophy,* 1990, **XV**, pp. 1–14
83 GEIST, A., BEGUELIN, A., DONGARRA, J., JIANG, W., MANCHECK, R., and SUNDERAM, V.: 'PVM3 user's guide and reference manual'. ORNL/TM-12187, May 1994
84 JENNINGS, N.R., and CAMPOS, J.R.: 'Characterising socially responsible agents'. Technical report, Department of Electronic Engineering, Queen Mary & Westfield College, 1995
85 LAASRI, B., LAASRI, H., LANDER, S., and LESSER, V.R.: 'A generic model for intelligent negotiating agents', *Int. J. Intelligent and Cooperative Info. Syst.,* 1992, **1**, (2), pp. 291–318
86 MÜLLER, H.J.: 'Negotiation principles' *in* O'HARE, G.M.P., and JENNINGS, N.R. (Eds.): 'Foundations of distributed artificial intelligence' (Wiley, 1996), pp. 211–229
87 ADLER, A.B., DAVIS, R., WIEHMAYER, R., and FORREST, F.W.: 'Conflict resolution strategies for nonhierarchical distributed agents' *in* GASSER, L., and HUHNS, M.N. (Eds.): 'Distributed artificial intelligence, vol II' (Pitman Press, 1989), pp. 139–192
88 KLEIN, M.: 'Supporting conflict resolution in cooperative design systems', *IEEE Trans. Systems Man and Cybernetics,* 1991, **21**, (6), pp. 1379–1390
89 BRATMAN, M.E.: 'Intention, plans and practical reason' (Harvard University Press, 1987)
90 COHEN, P.R., and LEVESQUE, H.J.: 'Intention is choice with commitment', *Artificial Intelligence,* 1990, **42**, pp. 213–261
91 VON MARTIAL, F.: 'Coordinating plans of autonomous agents' (Springer–Verlag, 1992)
92 FOX, M.S.: 'Beyond the knowledge level' *in* KERSCHBERG, L. (Ed.): 'Expert database systems' (Cummings Publishing Company, 1987), pp. 455–463
93 DOYLE, J.: 'Rationality and its roles in reasoning', *Computational Intelligence,* 1992, **8**, (2), pp. 376–409
94 LUCE, R.D., and RAFFIA, H.: 'Games and decisions' (Dover, 1957)
95 WELLMAN, M.P.: 'A market-oriented programming environment and its application to distributed multi-commodity flow problems', *J. Artificial Intelligence Research,* 1993, **1**, pp. 1–23
96 SMIT, R.A., and VERHAGEN, H.J.E.: 'On being social: degrees of sociality and models of rationality in relation to multi-agent systems'. Proc. AAAI-95 Fall Symposium Series Rational Agency: Concepts, Theories, Models and Applications, 1995
97 GAUTHIER, D.: 'Reason and maximisation', *Canadian J. Philosophy,* 1975, **4**, pp. 418–433
98 MARSCHAK, J., and RADNER, R.: 'Economic theory of teams' (Yale University Press, 1972)
99 MASON, C.L.: 'Cooperative seismic data interpretation for nuclear test ban treaty verification', *Int. J. Applied AI,* 1995, **9**, (4), pp. 371–400
100 BARBUCEANU, M., and FOX, M.S.: 'Capturing and modelling coordination in multi-agent systems', *J. Cooperative Inf. Syst.,* 1996, (to appear)
101 DE GREEF, P., and BREUKER, J.A.: 'Analysing system-user cooperation in KADS', *Knowledge Acquisition,* 1992, **4**, pp. 89–108
102 HICKMAN, F.R., KILLIN, J.L., LAND, L., MULHALL, T., PORTER, D., and TAYLOR, R.M.: 'Analysis for knowledge based systems: a practical guide to the KADS methodology' (Ellis Horwood, 1989)
103 WIELINGA, B.J., SCHREIBER, A.T., and BREUKER, J.A.: 'KADS: a modelling approach to knowledge engineering', *Knowledge Acquisition,* 1992, **4**, pp. 5–53
104 CHANDRASEKARAN, B.: 'Generic tasks in knowledge based reasoning: High level building blocks for expert system design', *IEEE Expert,* 1983, **1**, (3), pp. 23–30
105 CLANCY, W.J.: 'Heuristic classification', *Artificial Intelligence,* 1985, **27**, (3), pp. 289–350
106 HOROWITZ, E., and MUNSEN, J.B.: 'An expansive view of reusable software', *IEEE Trans. Software Eng.,* 1984, **10**, (5), pp. 477–487
107 MCDERMOTT, J.: 'A taxonomy of problem solving methods' *in* MARCUS, S. (Ed.): 'Automating knowledge acquisition for expert systems' (Kluwer Academic Press, 1988), pp. 225–256