Socially Rational Agents - Some Preliminary Thoughts

Lisa M. Hogg and Nick R. Jennings
Department of Electronic Engineering
Queen Mary & Westfield College, University of London.
{L.M.Hogg, N.R.Jennings}@qmw.ac.uk

Abstract

Rationality relates to making the right decisions and producing successful behaviour. The notion of building rational agents is one of the main aims of research into artificial intelligence. The current predominant approach is to develop agents which follow a decision theoretic notion of rationality in which the agent maximizes the expected utility of its actions. Although this is intuitively and formally appealing, it lacks applicability in real systems where the agent is faced with resource limitations. Furthermore, this view fails to adequately cope with the case in which an agent is embedded within a system of interacting agents. In such an environment, the agent has the further consideration of the effects of its actions on other agents and the effect of their actions on itself. Therefore to operate effectively in such environments the agents need a principle of social rationality. In this paper we outline our preliminary thoughts on devising such a principle and indicate how it helps the agent strike a balance between its individual needs and the needs of the overall system.

1. Theories of rationality

Rationality is all about doing the 'right' thing, where right equates to performing successful actions [1]. Agent rationality is concerned with intelligent decision making as defined by the mapping of percepts into actions. Thus determining whether an agent is behaving rationally can only be ascertained by examining the information that it possesses and by evaluating the success of the actions that it performs based upon this information. The predominant theory of rational decision making in agents is that of the economic principle of maximizing the expected gain of actions [2]. Decision theoretic rationality dictates that the agent should choose an action which will maximize the expected utility of performing that action given the probability of reaching a desired state in the world and the desirability of that state. However, this theory makes the assumption that the agent has both complete information and sufficient time to carry out the necessary reasoning. In reality, however, agents have limitations on their deliberation with regard to the resources they have available. Hence theories of decision making need to take such boundedness into consideration if the agents are to be applied in real systems. Recognising this shortcoming, work on bounded rationality [3], [4], [5] takes into consideration the fact that the agent is faced with resource limitations which affect its reasoning. Despite numerous attempts at overcoming the problems of resource bounds on reasoning, there is yet to emerge a definitive concept of bounded rationality which can be applied in real systems. In addition, most work on ideal and bounded rationality fails to recognize the importance of the fact that many systems are composed of several interacting agents and that the decisions that each agent makes have consequences on the others within the system. To this end, this work proposes an alternative view of rationality, which takes these factors into consideration. It also explores how resource limitations further affect the agents reasoning in this context.

2. Shortcomings of individual rationality in multi-agent systems

The successful combination of several autonomous, intelligent agents working together is the aim of research into multi-agent systems [6], [7]. Increasingly, the multi-agent paradigm is being used to build real, complex systems [8], [9]. Reasons for this include the inherent natural distribution of problem components and the maturing of distributed computing technology. In such systems, the agents are interdependent, due to resource limitations and problem interdependencies, andl need to interact with one another in order to achieve their goals. For example, due to lack of knowledge or problem solving capability, an agent may need to obtain the assistance from other agents to help it achieve a goal. To do this, an agent may use its knowledge about how other agents are dependent on certain resources in order to influence others to adopt one or more of its goals [10]. An example scenario would be the case where there are two transporter agents, each responsible for the delivery of some goods. One agent has a truck full of its goods, which it cannot empty alone. A second agent needs an empty truck to transport its goods to its customer. Given that the second agent needs an empty truck, the first agent can suggest that the second help it unload its truck, and then use the truck to deliver its goods. In another situation of interdependence, an agent may decide to work together as a team with others, as a necessary or more profitable means of achieving individual or system goals. Figure 1. displays the main decision making components and reasoning complexity of a social agent. Given the situation the agent finds itself in, it faces a number of choices which control its actions. Not only must the agent decide the benefit of combining forces with other agents to work more efficiently, but also it must determine if it acts alone how it can produce most benefit from its actions balancing individual and societal needs. Agents which merely follow the individual perspective of rationality only perform actions which bring themselves the most benefit. In multi-agent system research however, the design objective is to produce successful behaviour at both the individual and the system level.

We believe rationality needs to be considered not only from the individuals point of view, but also from the societal perspective. An agent adopting a selfish strategy may inhibit the achievement of system/social goals. For example if one agent has sole control over a resource which is required by others, then by selfishly using all of this resource itself it can inhibit other agents achieving their goals and hence the system producing the desired behaviour. It may even be the case that decisions taken from a more social perspective actually produce greater benefit for the individual than taking the individualistic point of view. Decision making at a societal level can thus be seen to be composed of a multitude of factors including current situation and individual and global utility considerations. To produce a socially rational choice the agent needs to be able to determine the effect of its actions on others by estimating the benefit that a course of action would provide. In order to do this, the agent needs to know the goals and the preferences of the other agents in the society. Using this information, the agent can determine how desirable the outcome of its actions are to others and hence make decisions which are more socially acceptable.

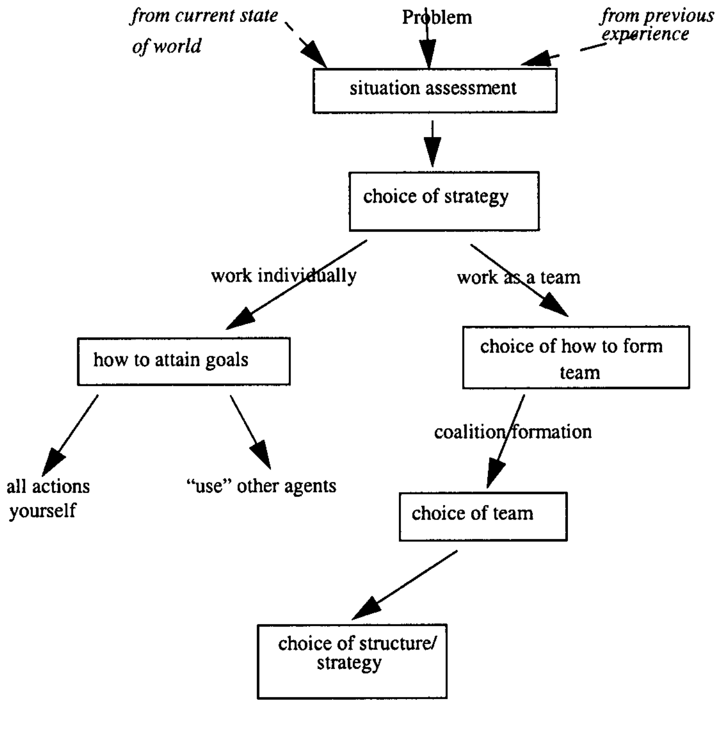

Figure 1: Flow of decision making in a socially rational agent

3. Towards social rationality

For agents situated in a social context, it is clear that they require a fundamental decision making principle to help guide their behaviour in the same way that individual utility maximisation works for asocial agents. This principle should take into consideration both resource bounds (to be practical) and task and social interdependencies (to interact effectively). Such a principle of social rationality provides the agent with a normative theory of decision making within a multi-agent context. The social agent, of figure 1, is faced with many decisions before taking action and we believe an adequately formulated principle of social rationality would assist at all its choice points. A preliminary attempt by Jennings and Campos [11] to define social rationality using individual and global benefits is as follows:

> ***Principle of Social Rationality:*** *If a member of a responsible society*[1] *can perform an action whose joint benefit is greater than its joint loss, then it may select that action.*

As with Newell's rationality hypothesis of the knowledge level [12], this principle conveys, at an abstract level, a normative theory of decision making which takes into account the benefits accrued from actions to the society that the agent inhabits. Joint benefit is defined as the benefit provided to the individual plus the benefit afforded to society as a result of an action. Similarly, joint loss is the individual plus societal loss of performing an action. Although this definition focuses the agent into choosing more beneficial actions from the societal viewpoint, it lacks concrete guidance in the choice of alternatives and the concept of maintaining a balance between individual and system needs. Thus, in order to be applied in real systems the definition needs to be expanded to show how the agent chooses between actions and how the situation that the agent finds itself in effects its decision.

As stated previously, the decision theoretic concept of maximizing the utility of an action is an intuitive and appealing way of conceptualizing decision making. It provides a way of evaluating a set of alternatives by considering the expected utility of each alternative in that set. Jennings and Campos' principle of social rationality, can be mapped onto a utility based definition in the following way. Consider a society of agents $S = \{A_1, A_2, \ldots, A_n\}$, and let B_{Ai} be the benefit afforded to agent A_i by the action a and B_S the benefit afforded to the society. Similarly, let L_{Ai} be the loss (or cost) the action produces for A_i and L_S the loss afforded to society. Hence the expected utility, $EU_{Ai}(a)$, of an action a to agent A_i is

$$EU_{Ai}(a) = (B_{Ai} - L_{Ai}) + (B_S - L_S) \qquad \equiv \qquad EU(a) = \mathbf{f}\,(EIU(a), ESU(a))$$

where EIU is the ***Expected Individual Utility,*** and is the standard decision theoretic notion of the expected utility of an action[2]. ESU is the *Expected Societal Utility* and is the expected benefit that an action a produces to the society. Function f shows the relationship of the two utilities from the agents' perspective and ultimately defines the characteristics of the agent. A social agent would use a function which places a greater weighting on the social utility of actions, while a self-interested agent would place more emphasis on the individual utility of actions.

As a first approximation, f might be defined to be the weighted addition of the individual and societal utilities:

$$EU_{Ai}(a) = w_1.EIU(a) + w_2.ESU(a)$$

where w_1 and w_2 are in the range [0,1] and are the importance the agent places on the respective utility functions.

Given a relatively large society, an agent is likely to find itself interacting with various other individual agents, and in some cases groups of agents, at different times. For example, if an agent does not have the necessary resources to carry out a task it may enlist the help of another agent, or agents, in the society to help it. Work on coalition formation investigates how agents calculate whether working

[1] A responsible society is defined as a system of autonomous agents which balance individual needs with those of the overall system when making decisions. This is equivalent to our social agent.

[2] $EIU(a) = \sum P(s, a)U(s)$. $P(s,a)$ is the probability of reaching a state s after performing action a, and $U(s)$ is the utility of that state to the agent

together as a team is profitable or not [13], [14]. Along similar lines, an agent may wish to perform actions which are potentially helpful to the others who it is interacting with (be it for selfish or social motives). An example of this would be if an agent thought that by performing an action which was beneficial to others now, it may prompt others to favour (help) it in the future. It is therefore possible to further distinguish between the agent's commitment to achieving benefit for the 'society' in general and the commitment to assisting those agents or groups with which it is currently interacting. Adding this consideration to the above equation we obtain:

$$EU_{Ai}(a) = w_1.EIU(a) + w_2.EPU(a) + w_3.ESU(a)$$

where EPU is the ***Expected Partners Utility*** (partners being those with whom the agent is currently interacting) and ESU is the benefit to the general society (this may be in terms of following social norms and conventions, or level of achievement of social goals):

$$EPU(a) = \sum_{A_p \in S} EU_{Ap}(a) \quad \text{and} \quad ESU(a) = \sum_{A_s \in \{S-A_P\}} EU_{As}(a)$$

An agent may not necessarily interact with the same agents all of the time. Hence the notion of partners is a dynamic concept, with the agent likely to interact with different sets of agents to varying degrees. From the previous equation, the utility to society is thus divided into utility of agents with whom the agent has some form of relationship (i.e. interacting and may be dependent in some way on each other) and the utility to the society in general (e.g. doing something for the common good). By making this distinction, agents can identify actions which are rational given the small community within the society that the agent is interacting with.

In deliberating in this manner, an agent can exhibit more socially rational behaviour in terms of the small groups of agents which it finds itself interacting with on a regular basis. However, consideration of the wider social context can be seen as an extra bound on an agent in the same vein as the resource limitations mentioned in section one. It takes time and resources to calculate the utility an action affords to others. Resources which agents may not always have. In such cases, it would be advantageous to have a flexible control mechanism which, allows the agent to tailor the amount of social reasoning it performs to its current situation. Thus, in times of heavy resource constraints where it may be impractical to take the effect of actions on others into consideration, the agent will make decisions based solely on individual benefits. In situations where it has more resources, it may be able to include the consideration of individual and partners utility for choosing action and in the most plentiful scenario, may use full social rationality. By having this flexibility the agent can manage its goal priorities (i.e individual and social goals) more efficiently and hence make its decision making more socially rational as its resource bounds increase. Additionally, if the agent has the ability to learn from previous decision making successes and failures, there is the potential that it could converge towards finding the right balance, depending on the situation, between individual and societal needs.

4. Conclusions and future aims

Rationality is a desirable property of agents since it provides a means of assessing and attaining intelligent behaviour. Previous work on rationality has concentrated on the benefits gained from making a decision based on an individualistic, selfish perspective. Given the increasing use of the multi-agent paradigm in tackling complex problems, some other principle of rationality needs to be considered which would guide the agent's actions within a multi-agent environment. Social rationality provides a principle by which agents make decisions which strike a balance between the needs of the system/ society and those of the individual agent. However, such a theory needs to also take into consideration the fact that the agent is bounded and that being 'social' is a limitation as well other considerations such as time and computational power. The ultimate aim of this research is to define a principle of social rationality which can be used to guide an agent's decisions making in realistic, multi-agent environments, and this paper represents a preliminary step towards this goal. The robustness of the principle under varying constraints, will also be explored and this will lead to the development of socially bounded rational agents.

5. References

[1] S. Russell, *Rationality and Intelligence*, 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, August: pp 950-957, 1995.

[2] J. Doyle, *Rationality and its Roles in Reasoning*, Computational Intelligence, Volume 8 No. 3, 1992.

[3] H. A. Simon, *A Behavioural Model of Rational Choice,* Quarterly journal of Economics, 69: pp99-118, 1955.

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

[5] E.Horvitz, *Reasoning Under Varying and Uncertain Resource Constraints*, Proceedings of the Seventh National Conference on AI, Minneapolis, August: pp111-116,1988.

[6] A.H. Bond and L. Gasser, *Readings in DAI*, Morgan Kauffman, San Mateo, California, 1988.

[7] M. Huhns, *Distributed Artificial Intelligence*, Pittman, 1989.

[8] R. Weihmayer and H. Velthuijsen, *Applications of distributed AI and cooperative problem solving to telecommunications*, in J. Liebowitz and D. Prereau (eds), Ai Approahces to telecommunications and network management, IOS Press, 1994.

[9] H.V.D. Parunak, *Applications of distribuited artificial intelligence in industry*, in G.M.P O'Hare and N.R. Jennings (eds), Foundations of Distributed Artificial IntelligenceWiley: pp139-164, 1996.

[10] C. Castelfranchi, *Social Power: A Point Missed in Multi-Agent, DAI and HCI,* Decentralized A.I., Yves Demazeau & Jean-Pierre Muller eds., Elsevier Science Publishers, B.V (North Holland), 1990.

[11] N.R. Jennings & J. Campos, *Towards a Social Level Characterization of Socially Responsible Agents*, IEE Proceedings on Software Engineering: pp 11-25, 1997.

[12] A. Newell, *The Knowledge Level*, Artificial intelligence, 18: pp87-127, 1982.

[13] A. Shehory and S. Kraus, *Task Allocation via Coalition formation among autonomous agents, 14th International Joint Conference on Artificial Intelligence, Montreal, Canada, August1995.*

[14] S. Ketchpel, *Coalition Formation Amongst Autonomous Agents,* in: Castelfranci Cristiano and Muller J-P (eds.), : From Reaction to Cognition, 5th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW '93, Neuchatel, Switzerland, August 25-27, 1993., Selected papers, Lecture Notes in Artificial Intelligence 957, Springer Verlag, Berling, Heidelberg, 1995, 73-88.