Socially Rational Agents

L. M. Hogg and N. R. Jennings
Department of Electronic Engineering, Queen Mary & Westfield College,
University of London, London E1 4NS, UK.
{L.M.Hogg, N.R.Jennings}@qmw.ac.uk

Abstract

Autonomous agents are designed to carry out problem solving actions in order to achieve given, or self generated, goals. A central aspect of this design is the agent's decision making function which determines the right actions to perform in a given situation to best achieve the agent's objectives. Traditionally, this function has been solipsistic in nature and based upon the principle of individual utility maximisation. However we believe that when designing multi-agent systems this may not be the most appropriate choice. Rather we advocate a more social view of rationality which strikes a balance between the needs of the individual and those of the overall system. To this end, we describe a preliminary formulation of social rationality, indicate how the definition can vary depending on resource bounds, and illustrate its use in a fire-fighting scenario.

1. The Case for Social Rationality

Rational agents make decisions about which actions to perform at what times in order to best achieve their goals and objectives. The exact nature and the underpinning principles of an agent's decision making function have been studied in a range of disciplines including philosophy, economics and sociology. This work has demonstrated that the design of the decision making function is the critical determinant of the success of the agent [Doyle 83]. The current predominant view is to equate rational decision making with that of maximizing the expected utility of actions as dictated by decision theory [Horvitz et al. 88] (although see [Castelfranchi and Conte 97] for a critique of this position).

This utilitarian view of rationality may be adequate when all that is being considered is the design and success of a single agent. However when designing a system in which multiple agents need to interact (cooperate and coordinate) in order to achieve *both individual and system goals*, we feel that such a solipsistic view may not be the most appropriate. For the moment we limit our claim to the case in which the designer wishes to build a (complex) system using autonomous agents[1]. Relevant examples include process control systems, telecommunications management systems, business process management systems, air traffic control systems, and manufacturing systems. Thus we are not talking about pure distributed problem solving systems nor pure multi-agent systems. In the former case, the sole concern of the designer is with the overall performance of the system and the performance of the individual agents is not important. In the latter case, the concern of the designer is with the performance of the individual agents and the system level performance is left to emerge out of the interplay between the constituent components. Rather we are concerned with a hybrid case in which the designer wishes to exploit the conceptual power of autonomous agents (as in the multi-agent systems view), but wishes to achieve system level objectives (as in the distributed problem solving case). In this hybrid context, we feel that a view of rationality which considers, and attempts to strike a balance between, both individual and system level goals is both more natural and more likely to succeed[2] (cf. the view of the market-based computing community [Wellman 93]).

This more social view of rationality means that an agent has to consider the implications of its choices on other, and sometimes all, agents in the system. That is, agents need to be given a social perspective to aid their decision making and improve their performance [Castelfranchi 90]. Several examples of such social decision making functions have been identified [Cesta et al. 96; Kalenka and Jennings 97]; however we feel that the conceptual foundations of these functions need to be analysed in greater depth. To this end, we start from the following decision making principle [Jennings and Campos 97]:

> ***Principle of Social Rationality:*** *If a socially rational agent can perform an action whose joint benefit is greater than its joint loss, then it may select that action.*

Joint benefit is a combined measure which incorporates the benefit provided to the individual and the benefit afforded to the overall system as a result of an action (*mutatis mutandis* for joint loss). Although this definition focuses the agent into choosing more beneficial actions from the societal viewpoint, it does not provide concrete guidance in the choice of alternatives nor does it provide a framework for maintaining a balance between the individual and the system needs. Thus, to be applied practically, the definition needs to be expanded in these directions.

Due to its intuitive and formal treatment of making decisions from a set of alternatives under uncertainty, we will use the notion of expected utility of actions in deriving a more descriptive notion of choice within a multi-agent environment. From the aforementioned principle of social rationality, to calculate the expected utility (EU) of an action α, an agent needs to combine (using some function f) the

[1] Such systems are closed in the sense that the designer knows precisely which agents are part of the system and that he has control over their decision making functions. Hence exploitation by outside agents is not a concern.

[2] Whilst it would be possible to manipulate the agent's individual utility (or goals) so that it incorporates a measure of social awareness, this would simply be hiding the underlying social principles behind the numbers.

individual utility (IU) afforded to the agent which performs α and the social utility (SU) afforded to the overall system when α is executed:

$$EU(\alpha) = f\,(IU(\alpha), SU(\alpha)) \qquad \text{(equation 1)}$$

The exact nature of the combination function determines the characteristics of the agent: a selfish agent places more emphasis on its individual utility, an altruistic agent places greater emphasis on its social utility, and a socially rational agent tries to strike a balance between the two.

The calculation of SU can be further distinguished by differentiating between the different social relationships in which the agent is engaged. Thus, for example, a particular agent may work in a team with a small number of agents, a loose confederation with a larger number of agents and hardly at all with the remaining agents. Let the set of social relationships in which a particular agent (a) is engaged be denoted by λ_a ($\lambda_{a,1}$, $\lambda_{a,2}$,.... $\lambda_{a,n}$), (φ_1, φ_2,...φ_n) be a's rating of the importance of each of these relationships, and $SRU_{\lambda a,i}(\alpha)$ be the social relationship utility afforded to the agent in $\lambda_{a,i}$ by the execution of α. Given these definitions, and replacing f in equation 1 with an additive function in which the weights of the individual and social utilities are respectively κ_1 and κ_2, the expected utility of an action α is given by the following equation:

$$EU(\alpha)= \kappa_1 * IU_a(\alpha) + \kappa_2 * (\sum_{\forall i \in \lambda_a} \varphi_i * SRU_i(\alpha)) \qquad \text{(equation 2)}$$

This equation allows the agent to alter the balance it maintains between the local and the global needs (by varying κ_1 and κ_2), enables new social relations to be formed and old ones to be removed (by adding/deleting elements of the set λ_a), and allows the importance of the different social relationships to change over time (by varying the social relationship weighting factors φ_i).

One problem with the use of decision theory in general is that practical agent implementations do not always have the time, the computational power or the necessary information to conform to the ideal. To counter this problem, work on bounded rationality [Simon 57] has investigated techniques which enable agents to converge towards being rational (or as rational as possible) given resource constraints [Horvitz 88; Russell and Wefald 91]. Implementations of socially rational agents will be faced with further resource limitations—since they need to consider the effects of actions on others as well as on themselves. Therefore it is important that techniques are developed which enable resource-bound socially rational agents to be designed. Presently, we view this problem as being best handled by the use of a meta-level on the agent architecture which decides the amount of computation which should be allocated to determine the social utility part of equation 2 [Hogg and Jennings, 97]. Thus if the agent is severely limited it may just compute IU(α), given more time it may compute SRU(α) for the most important social relationships and as more resources are available so the agent may calculate the utility for the less important social relationships. This process can be extended until all social relations have been incorporated, in which case the agent has reached the ideal case of social rationality.

2. Socially Rational Fire-fighting

To test our ideas of social rationality, we are using the Phoenix fire-fighting simulation [Cohen et al. 89]. In this system, several fireboss agents are in charge of a number of fire-fighting resources (e.g. bulldozers, helicopters, etc.) with which they protect their designated area of the park. There are also some shared resources such as fuel carriers. The main goal of each fireboss is to minimise the amount of land lost to fires in its area. The overall system goal is to minimize land lost for the entire park. It can be seen that individual and system goals are highly inter-related and can be traded off depending on the characteristics of the agents.

In this application, the rational decision function has to determine when the agent should deploy its fire-fighting resources to fires in its local area (advancing local goals) and when it should lend its resources to other firebosses (advancing system goals). In most cases, fires are fought by a bulldozer cutting a trench around the fire to stop it spreading further. Simple fire-fighting plans involve different ways in which various numbers of bulldozers can be combined. Generally, however, the more bulldozers that work on a fire, the lower the amount of land lost. We assume that each fireboss retains at least one bulldozer under its command at all times in case fires occur in its territory. Firebosses request loans of fire-fighting resources when: i) they do not have sufficient resources to fight the fire alone; and ii) a loan would significantly increase the amount of land saved (in our initial testing we are using 30% as the significance threshold).

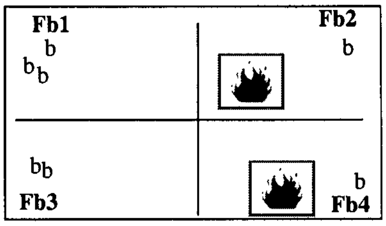

Figure 1. Example Scenario

In the Figure 1 scenario, fires have broken out in two different areas of the park. The respective firebosses—fireboss 2 (Fb2) and fireboss 4 (Fb 4)—each have a single bulldozer initially. They calculate a projection[3] for the fire given that they have one bulldozer (their current situation), and another for the case in which they have two bulldozers (in the case they can borrow one). Projections with more bulldozers could be carried out, but we limit ourselves to borrowing one for illustrative purposes. Fb2 calculates it could save 40% more land if it had an extra bulldozer and Fb4 50% more land, so they both try and obtain resources from one of the other firebosses (since this would result in a

[3] A projection is a prediction of how far the fire will spread up to a time t.

significant saving). Initially, all firebosses know of the existence of all other firebosses and of their initial resource capacity. However, they do not know what kind of attitude each has to lending resources, nor how reliable their estimates of land savings are likely to be. So making a choice initially of which agent to ask is relatively random and based on the quantity of resources a particular agent has. Assume, for example, both Fb2 and Fb4 send a request to Fb1. Fb1 has no preconception of either agent (their social weights are the same) and so its choice is based upon the agents' predicted saving of land:

EU(b->Fb2) = 40 EU(b->Fb4) = 50

Thus, Fb1 accepts the request of Fb4 and rejects Fb2's. As time passes and interactions occur, firebosses build up relationships with one another. These interactions can be used to vary the weightings of the importance of the various social groupings used in the socially rational decision making function. Thus, in the same situation as before, but after several interactions, Fb2 may have built up a rapport with Fb1 based on several successful cooperations. Meanwhile, Fb4 may not have proved to be reliable, or might not have reciprocated in giving help, so its social weighting has diminished. Hence given the same requests, Fb1 may come to prefer Fb2 over Fb4.

In a variation of the above scenario, imagine a fire breaks out in Fb1's quadrant just before it receives the requests from Fb2 and Fb4. Fb1 must now weigh the cost of attending its own fire against the utility of loaning out its resources. In making this decision, Fb1's meta-level reasoning may determine that only firebosses with high social weightings should be considered in order to save time in deliberation (so only Fb2's utility is used in computing SU). Finally, in the case where the fire in quadrant 1 is serious, Fb1's meta-level reasoning may determine that it is imperative that this fire is dealt with immediately and hence no computation of the social utility is performed (κ_2 is set to zero).

3. Conclusions and Future Work

The question of how to achieve rationality in practical agents is still an open one. Within Distributed AI, the concept of rationality becomes even more complex as we wish to achieve coordination in systems composed of multiple autonomous agents. We examined rationality from an individual agents perspective and indicated how this can be made more social so that both the individual and the overall system perform well. However, to be useful in a practical context, social rationality needs to have the notion of resource boundedness built into it—since being social and taking others into consideration, is itself a bound on the agent. We have advocated that one way of overcoming such resource constraints is to prioritize decision making according to the social relationships in which the agent is engaged. We believe that by tailoring decision making in this way, the agent can efficiently manage its resources and strike a balance between individual and system goals. Our next step is to investigate varying scenarios of resource constraints to explore how the social rationality principle and the agent's meta-level reasoning can be used to attain flexible context sensitive behaviour.

4. References

Castelfranchi, C., 1990. 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) pp49-62.*

Castelfranchi, C. and Conte, R., 1997. Limits of Strategic Rationality for Agents and M-A Systems, Model-Age Workshop.

Cesta, A., Miceli, M. and Rizzo, P. 1996. Effects of different interaction attitudes on multi-agent system performance. In de Velde, W. V. and Perram, J. W., eds., *Agents Breaking Away* LNAI 1038, Springer: Berlin: pp128-138.

Cohen, P.R., Greenberg, M.L, Hartand, D.M. and Howe, A.E, 1989. Trial by Fire: Understanding the Design Requirements for Agents in Complex Environments, *AI Magazine* pp32-48.

Doyle, J. 1983. What is rational psychology? Toward a modern mental philosophy. *AI Magazine 4(3):pp50-53.*

Hogg, L. M. and Jennings, N. R. 1997. Social Rational Agents— Preliminary Thoughts. *Proc. of Second Workshop on Practical Reasoning and Rationality*, Manchester, UK, pp160-162.

Horvitz, E., 1988. Reasoning Under Varying and Uncertain Resource Constraints, Proc. of the 7th National Conference on AI, Minneapolis, MN, Morgan Kauffman, pp:111-116.

Horvitz, E. J., Breese, J.S., Henrion, M. 1988. Decision Theory in Expert Systems and Artificial Intelligence, *Journal of Approximate Reasoning, 2, pp247-302.*

Jennings, N.R and Campos, J. 1997. Towards a Social Level Characterisation of Socially Responsible Agents, *IEE Proceedings on Software Engineering, 144(1):pp11-25.*

Kalenka, S. and Jennings, N. R., 1997. Socially Responsible Decision Making by Autonomous Agents. *Proc. 5th Int. Colloq. on Cognitive Science*, San Sebastian, Spain.

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

Simon, H. A. 1957. A Behavioural Model of Rational Choice. *Quarterly journal of Economics, 69:pp99-118.*

Wellman, M. P., 1993. A Market-Oriented Programming Environment and its Application to Distributed Multicommodity Flow Problems, Journal of AI Research, 1:pp1-23.