Chapter 2

Altruism and Co-operation

2.1 The Problem With Altruism

Agent altruism[1] and co-operation[2] are beneficial[3] within Multi-Agent Systems (MAS). Since agents often require other agents to help them achieve their goals, a society composed of eager co-operators and altruists would intuitively appear to produce a better society (in the sense of many agents achieving their goals). Simulation experiments have demonstrated the benefit of such pro-social agent behaviours within MAS [25], [31], [17], [24], [100].

Many workers within Distributed Artificial Intelligence (DAI) start from the assumption that agents should rationally act to achieve their individual goals[4]. Actions, therefore, have utility to the extent that they result in the achievement of individual goals. Such a conception of rational action was formulated clearly by von Neumann and Morgen-

[1] Altruism may be defined as action by an agent which benefits some other(s) at the expense of the actor.

[2] Co-operation may be defined as co-ordinated agent action such that all co-ordinating agents benefit.

[3] Here the term "beneficial" is used in the utilitarian sense of allowing more agents in a MAS to achieve their goals than would be the case without altruism and co-operation.

[4] However, as will be seen later, this produces problems when groups of agents are required to co-operate or behave altruistically. Consequently many DAI protocols are not based on individual rationality but on engineered solutions.

stern when they were creating the area known as Game Theory:

> "we wish to find out mathematically complete principles which define "rational behavior"... the discussion which follows will be dominated by illustrations from chess, matching pennies, poker, bridge, etc. [These studies] have their origin in the attempt to find an exact description of the endeavour of the individual to obtain a maximum of utility, or, in the case of the entrepreneur, a maximum of profit." [160].

Given this "what's in it for me?" conception of rationality, an individual rational agent clearly can not be relied upon, *by definition,* to act in an altruistic way. However, co-operation is not ruled out so long as this maximises utility for those agents involved. However, from such a conception of rational action it is important to realise that, given the opportunity, a rational agent will cheat, subvert and coerce others if this increases individual utility. Practically, work within DAI has generally side-stepped this problem by designing agents which do not have the option of cheating or coercing other agents [166]. But that kind of "fix" is only applicable to cheating which is envisaged at the design stage and does not preclude the possibility of an intelligent rational agent finding a novel anti-social method which is not precluded by the design[5]. Such socially negative aspects of individual rationality, and the benefits of altruistic behaviours, have led some DAI researchers to reconsider "what's in it for me?" individual rationality as a basis for action (see section 2.2 below).

From the perspective of human societies, such a conception of action is intuitively appealing. It appears that individuals *do* act to maximise utility, *do* cheat and *do* coerce when given the opportunity. However, there are many situations in which seemingly non-rational altruism (e.g. consider someone entering a burning house to rescue a stranger) and co-operation (e.g. paying for a newspaper at a self-service kiosk) occur. It could be argued that without these kinds of unenforced altruistic and co-operative behaviours, society could

[5]Although it must be stated that agents within existing MAS are currently far from this level of intelligence.

not exist at all. The rationality of Game Theory appears to overstate the selfishness of human individuals[6].

2.1.1 The Prisoner's Dilemma

The Prisoner's Dilemma (PD) game, minimally and abstractly, captures a common social dilemma. The pursuit of self-interest by the individual rational player leads to a poorer outcome for the players as a whole. This is an example of *sub-optimisation* [83]. Sub-optimisation results when a system as a whole (say a society of agents) can not be optimised by simply allowing each individual sub-system to optimise (individual agents). The game specifies that the highest abstract joint utility is produced during mutual co-operative interactions. However, an individual can "exploit" a co-operative other and get an individual advantage. Mutual exploitation produces the lowest utility. Table 2.1 shows the payoff grid for the two player PD game. Each player chooses from either C (Co-operate) or D (Defect). If both agents choose C then both get payoff R (Reward). If both choose D then both get payoff P (Punishment). If one chooses D and the other C then the defector gets T (Temptation) and the co-operator gets S (Sucker). When T>R>P>S and 2R>T+S the dominant strategy is D for both players since whatever an opponent chooses a player always does better choosing D. Individual rationality yields an outcome that is sub-optimal for both players.

Consider the collective production of a resource. If the collective co-operate over the distribution / production of the resource then all benefit. If, however, some individual(s) are not co-operative and "free-load" during production or take more than a "fair share"

[6]This is obviously a contentious issue. It is always possible, for example, to restate seemingly altruistic acts in the language of selfishness: "I gave money to charity to feel good about myself" or "He gave money to charity so that he would get an OBE". However, this kind of interpretation of altruistic acts still begs the question of why such an action would make one "feel good" or get an OBE. If these acts are simply irrational why is such behaviour rewarded (either by self or others)?

		Player 1	
		C	D
Player 2	C	(R,R)	(S,T)
	D	(T,S)	(P,P)

Table 2.1: The payoff grid for the one-shot Prisoner's Dilemma.

(by whatever means) during distribution, they benefit at the expense of those who were co-operative. An individual utility maximiser is therefore tempted not to co-operate (if it can get away with it). But if everyone becomes un-cooperative then the collective resource ceases to exist. This situation has been summed-up by Hardin in the phrase "the tragedy of the commons" [80]. If individuals act in self-interest then communally produced / owned resources are all possible candidates for the "tragedy" in which co-operation vanishes and all are worse off. Hardin used the example of the "environment" as a collectively owned resource, making reference to pollution and overpopulation as examples of the "tragedy" in action.

In the context of individual rationality, two agents playing the PD would select the D strategy. This is the so-called "Nash" equilibrium [59]. Essentially if an individual agent changes its strategy from C to D then it will reduce its utility since the temptation T is more than the sucker payoff S. So the "what's in it for me?" conception of rationality falls prey to the problem of sub-optimisation in the PD scenario[7]. Self interested behaviour yields a results in which agents fail to co-operate and both do worse.

Interestingly, D is not always chosen in experiments with humans playing the game for financial benefit [135].

[7]See however, Danielson [33] for a computational model of how "irrational" co-operative behaviour can be produced in the PD by agents rationally choosing a strategy based on evolved subjective preferences which do not track the objective payoff matrix.

2.2 Social Rationality

If agents are modelled as individual, self-interested, rational, goal directed systems, intentionally altruistic behaviours are excluded *by definition*. Behaviours which are not directed to the achievement of individual goals are ruled out. Cheating and subverting other agents is encouraged if this benefits the individual in the achievement of goals. This is demonstrated in the Prisoner's Dilemma (PD) presented above. In this scenario, individual rational agents will always fail to co-operate and hence both do worse than if they had co-operated. To overcome this problem of sub-optimisation it has been suggested that agents in MAS should be designed to be "socially rational":

> "Principle of Social Rationality: If a member of a responsible society can perform an action whose joint benefit is greater than its joint loss, then it may select that action." [85].

This "what's in it for everyone?" rationality presupposes that individual agents can monitor the progress of the social benefits of their actions. Often, social goals concern global properties of the population (e.g. "maximise production" in a manufacturing system) which would at best be costly for each agent to monitor and at worst impossible. The inclusion of social goals at the individual level presupposes that agents can determine how their actions relate to the achievement of the social goal, which in a complex society would be cognitively costly and possibly impractical. In situations with temporal constraints there may not be time to perform a complex social cost / benefit analysis of an action. However, putting such practical problems to one side, social rationality generalises and makes explicit what is implicit and ad-hoc when individual rationality is "fixed" to avoid cheating behaviours in specific MAS contexts (as described previously). That is, that social level optimisation should be given priority over individual level optimisation.

It would seem that social rationality would only be of benefit in a society where the majority of agents practiced it. From the perspective of human societies[8], social rationality overstates the beneficence of individuals. A socially rational individual would starve to death if that would save two others from starvation. In the context of the Prisoner's Dilemma a socially rational agent would always co-operate if it had no knowledge of which strategy the other agent was going to select[9]. As might be expected co-operation is not always chosen in experiments with humans (playing the game for financial benefit) [135].

It can be seen that social rationality only works if most agents of a society are also socially rational. Social rationality makes sense from an engineering perspective (for which it was proposed) in a closed society in which agents can not deviate from the principle. However, in an open society, or one in which agents may adapt their behaviours, socially rational agents would reward and encourage cheating in other agents. Cheating would reduce the optimality of the society as a whole.

2.3 Evolutionary Optimisers

2.3.1 Rational Action is Complex

Individual rationality will often produce societies which are very sub-optimal. It requires that agents calculate the effects of their actions before they act and select the one which produces the highest individual utility. This latter requirement is non-trivial when the actions of others have to be taken into account. Game Theory concerns itself with these non-trivial problems. As a mechanism for action in MAS, individual rationality is

[8]It should be noted that social rationality has not been proposed as a theory of human action by anyone.

[9]If an agent chooses to co-operate then the two possible joint outcomes of the game would yield payoffs of 2R (if the other co-operated) and T+S (if the other defected). Against players selecting strategies randomly this would give an expected average joint payoff of $\frac{2R+T+S}{2}$. Conversely, if an agent chooses to defect then the joint expected average would be $\frac{2P+T+S}{2}$. Since R>P, a socially rational agent would therefore choose to co-operate if it had no knowledge as to how the other player would behave.

problematic, as it may lead anti-social behaviours. As a theory of human action individual rationality cannot account for altruism and certain kinds of co-operation.

Social rationality would, in principle, produce more optimal societies but requires that agents can calculate the social effect of their actions. This would appear to be a very complex task, more complex than the individual case. As a mechanism for action in MAS, social rationality, therefore, places heavy computational demands on agents. It also requires that the majority of agents practice it for it to be of benefit. Socially rational agents would be easily exploited by individually rational agents. As a theory of human action, social rationality cannot account for much anti-social behaviour and the pervasive nature of self-interested behaviour.

2.3.2 Cultural Evolution

A radically different way of conceptualizing agent action is in the form of evolutionary dynamics. If agents are viewed as optimisers, with the ability to observe and imitate others who are relatively more successful than themselves, then the behaviour of agents in a society can be understood as an evolutionary process acting on imitated behaviours.

A detailed empirical analysis (involving field work) of such imitative processes was conducted by Lansing [107]. Lansing studied the subak based system of rice farming in Bali. The subaks (territorial units of farmers) need to co-ordinate their cropping plans to optimise rice yields. Lansing demonstrates that the imitation of cropping plans from more successful neighbouring subaks is sufficient to produce the actual pattern of co-ordination manifest in the rice growers cropping plans. Lansing used a computational model to simulate the spread of cropping plans between subaks. Each subak was treated as a single adaptive agent, adapting its behaviour via imitation.

Essentially, those behaviours that produce more utility are copied by more agents than less beneficial behaviours. Such a process therefore selects for beneficial behaviours via an evolutionary process at the cultural level. Agents can dispense with complex calculations concerning which actions will produce the higher utility and simply copy those behaviours that are observed to produce the highest utility for other agents. If some method allows for the introduction of novel behaviours (say occasional creativity or random mistakes) then an evolutionary process can occur in which utility equates to fitness. Such a process will select for novel behaviours which increase individual utility without the "overheads" of individual rationality. In most scenarios this kind of evolutionary mechanism will converge to behaviour consistent with that of a "what's in it for me?" individually rational agent [149].

Evolutionary optimising agents, though solving the problem of the complexity required of an individually rational agent, would still be stuck with the anti-social behaviours that reduce the optimality of the society. As a theory of human action we are still left with the problem of explaining altruistic and co-operative (where cheating is possible) behaviours.

2.3.3 Memes

The unit of heredity (the replicators) in such an evolutionary process are the imitated behaviours. By analogy with genes (the replicators of Darwinian evolution) these units have been termed "memes" [35]. Memes represent units of cultural information. The meme concept goes beyond simple imitation, implying that all culturally learned information can be viewed as a large collection of interconnected (yet in some sense, distinct) memes. Given the assumption that agents are optimisers that are able to imitate, copy or learn memes

which increase utility[10], then many results from Darwinian theory can be imported into a theory of cultural evolution - or "memetic evolution"[11]. This model of cultural evolution is often adopted by economists within the areas known as "evolutionary economics" and "evolutionary game theory" [13], [14][12].

2.4 Survival of the Nicest?

As stated previously (section 2.3.2), cultural evolution under the optimising assumption tends to produce agent behaviour that converges to the "what's in it for me?" form of individual rational action which often precludes social level optimisation. However, this is not always the case. Within biology several evolutionary mechanisms have been advanced to explain social level optimisation, mainly as a response to the apparent altruism observed in nature. Here three such mechanisms are summarised, they are: kin selection, group selection and reciprocal co-operation[13]. For each a brief description is given and then some of their shortcomings are outlined when applied as an explanation of the emergence of altruism and co-operation in large social systems (such as human societies), and, by implication, their applicability within large scale MAS.

[10] Although it can be argued that this assumption appears unrealistic for complex MAS or human societies (see later).

[11] In this view, "memetic" evolution produces results similar to standard Darwinian evolution at the agent level. Since it is assumed that memes which improve the utility of agents will spread due to agents being optimisers and accepting only those memes that improve utility.

[12] It is perhaps no surprise that this kind of conception of cultural evolution is appealing to economists. The "agents" in their models are often interpreted as firms or even nations. Under these kinds of interpretations the optimising assumption seems more plausible than its application to individuals. Also the conception of the optimising agent does not break with the rationality definition given by von Neumann and Morganstern (see section 2.1) and consequently stays within the assumptions of classical game theory.

[13] Sometimes known as "reciprical altruism". However, this is generally consdered to be a misnomer since the mechanism cannot account for true altruism where an individual actually sacrifices utility for another. In fact, reciprocal co-operation cannot even account for co-operation in the one-shot Prisoner's Dilemma game.

2.4.1 Kin Altruism

Kin altruism as formulated by Hamilton [79] states that co-operation and altruism among close kin relations can emerge from selfish replicators because natural selection operates on replicators (genes) not individuals. In computational studies, such behaviours have been observed experimentally [132]. Assessing the ability of an individual to assist in the propagation of its genetic material by helping others with the same material is termed "inclusive fitness". It offers an explanation for some seemingly "evolutionary irrational" behaviours such as self-sacrifice for close kin relations. The power of kin selection to produce co-operation and altruism can be seen in the social insects (wasps, bees and ants). Since reproduction is centralised (in the queen), individuals within the colony share most of their genetic material - they are all essentially siblings. In a cultural context, such centralised reproduction of culture does not occur. For application of kin altruism to culture (memetic kin) individuals need some mechanism of identifying others who share many similar memes (i.e. to identify memetic kin). Put simply, "Hamilton's Rule"[14] states that an individual would sacrifice its genetic fitness if this is likely to increase a sibling's fitness by more than twice the sacrifice. In a genetic context, the concept of relatedness is unproblematic. But how does this translate to a cultural context? What does it mean to say that "another shares 50% of my memes"? And more importantly, how could I know? These issues are discussed later (see section 2.6).

[14]Hamiltons Rule states that $\frac{C}{B} < b$. This means that the cost C to the donors utility (fitness in a genetic context), over the benefit B to the recipients must be less than b, the probability that the recipient has the same meme (or allele in a genetic context). In the genetic context, the rule means that an individual may be prepared to reduce its fitness by one unit if this increases the fitness of a sibling by two units (since siblings share 50% of the same alleles and hence $b = 0.5$).

2.4.2 Group Selection

Group selection argues that selection can operate at the group level. Given distinct (i.e. rarely interacting) groups, those groups which are more co-operative and better organised socially will do better and be selected for, thus replacing groups composed of more selfish individuals. As is often argued by evolutionary biologists there is a flaw in this line of reasoning. Though better organised groups will do better, this applies to all individuals within the groups, not just those who behave co-operatively or altruistically. So, although a co-operative group may grow, those within the group who are not co-operative and "freeload" will do even better, eventually eliminating co-operation. Group selection in the genetic and memetic contexts is therefore generally dismissed[15]. However, one mechanism by which group selection may be able to operate involves the way in which the group is changed by selfish behaviour. Consider groups which quickly "dissipate" members when selfishness is high and quickly recruit new members when altruism is high. Under these conditions a kind of group selection may select for altruism. This concept is expanded later (see section 8.2.5 in chapter 8 and chapter 9).

2.4.3 Reciprocal Co-operation

Reciprocal co-operation [158] relates to individual conditional behaviour towards other individuals. An individual will help another only if it expects the other to reciprocate. If the other does not reciprocate then co-operation is suspended. Axelrod [5] shows that such a strategy can be successful in *repeated* games of the Prisoner's Dilemma. He showed that the simple strategy of Tit-For-Tat (copying the last game strategy played by an opponent) could takeover a population of un-cooperative individuals under the conditions that: 1)

[15] However some argue this orthodoxy has become dogma and is often in error (see [165]).

interactions continue long enough to make reciprocal "punishment" effective and 2) there are enough reciprocal co-operators within the population initially. Essentially this "you scratch my back and I'll scratch yours" future orientated strategy can outperform pure selfishness given that reciprocators meet each other in the future.

Can reciprocal co-operation explain how populations can solve the social level sub-optimisation problem? There are several problems with the reciprocal argument as a general explanation of co-operation. Firstly, although Axelrod proves that reciprocal strategies are "collectively stable", he does not prove that they are strictly evolutionarily stable [120]. Briefly, although a reciprocal co-operative strategy can invade and dominate the population and can't be bettered, it can be equalled and invaded by pure co-operative strategies which in turn can be invaded by pure defection (selfish strategies). Secondly, reciprocal co-operation only becomes beneficial when interactions are for some repeated period with identifiable partners. Such constraints do not apply to many large groups of individuals. Their interactions with others are sometimes singular, or sporadic with large gaps in which interactions take place with many others. In such a setting, reciprocal co-operative strategies would only be beneficial in those interactions which continued long enough for retaliation to be effective. Reciprocal strategies place a heavy cognitive burden on each individual. To make them work, individuals need to recognise each individual and memorise past interactions (at least the last interaction). This kind of restriction would indicate that sustainable group size based on reciprocal altruism would be limited by cognitive capacity[16]. One method of tackling this would be the formation of relatively fixed subgroups of "interaction partners" [127], [84] but this would require a highly structured form of interaction which limits the generality of this form of co-operation. Reciprocal co-

[16] A detailed discussion (including empirical evidence from human societies) of group size and group organisation is given by Johnson [99]. Specifically he shows that larger groups require more complex organisational structures.

operation can not account for co-operation with strangers in a one-shot Prisoner's Dilemma game (i.e. a non-repeated game).

2.4.4 Summary

Three mechanisms that may produce co-operation under natural selection have been examined: kin selection, group selection and reciprocal co-operation. Although each offers explanations of some of the kinds of the social behaviours of interest neither seems to offer a general framework applicable to human social systems or MAS. Kin selection only applies to highly related individuals and culturally there are problems of identification, group selection is too open to "free-loading" and reciprocal co-operation does not explain true altruism or scale-up well to large groups.

2.5 The Unjustified Optimising Assumption

In the previous sections cultural evolution was equated with genetic evolution under the assumption that agents were optimisers, imitating the behaviours of others (memes) that produced highest utility. Although plausible in many contexts (e.g. the rice farmers of Barli [107]) this assumption appears to be unjustified as a general assumption. It assumes that agents have the ability to recognise and imitate actions which produce higher utility for others. Such a mechanism requires that agents can compare utilities with others and identify the memes others are using to produce a higher utility. Such assumptions are unproblematic from a genetic interpretation since fitness is equated with utility by definition and the optimising process requires no agent action other than reproduction. However, in a cultural interpretation, utility is equated with the achievement of goals and imitation or learning of fitter memes requires significant cognitive effort: agents must be identified

with higher utility, memes must be identified which produce that higher utility. By making the optimisation assumption in cultural evolution, evolutionary economics[14][17] assumes optimisation occurs *but does not model the underlying process.* The assumption that some underlying process will produce results equivalent to the optimisation assumption may well be true in many scenarios, but this needs to be demonstrated rather than assumed. The optimisation assumption misses out several well known cultural phenomena such as *frequency dependent bias* - the tendency of agents to copy memes which are encountered in others frequently, and *individual bias* - the tendency of agents to copy from some agents more than others [19], [23]. The unconstrained optimising assumption (often used within evolutionary economics) assumes that the entire population may freely copy behaviours from one another and consequently does not address issues of cultural or spatial "boundaries" [7], [126] where agents may only copy from those who are "culturally similar" or spatially near. Neither can the optimising assumption take account of issues of attachment to long held memes or the success of "active proselytizing" of memes to others which may not enhance utility. The optimising assumption, consequently, has a rather Herculean view of agent ability. Allowing the agent to identify and copy only those memes which are of benefit. To be fair, the optimising assumption allows for simple models which are often analytically tractable and can draw on existing genetic evolutionary theory. However, it is argued here that by reducing cultural evolution to genetic evolution much of what is distinctive and interesting about cultural evolution is lost. Specifically, the idea that *memes may evolve which are good at getting themselves replicated but may not be of benefit to the agents which host them*, is not consistent with the view that only memes which are beneficial to agents get replicated. This so-called "selfish meme" or "memes-eye" view of cultural evolution is often advanced

[17] Not all work within evolutionary economics makes such assumptions.

as a novel way to think about culture and throw light on seemingly irrational behaviour (see [15], [68]).

Computational MAS simulation models can be constructed (artificial societies) which begin to address some of these issues by *modelling the underlying process of cultural evolution*. By relaxing the optimising assumption, and including simple mechanisms that capture some of the distinctive characteristics of cultural evolution (as simply as possible), it is possible to observe whether the optimising assumption would produce the same results for a given cultural scenario.

Two of the three artificial society models presented in this thesis implement agents which are "satisficers" rather than optimisers. Satisficing agents attempt to achieve a satisfactory level of utility beyond which they do not continue to optimise.

2.6 Behaviours and Tags

A necessary condition of kin altruism (see section 2.4.1 above) is the ability to distinguish between kin and non-kin. Without such a mechanism, altruistic behaviours could not be directed towards the correct individuals. Many mechanisms may be used. Any mechanism would need to balance simplicity (requiring little effort) and effectiveness (making few mistakes). In a given setting, kin altruism could only function if the correct balance was established. An overly simplistic mechanism may be too ineffective, a highly effective mechanism may be too costly (outweighing the benefits of altruism). It would seem that some ability to recognise a genotypical "marker" (expressed as a phenotypic feature) could be one mechanism[18]. This mechanism implies that the genotype of such an individual contains both the mechanism for reading the marker and also the marker itself. Another

[18]Such a mechanism termed "the green beard effect" has been theoretically speculated [35] and recently empirically observed [101].

might be to quickly "imprint" distinctive characteristics of kin after birth. An even simpler method, possible when individuals occupy fixed spatial regions, could involve a spatial bias (those spatially close are considered as kin).

2.6.1 Memetic Kin

Memetic kin altruism can only function if memes can induce individuals to distinguish between memetic kin and non-kin. Consider a meme which induces a host to recognise those others who posses the same meme. If the meme also induces co-operative or altruistic behaviour between the two hosts, and if higher performance or utility effects meme propagation positively (i.e. hosts with higher utility are more likely to propagate their memes than those with lower utility), then such a meme would tend to be successful[19]. However, this depends on a host identifying that another possesses the same meme. The SwapShop model (see chapter 5) demonstrates this process.

2.6.2 Surface and Hidden Memes

How can agents recognise memetic kin without being able to see anothers' memes directly? Here I distinguish between two kinds of meme: 1) a surface meme (or feature), 2) a hidden meme (or conditional behaviour). Both of these could be the expression of some belief or behaviour. If a host believes that by wearing a white hat its life is prolonged, this produces the surface meme expressed as the wearing of a white hat (assuming the host wants to live). If the belief is passed to others they will also wear white hats. Alternatively others may simply be influenced to wear white hats because everyone else is, without subscribing to the belief. In both cases the surface meme has been replicated. A hidden meme is one which can not be observed in a direct sense. Consider the behaviour "co-operate with those

[19] A memetic form of the "green beard effect". This kind of process is theoristed by Allison [2].

wearing white hats" which might be the expression of a belief "those with white hats are friends". It would seem that this kind of hidden meme could not be copied directly since simple observation might not be sufficient to determine the underlying behavioural rule being employed.

A surface meme can become a signal which activates or triggers a behaviour based on a hidden meme. This kind of behaviour, selecting behaviours relative to tags (or social cues) may be seen as a restricted form of stereotyping [130], [109]. It may be seen as a boundedly rational mechanism employed by agents to structure interaction with strangers (see chapter 6).

Certain combinations of surface and hidden memes (tags and conditional behaviours) when combined may produce beneficial coordinating activity between hosts. So the combination of the surface meme of white hat wearing and the hidden meme of white hat co-operation would seem to make a fortuitous combination. White hat wearing becomes a "marker", "tag", "label" or "social cue" which identifies an in-group: the group of white hat wearing co-operators. Of course, a host may wear a white hat without co-operating, exploiting the convention (free-riding). Does the possibility of such a process preclude the emergence of co-operative tag / behaviour pairs? Allison [2] speculates that such problems can be overcome via "cultural packaging mechanisms"[20]. But given that the surface tag and hidden conditional behaviour are represented as separate memes how can this be achieved?

2.6.3 Cultural Packaging

When cultural and behavioural interaction involves either a spatial or cultural bias, this may produce spatial or cultural regions[21] of shared memes. Cultural bias, involves

[20] A cultural package might be termed a "meme bundle" or a "meme complex" [40], [15].

[21] By "cultural region" is meant a set of agents demarcated by some set of common surface memes.

interaction biased towards those who are already culturally close (based on surface memes). Spatial bias involves interaction biased towards those who are spatially close. The function of both of these processes is the same: by biasing interaction to a subset of the population, groupings of hosts may form with sets of shared memes.

Assuming that regions exist with shared surface and conditional behaviour memes, regions which form beneficial coordination between the two should out-perform regions which do not. If groupings which produce high utility via in-group co-operative behaviours are more stable than those with low in-group co-operation then the members of the low utility groups may be absorbed into the higher utility groups. However, a "free-rider" within a group who discovers how to exploit any co-operative convention would appear to break down such a process (e.g. the Prisoner's Dilemma).

In the context of optimising agents, both spatial [127], [102], [93], [84] and tag based [143], [88] biasing have been discussed and explored. The results obtained from spatial studies demonstrate that increased localisation of learning and interaction promotes the success of co-operative strategies. This is due to the insulation of clusters of co-operative strategies from contact with non-cooperative strategies. Riolo [143] gives results of simulations demonstrating the effect of interaction biased by tags. In his study, he represents tags as a single real number attached to individuals. His results indicate that biasing of interaction by tag similarity promotes co-operative strategies in the Iterated Prisoner's Dilemma (i.e. the repeated or iterated PD game). He does not however, explore the effect of biasing learning by tag, since learning in his simulations is based on a population level optimising algorithm. Riolo's results show that co-operation is only promoted in the Iterated Prisoner's Dilemma, where reciprocal co-operation is possible and therefore does not demonstrate true altruism. It is also unclear from Riolo's results as to the underlying mechanism by which

tags actually promote co-operation.

These studies capture some of the aspects of the meme processes discussed above. However in the context of the replication of strategies, they prescribe that host utility is the only determinant since they make the optimising assumption and *do not model the underlying process*[22]. As previously stated such optimising models do not consider social pressure in the form of frequency dependant bias (being predisposed to follow the crowd). Often those studies which do attempt to capture these phenomena [125], [7] study the movement of passive attributes, with no direct effect on agent behaviour (i.e. purely surface memes or tags). It is claimed that it is the combination of these two mechanisms that characterises meme processes. In the SwapShop (chapter 5) and the StereoLab (chapters 6 and 8) these processes were modelled based on a satisficing agent model.

2.7 Summary

In this chapter the problem of sub-optimisation in agent societies has been described. Various solutions to the problem of sub-optimisation, from evolutionary biology have been examined and their shortcomings discussed. A memetic conception of culture has been proposed in which an agent optimising assumption has been questioned. Mechanisms based on group formation and cultural markers (labels or tags) which may promote co-operation and altruism between strangers have been outlined, and a game theoretical formulation of sub-optimisation in the form of the Prisoner's Dilemma has been introduced. In the following chapters artificial society models are constructed that explore and make concrete these concepts by investigating the conditions under which groups and tags promote altruism and co-operation within cultural evolutionary scenarios.

[22]This is not a criticism of Riolo's model since his interpretation is one of genetic evolution applied to "animals searching for interaction partners".