The evolution of strategic thinking

Adam Morton

in Peter Carruthers and Andrew Chamberlain, eds., *Evolution and the human mind: Language, modularity and meta-cognition.* Cambridge U.P. 2000, pp 218-237

0. where I'm going The theme of this chapter is that some seemingly arcane philosophers' disputes about the nature of rationality are relevant to questions about the evolution of *strategic thinking* in our species. By strategic thinking I mean, roughly, the kinds of cognitive processes that an agent uses in order to choose an action or a strategy in a situation involving other agents, taking account of the fact that the others are also choosing acts or strategies in the light of the same factors. There is a fairly familiar connection between strategic thinking and evolution in terms of considerations governing the choice of strategies that species can adopt in their interactions among themselves, for example with Maynard Smith's concept of an evolutionary stable strategy. A crucial question will be the relation between the criteria governing choice of strategy for an evolving species and for individuals in the species. The strategy of the paper is first to present reasons for dissatisfaction with the central tool of present theories of strategic thinking, the concept of an equilibrium. There are equilibria which it would seem that intelligent creatures should avoid rather than gravitate towards. Then I argue that there is really no alternative to thinking in terms of equilibria, but that by using the concept carefully we can see how evolving intelligent creatures could cope with the situations that make it problematic. [*]

1. strategy and equilibrium When intelligent agents interact the consequences of each agent's decisions depend not only on the state of the world and chance but also on the decisions that other agents make. Strategic thinking is the thinking that leads to decisions which take account of the factors that affect other agents' decisions. This is the standard meaning of the term in economics, game theory, and philosophy, where strategic choice is contrasted with parametric, that is, non-strategic, choice.

[*] Versions of this paper were read to the Hang Seng workshop and conference, and received a wonderfully useful discussion. I am particularly grateful to Alex Barber, Peter Carruthers, Andrew Chamberlain, and Robin Dunbar for comments.

Biologists sometimes use 'strategic' so that the emphasis is instead on strategic deception, which is certainly an aspect of strategic thinking, but not central to it as it is understood by these other disciplines. (From the perspective of game theory deception presumably falls under the general topic of situations where agents have incomplete information, which is a very live but also rather controversial and unsettled part of the subject. See Myerson, 1991, chapter 2 sections 7 to 9 and chapter 4.) The most important thing to come to terms with about strategic thinking is this: *strategic choice cannot rely on beliefs or probabilities about the other agents' choices.* The reason is clear. The agent herself is thinking through what to do and so are the other agents. She does not know in advance what *she* is going to do, and as she deliberates her expectations about what acts she might choose may change completely. Moreover her decision depends on what she expects the others to do, and their decisions depend on what they expect her to do. So as everyone concerned thinks out what they and the others may do the probabilities of their actions change, until finally they crystallise in a decision. There are no stable probabilities to guide a decision; they only emerge once a decision has been made.

This basic fact has an important consequence: it is not evident what we might mean by saying that one strategic choice is rational or optimal. In non-strategic choice we have a vague handle on rationality in terms like 'the rational choice is the one which is most likely to maximize satisfaction of an agent's preferences.' But both 'most likely to' and 'maximize' here are implicitly probabilistic: the natural way of making the idea precise is in terms of expected utility. The best act is then the one whose average outcome is greatest, taking into account all possible future states of affairs but giving more probable ones more weight than less probable ones. To do this we have to think about what kinds of probabilities we are talking about, but with strategic choice we have deep problems before we have a chance to ask these questions. Here is an example that brings out several basic points. It is a 2-person game for two players, I and II, diagrammed in the usual matrix.

G1

	II	
I	**F**riendly	**U**nfriendly
Adventurous	(10,10)	(-100,5)

Defensive (6,2) (0,0)

What should agent I choose? One line of thought runs: in the absence of probabilities she should act with extreme caution, choosing the act whose worst consequences will be least bad for her. This is the minimax strategy. In this case this means choosing **D**, since the worst possible consequence of **A** are very bad. Or she might assign probabilities via an indifference principle: 50/50 that II will chose **F** or **U**. Then the act with the higher expected pay-off will be **D**, so again she will choose that. But both of these strategies have a serious flaw, when one adopts the strategic attitude. Agent I can consider what the situation looks like from II's point of view. Then she can see that he is very unlikely to choose **U**, for whatever I chooses II will be better off if he has chosen **F**. **F** is II's dominant choice; he will only choose **U** by mistake. So if I thinks that II is informed, intelligent, and paying attention (I am avoiding 'rational'!) she will choose **A**.

The combination (**A**, **F**) is an *equilibrium outcome* of the game in that if either player chooses their part of it then the best response of the other is to choose their part too. Or, equivalently, it is stable in that if either player knew that the other was going to choose their part they would choose their own part too. (Perhaps of interest: the equilibrium outcome in this case is also each agent's best response to the other agent's minimax choice.) There are economists and others who equate the equilibrium outcome with the rational choice. But this is not a position that is at all forced on us; it smacks more of prescriptive definition than of analysis. (It was not the attitude in the early days of the subject, except perhaps in the special case of zero-sum games, and it is increasingly doubted among contemporary writers. It seems to have had a temporary semi-orthodoxy though. See Luce and Raiffa, 1957 chapter 2; Myerson, 1991, Chapters 3 and 5; Skyrms 1991, chapter 2.)

Equilibria are particularly attractive when we move away from the original domain of game theory - one-off unrepeated choices by individuals whose earlier and later interactions are irrelevant - to consider situations in which agents make the same choices in the same situations repeatedly. Suppose that two agents meet in a situation like the one above and the combination of their choices is not an equilibrium. Then one or another of them would have done better by making a different choice. So it may be that next time that agent will make that different choice. Or both may. And so on. As a result it is plausible (*not* inevitable) that they may eventually gravitate to

a combination of choices in which each agent's act is the best response to that of the other. This situation will be stable in that no agent will have a motive for deviating from it as long as the other does not deviate. (Note the qualification.)

Or imagine a different model: both agents tend to choose by considering their previous action and thinking how they could have done better. When this doesn't tell them what to do they choose at random. Suppose that one way or another they arrive at an equilibrium. Then they will stick there, since neither can improve on the situation as long as the other acts as they did the previous time.

The stability of the equilibrium will be greatest if it is strict, that is, if any agent will do strictly worse by unilateral deviation. The equilibrium is non-strict if unilateral deviation may get some agent an equally good result though not a better one. The standard equilibrium concept is non-strict, but we can obtain more stability while not moving all the way to strictness. We can do this by requiring that any unilateral deviation set off a series of changes which takes all agents back to the equilibrium. One agent shifts to a just-as-good alternative, which changes the situation of the other who can now do better by herself shifting, putting both agents in a situation where they can do better by shifting back to where they started.

This can be put more formally. (And also more off-puttingly. All that really matters is the informal explanation I've just given.) A combination (a,b) of choices for two agents I and II - the extension to n agents is no problem - is a *stable equilibrium* iff (i), (ii) and (iii) are true.

(i) there are no choices x, y such that (a, x) is strictly better for I than (a,b) or (y,b) is strictly better for II than (a,b).

(ii) if there is a choice c such that (a,c) is better for II than (a,b) then there is a choice d such that (a,b) is better for I than (a,d); (a.d) is better for I than (a,c); and (a,b) is better for II than (d,c).

(iii) if there is a choice d such that (d,b) is better for I than (a.b) then there is a choice c such that (a,b) is better for II than (a,c); (a,c) is better for II than (a,d); and (a,b) is better for I than (d,c).

((ii) and (iii) are just I/II mirror images.)

A special case of stable equilibrium is well know in biology as an *evolutionary stable strategy* (ESS). (The classic source is Maynard Smith, 1982. See also Skyrms, 1996, chapter 3.) We get an ESS if we imagine the interactions to be between biological individuals who gain or lose reproductive fitness relative to one

another. We then imagine that all the individuals form part of one rough species-with-variants. We can then ask which variant will do best in interaction with others from this pool. Since individuals are part of the same rough species we can have some idea of "same" and "difference" as applied to their actions. We can also suppose that the payoffs are symmetrical in that the gain in evolutionary fitness to I when she chooses strategy a against II's b is the same as the gain to II when she chooses a against I's b. Then the three conditions above reduce to two. A strategy x is an ESS iff (i) and (ii) below are true, where (z,w) > (t,s) means that the payoff of (z,w) to the agent choosing z is greater than the payoff to that agent of (t,s). Similarly for =, =.

(i) for any y (x,x)= (x,y).

(ii) for any y if (x,x) = (x,y) then (y,x) > (y,y).

An ESS is a special case of a stable equilibrium, where the biological context allows us to simplify by assuming a symmetry in the payoffs and assuming that we can identify actions of one agent as the same as actions of the other. But the importance is the same: any deviation from equilibrium leads to a chain of deviations, which leads back to the starting point. The consequences are clearest when deviations are mutations. Start with animals interacting via some action x performed by all of some species and then introduce a mutation which interacts via a different act y. If x is an ESS (i.e. x against x is a stable equilibrium) then the mutation is unlikely to get established. Conversely if we begin with animals interacting via some action y which is not an ESS then a mutation to an action x which is an ESS is likely to get established, and in fact to displace y. In this connection it is worth noting that a set of choices is an equilibrium, or a strategy an ESS, only with respect to a given set of alternatives. Add another choice and an equilibrium may cease to be so. Thus a strategy x may be an ESS as long as some rival y is not available, but as soon as a mutation makes y available then x may cease to be an ESS. It may be that no strategy then is.

It appears then that at a very abstract unrealistic level we can say that a species tend to evolve towards ESSs, and that a long-term departure from a ESS is likely to be to another one. This is like saying that combinations of rational actions tend to be stable equilibria. But this latter claim is not clearly true. In fact, on the face of it, it seems false. For it seems to say that, in trying to choose for the best, rational agents

will tend toward situations that are stable equilibria. The problem about this claim is that there are many stable equilibria which are decidedly *not* in the best interests of the agents concerned. Human agents very often do not choose them. So how could they be required by rationality, or inevitable results of trying to choose for the best?

The standard example to make this sort of point is the familiar prisoner's dilemma, or PD.

PD

	II	
I	**C**ooperate	**D**efect
Cooperate	(5, 5)	(0, 9)
Defect	(9, 0)	(1, 1)

The payoffs have the right kind of symmetry for there to be a ESS, and indeed **D** is a ESS (and thus a stable equilibrium). But it is not optimal, in the following sense: if a pair of agents both choose **D** they will do less well than if they both choose **C**. It *is* optimal in a different sense: whatever one agent chooses the other will be better off choosing **D**. The sense in which it is not optimal is relevant to actual human choice in that in many prisoner's dilemmas people do choose the action corresponding to **C**. They do so trusting that the other person will too, either from moral principle or because they see that it would be best for both of them if they both so chose. We have agreed you'll send me a first edition of *The origin of species*, if I send you £2,000. **C** is keeping ones side of the deal and **D** is doing nothing hoping the other will be suckered into sending for nothing. Most people most of the time will do what they have agreed to. Especially if, as in a PD, it is worse for both if neither keeps their word than if both do.

It is worth remembering now the doubts expressed earlier about identifying rational choice with choice that conforms to an equilibrium. Those doubts had much less hold when we considered not what choices people would make one-off, but what choices they would make repeatedly. An equilibrium is then a situation which once reached is unlikely to be departed from. And many PDs do indeed have this quality. If in the past you have not sent the money for transactions we have agreed, then the

next time round I am not going to send the goods. Knowing this you won't send the money, and we won't do business.

Switch back now from individual choice to evolution. Both senses of optimality are relevant. The first sense applies when we have two populations of animals interacting in a PD one of which interacts via **D** and one via **C**. Then the one interacting via **C** will do better than the one interacting via **D**. So the C-interactors will do better, have more grandchildren, than the D-interactors. The second sense applies when we have a mixed population of C- and D-interactors. Then as long as everyone interacts with everyone else the D-interactors will do better. The C-interactors may wish that the D-interactors were not around, but as long as they are, in any numbers, the best policy is to join them.

In fact, the first of these situations is unlikely to be sustained. It takes just a few D-interactors to have got within the walls – by invasion or mutation – to ruin the D-interactors haven. The situation can only be sustained either by some mechanism that constrains C-interactors to interact only with other C-interactors, or some fact of the situation that makes D- against C-interaction less good than it appears for D-interactors. (An example of the latter would be shared genes, so that in clobbering a C-interactor a D-interactor would be lessening the survival of its own genes in the long term. But this, like other similar facts, covertly changes the payoffs so that we no longer have a PD.) As a result a population of C-interactors is likely to evolve towards a mixture of D and D interaction, in which the proportion of C interactors is likely to decline, even though this decline will result in a lessening of the general fitness of the population. (The matter of proportions is delicate, see Skyrms 1996, chapters 1 and 4.)

Some stable equilibria, ESS, are thus as much traps as improvements. They are like the situation of would be traders who have no reason to trust one another enough to do business. Once a population finds itself in such a situation there is a strong pressure to remain in it, in spite of the fact that there are combinations of actions that are better for all concerned. But these better combinations are not stable.

2. choosing a choice-mechanism We are facing some challenges to naïve convictions. We naïvely think that evolution results in better-adapted animals. And we naively think that rationality allows agents to choose the options that are in their

best interests. So we might think that when an ESS has the trap quality of a PD then somehow or other evolution will find a way to a situation in which combinations of better actions will have the stability they lack. And we might think that rational agents will find ways of cooperating in PDs and similar situations without running the risk of defection by others.

These are indeed naïve ideas. They do not follow from evolutionary theory or from any standard account of strategic decision. But putting them together, we might well make a conjecture about the evolution of rational decision. Ways in which individuals choose their actions are subject to evolution, like everything else. So we might conjecture that ways of choosing that avoid PD-like traps will be favoured in evolution. Creatures that develop them will have an advantage over creatures that do not.

The purpose of this section is to show that this conjecture is false, at any rate on one way of understanding it. What a pity. The argument is fairly abstract, and it leaves us roughly where we were. So if the question does not interest you, skip to the next section, which is less abstract and less disappointing.

Suppose we have an organism which interacts with conspecifics and mutants in strategic choice situations and in its evolution all possible ways of making choices are tried out. Which ones will be selected? Put less carefully: what is the optimal way of making strategic choices?

My way of answering this question is a traditional 'backwards' one: I assume that there is such an optimal way and then I deduce what properties it must have. So suppose that we have a population of individuals, each of whom has preferences over a range of outcomes which can result from the effects of combining their actions with those of others. (The preferences may be for outcomes with greater reproductive benefit, though the argument does not assume this.) And suppose that each individual has an 'oracle' which given information about the situation, including the preferences of other agents, pronounces on the action to choose. In the evolution of the species we can imagine that a variety of oracles have developed, and we can imagine that each oracle-bearing agent interacts with others bearing variant oracles. For example there is the oracle EQU which always suggests choosing an action which is an equilibrium solution to the game; (Suppose that if there are more than one equilibrium EQU makes a random choice between those which are best for the greatest number of agents.) There is the oracle MINIMAX which always suggests

choosing that action whose worst consequence is least bad; there is the oracle COOP which says 'in a prisoner's dilemma, choose the cooperative option'. (COOP could be extended to an oracle giving advice about all games in many different ways.) And we can wonder what will happen when, say, MINIMAX-respecting individuals interact with EQU-respecting individuals.

Suppose then a strategic choice situation s involving two individuals and consider a 'metachoice' between two oracles O and O' prior to choosing between the acts that s offers. This sets up a new situation s[O, O'] in which the choices are which of O and O' to apply to s and the outcomes are the outcomes that would result from the combinations of the choices between acts in s that would be made from these metachoices. (First you decide which oracle to listen to, and then you do what it says.)

Sometimes given a situation s, and a prior choice of oracles O and O', one of O and O' will clearly be an ESS of s[O, O']. For example if s is PD as above, and the oracles are EQU and COOP then both agents will choose **D** if they have first chosen EQU and **C** if they have first chosen COOP. The matrix of the resulting outcomes is another prisoner's dilemma with (EQU, EQU) as a strict equilibrium, so EQU is an ESS. On the other hand if s is G1 above and the oracles are EQU and MINIMAX, then agent I will choose **A** if she has chosen EQU and will choose **D** if she has chosen MINIMAX, while II will choose F either way. (Note that the resulting matrix is not the same as that of G1.) There are two equilibrium outcomes of G1[O, O'], (EQU, EQU) and (EQU, MINIMAX). So EQU, is not an ESS of this two-stage situation.

This may seem slightly paradoxical. In the situation in which going for the equilibrium is on reflection the right thing to do (assuming the other player is rational), G1, EQU is not an ESS, while in PD, in which for most people the rationality of the equilibrium is somewhat puzzling, EQU is an ESS. The puzzle diminishes when you consider that an ESS is the situation that we can expect agents to gravitate to in the long run, if they are continually trying alternatives and sticking to them if they pay off. It is no surprise that a choice-procedure that leads to defection in a PD is like this. It may be more surprising that EQU in G1 is not an ESS. But this is really a limiting case. Choosing equilibria is clearly optimal for I in that it is a dominant strategy - it is better whatever II chooses - while II is strongly indifferent between the two choice-methods - they have exactly the same consequences for her. So if we begin with a population of equilibrium choosers interacting via G1 and

consider possible mutations to minimax the individuals playing the II-role would mutate quite freely while the individuals playing the I-role would stick to equilibrium choosing. If individuals switched from II-role to I-role they would quickly cease mutating. The II-role mutations don't really count, in that the consequences for the agent of choosing the two methods are exactly the same. From the point of view of G1 EQU and MINIMAX are the same strategy.

This discussion suggests a general result, which is in fact true.

If in a given 2 person game s EQU and some other method M lead to different acts for each of the agents concerned then EQU is an equilibrium of the game g[EQU, M].

Proof: Suppose that we have a game s such that EQU is not an equilibrium of g[EQU,M]. Then there is an act *a* of s such that if one agent chooses M and as a result chooses *a* in G, and the other agent chooses EQU then the outcome *o* is better for the first agent as the outcome *e* if they had both chosen EQU. So *o* better for the agent as *e*. But the result of their both choosing EQU is an equilibrium of s, by definition of EQU. So *e* is at least as good for the agent as *o*, by the definition of an equilibrium. Contradiction.

If we are thinking in terms of evolution then ESS has more interest than bare equilibrium. So we might wonder if EQU will in general be an ESS when oracles are being chosen and used. Not surprisingly, it will not always be. There are games s with a single equilibrium that is not a stable equilibrium such that for some O EQU is an equilibrium but not a ESS of s[EQU, O]. (I won't give details. For all I know this may be true of all such s.)

None of these results are very profound, and indeed there is a disappointing quality to them. It looks as if an evolution of choice mechanisms will favour some (realistic approximation to) an equilibrium-choosing procedure. And this makes it appear as if the advantages of non-equilibrium choice, for example in prisoner's dilemmas, can never evolve. That appearance is in part due to the naive abstraction of the discussion so far.

3. not taking ESS too seriously The argument for the conclusion that we can expect a species to evolve towards an ESS is simple. So are the reasons for expecting

strategic thinking to evolve towards the choice of equilibria. But their premises are quite strong, and it is not obvious when they are satisfied. To see this, consider some consequences that do *not* follow from the idea of an ESS.

It does not follow that the species will evolve to the equilibrium *soon*. In fact the considerations that suggest that the equilibrium is a likely eventual destination also suggest that in some cases it may take a long time to get there. For example suppose that starting from many particular combinations of strategies there are many alternative strategies that each participant can change to, which present advantages over the beginning combination. Then one possibility is a long migration from one combination of strategies to another, arriving at the equilibrium if at all after many intermediate stops. It may be that from some of these half-way halts the equilibrium is in actual biological terms rather remote. (From the starting point fairly simple mutations would get to the equilibrium. But if instead we mutate to an alternative the mutations that would then produce the equilibrium may be much less simple, or biologically plausible. If you can get easily from A to B and from A to C it does not follow that it is easy to get from B to C. London, Bristol, and Cambridge.)

It also does not follow that there could not be much better equilibrium outcomes for the species than the ESS. Suppose for example an ESS in mating behaviour in which males defend territories and females try to estimate which territory is biggest and best defended. Members of the species might do much better if males could produce infallible proof of their fortitude and females of their fertility. (Or if they could scan one another's DNA, or if they could register with a discrete mating agency.) But this is just not biologically possible for them: there is no pattern of mutation or learning, available to those animals at that time, that is going to get them there. To conclude that the species will tend to the ESS we must assume that the situation of which the ESS is the equilibrium is composed of strategies which are biologically possible developments for the species. In fact we have to assume that the situation includes all the biologically possible developments for members of that species at that time. If not, there may be other combinations of strategies to which the species could move from its initial situation rather than to the ESS, or to which it could move after arriving at the ESS.

(This point combines powerfully with the observation above that accessibility of B and C from A does not entail accessibility of C from B. Suppose a fairly tight construal of what is biologically possible, so that the ESS is in fact a likely outcome

of evolutionary forces. Suppose also a large number of possible strategies that the animals could adopt. Then there will be a great danger that though all these strategies are possible starting from some actual initial situation and although one of these is an ESS, that ESS is not itself possible starting from some of the alternatives. This would lessen the explanatory force of the model considerably. To avoid such worries, biological applications of game theory usually work with an extremely tight understanding of what is a biologically possible alternative to an existing situation.)

These two provisos should make us hesitate before using the arguments of the previous section to show that creatures involved in strategic choice will in any realistic sense gravitate towards equilibrium-choosing decision procedures. For there are many many ways of making decisions, some differentiated from others in subtle ways with finely balanced advantages and disadvantages. And from some presumed starting point of a strategic species most abstractly conceived decision-making methods will not be available. So what we should really expect is that decision-making procedures will move towards very local equilibria, stable only given a narrow range of alternatives. And we should expect that as other features of the species and of its environment change these will cease to be optimal and will shift to other equally local good-enough solutions. We cannot be surprised if sometimes the evolution to such a local equilibrium blocks the way to a more global one: an adequate decision-making method undermines the pressure to discover an ideal one. (It is for this reason that evolutionary theory is compatible with the contingency in evolution defended by Stephen Jay Gould. See Gould, 1989, and Sterelny, 1995.)

Now reconsider in this light prisoner's dilemma-type situations, in which the equilibrium is not the best outcome for any of the participants. Consider two extreme perspectives.

On the one hand consider situations involving two animals with hard-wired social choice mechanisms. Ants, for example. Suppose that the payoffs for combinations of cooperative and defecting behaviour are as in the prisoner's dilemma, but that the choice of one behaviour or another is completely determined by which group (nest, colony, or band) the individuals come from. Then individuals cannot but act cooperatively to individuals of the same group, and cannot but act defectingly to all others. What we have then is not accurately described as a prisoner's dilemma at all. For the options actually available to individual animals never form the required structure: on each occasion either cooperation or defection is simply not an option.

And the mechanism that selects either behaviour on the basis of a classification of the other is, we may suppose, effective and stable. It is an ESS among the alternatives within evolutionary reach of the species.

At the other extreme consider situations involving ideally rational agents. Again suppose that there are options available to them which define outcomes with the shape of a prisoner's dilemma. The situation of the agents only is a prisoner's dilemma, though, if these are *all* the options, and if the agents really are ideally rational this is very rarely the case. A crude way to express this is to say that if agents are rational they will have created some of the social apparatus of rational social life: contracts and means of enforcing them, motives for third parties to react appropriately to violations of conventions, and so on. (*We* are not ideally rational agents, so we can only gesture at the apparatus of fully rational life.) So when two of them approach a potential prisoner's dilemma they can either take steps to make sure that the actual payoffs do not fall into that pattern or there are other options besides cooperation and defection. (For example they can do the former by tying behaviour in the PD-like situation to choices in later situations: if you defect now my friends will defect on you later. And they can do the latter by signalling their intentions to cooperate and to behave vindictively if cooperation is not matched. Signalling often requires preparation: in practice the two kinds of measure are hard to distinguish)

There is a more subtle argument for the remoteness of prisoner's dilemmas for ideal agents. If they are ideally rational they can do game theory. And they can deduce what *would* be a set of institutions that would guide agents like themselves through problems of acting for mutual benefit. For example they can think through the various conventions that could regulate future behaviour of third parties to someone who has defected in a prisoner's dilemma, and come to conclusions about which ones are best. As a result, even without any explicit communication or convention each agent will know what reactions, and alternatives, and sequels would be the object of a convention, and each agent will know that each other agent will know this and that she knows it. (It will be mutual knowledge.) The conclusion is surprising: if there were conventions which ideal agents would adopt in order to disarm prisoner's dilemmas (and other similar situations) before they arose, then such agents would not have to explicitly adopt them; the mere knowledge that they would be adopted can do all the work. Ideally rational agents would be pre-insulated against all but the most unforseeable prisoner's dilemmas. (Arguments like this are

sometimes presented with an ideological slant: smart people don't need governments. For a clean and unrhetorical expression of the attitude see Sugden, 1986.)

The prisoner's dilemma is thus not a major issue either for animals with fixed socially routines or for ideally rational agents. It will only arise as a persistent feature of life for creatures who choose their actions on the basis of expectations about one another's choice, but bring an imperfect rationality to such choices. And that certainly sounds like us! But notice how the context has changed. We began by fearing that the evolution of strategic choice would converge on patterns that condemn agents to the consequences of double defection. Now we see that evolution can tend towards choosing equilibria, while making some uncomfortable equilibria very rare. For sufficiently evolved agents would make sure that they were rare. PD-like situations are a feature of the middle stages of evolution.

4. coalitions Creatures of active but limited rationality, such as human beings and many of their precursors, will make strategic decisions by a complex improvised combination of innate social routines and attention to the likely consequences of possible actions. There is an aspect of strategic choice where explicit strategic thinking and primate social capacities combine in a revealing way: that of the formation of coalitions. A coalition arises when two or more agents can by choosing suitable actions obtain a benefit in a situation involving a greater number of agents. What I shall argue is that the formation of coalitions - something which can be based on innate primate sociality - is systematically related to the measures by which rational agents can ensure that they are not caught in PD-like situations.

The analysis of coalitions plays a large role in the application of game theory to economics. (See Myerson ,1991, chapter 9.) The standard theories assume an interestingly richer apparatus than is usual in the game theory of individual agents. The basic assumption of the theory of coalitions is that each agent will perform an action which is part of a pattern of actions by some coalition which maximizes the total pay-off to the coalition. This assumption bears a delicate relation to the simpler assumption that each agent will perform an equilibrium action. (It does not follow from it, without additional assumptions. Its consistency with it is not obvious. Their relations are worth investigating for their own sake.) Using it, though, we can get very plausible and common-sensical predictions about the coalitions that will form in

particular situations. In particular, we can predict that people will cooperate with one another in order to form a profitable coalition against others, even when so doing requires them to forego the advantages of free-riding or defection. In effect, in a coalition individuals find their way to cooperation in PD-like situations.

This standard theory leaves the basic question unanswered. It just assumes that individual rational actions can form coalitions, without explaining how. But in some cases it is not too hard to see how coalitions can work. Consider a case which illustrates how a coalition can solve a PD.

Two proto-humans, Cain and Abel, are considering whether to fight over a carcass or to share it. Either could be aggressive (**D**) or pacific (**C**). If either is aggressive and the other is pacific then the aggressor gets the carcass to himself and the other is badly beaten. If both are aggressive then both are bruised, but not badly beaten. If both are pacific then they share the carcass. They would both rather share the carcass than get bruised. Then the structure is that of a PD and seems to have the same sad consequences. If either is pacific then he will be beaten by the other. So they will both fight, and both will be bruised, though each would have preferred a peaceful division of the carcass.

Now add one more element. An old ferocious patriarch, Adam, who will hear any quarrel and interfere, beating both participants and taking anything worth having. In Adam's presence the situation is no longer a PD. Aggression leads to a worse outcome than pacificity, whatever the other does.

Now suppose that Cain and Abel see a carcass that they might share or fight over. They can sneak towards it, or stride in a noticeable way which will catch Adam's attention. If they sneak then they are in the original PD. If they stride then they are in the variant situation in which Adam enforces cooperation. The predictable outcome of sneaking is thus mutual bruising and that of striding is an equal share of the carcass. (I'm assuming that they can share it and move off without bringing Adam down on them, but that a fight will make it impossible to avoid him.) So they will stride and share.

There is something magical about this simple conclusion. The presence of Adam allowed them to achieve exactly the cooperative outcome that they could not achieve in his absence, even though he did not perform any action or get any benefit. The fact that if they had acted differently he *would* have intervened does the trick, ensuring that they do not and he does not.

Notice though the calculation required. Cain and Abel must consider their actions several moves ahead, and each must consider how the other will consider their future actions. Each has to know that the other will see the advantage of making Adam aware of their presence. Moreover they have to be able to stick to a plan once they have made it. Once they get to the carcass they have to be able to refrain from fighting. And each has to know that the other has this degree of resolution. Though the plan is simple it requires quite a lot of thought.

The three-person game with Cain, Abel, and Adam could be written out as a fairly complicated-looking matrix with a PD like that in section 1 embedded in it. The complex action of striding and sharing would be an equilibrium, in fact a stable equilibrium. (To get everything to work out exactly right some more details would have to be specified. I am assuming the main point is easier to see without them.)

Cain and Abel have in effect managed to use Adam's presence as a way of making an enforceable contract between them. It is interesting how little of Adam's real physical presence is needed. A mythical Adam would do almost as well.

5. stratification Agents can form coalitions in the presence of third parties in ways that enforce cooperative behavior. In this final section I shall argue that we have here a central feature of strategic thinking in all primates, and one which marks human social organisation in characteristic ways.

You are a subordinate male and you want to displace the dominant male, who is distressingly more powerful than you. So you make an alliance with [the help of] another subordinate by grooming, sharing food, and other small cooperative acts. Then in the showdown with the boss you call for help and (all going well) get it. You can make it more likely that you will get help if you advertise to the dominant male your alliance with your confederate, as it makes it more likely that [he/ the confederate] will be also be the object of his aggression.

Or, you are a female and can see a future when you are going to have trouble feeding and protecting yourself. You have just given birth, and you raise your infant in such a way that she is devoted to you and very likely to take care of you later. You can make it more likely that she will remain bonded to you if you make common cause with her in bettering and defending your and her status.

In the first of these cases the two males transform each other by forming an alliance, in the second the mother transforms her daughter in the ways that parents have always transformed children. The mechanisms at work here are all evolved primate features. They are all shortcuts for strategic thinking, necessary for creatures which live by social strategies in which everyone has to take account of everyone else's plans, but who have limited powers to think through the resulting complexity. Thinking in terms of coalitions and bonding-inducing transformations seems to be a basic element of the primate approach to this problem. Let me put the problem in perspective.

All primates live deeply social lives in that each lives their life in the company of others and at any moment is aware of the activities of others. In any primate group at any time there is a lot of watching going on, and a lot of it is triadic or more: *a* is keeping a careful eye on what *b* is doing to or with *c* (or with *c* and *d* and *e*, etc.) These n-adic trackings are cognitively demanding and account for part of the size of primate brains. They obviously set the stage for strategic thinking when we introduce any capacity to anticipate the actions of others. For if *a* is forming an expectation of *b*'s action based on her information about b, that information will include the fact that *b* is keeping track of others, who typically include *a*. So *a*'s expectations about *b*'s actions must take account of the fact that *b* is also forming expectations, probably about *a*. (See chapter 13 of Byrne, 1994, chapters 3,4,5 of de Waal, 1996, Dunbar 1988, chapters 2 and 3 of Dunbar, 1996, Smuts 1986.)

There is a clear ratchet factor here. When an increase in the power to anticipate others in such a context becomes the norm it makes the job of anticipating harder, necessitating more anticipatory power, and so on. We can expect then that as primate social life becomes more complex individuals will usually be at the limits of their cognitive capacity. In particular, the number of individuals whose possible strategies they can anticipate will be limited by memory and processing limitations. (With n agents each of whom has m possible actions there are m^n cells to the matrix - the complexity you have to think through rises as the power of the number of agents involved.) So individuals will be torn between on the one hand the advantages of co-ordinated actions, which increase with the number of coordinating individuals, and on the other hand the danger of getting into unmanageable complexity, which also increases with the number of individuals.

There is an obvious solution to the problem, which marks all primate social life from baboons to humans: stratification. Keep the coalitions no larger than required for the activity in question. Engage with a minimal number of individuals for the required purpose, while having a greater number of individuals available in case of need. To make this a bit more precise suppose that the social tasks individuals in a species face come in kinds, and that for each kind the benefit to the individual in cooperating with others is linear in the number of individuals involved - so the benefit from cooperation with n others in task i is given by $b = k_i n$ - while the cognitive cost is exponential in the number involve - so the cost of interacting with n others is given by $c = K m_i^n$. (Perhaps m_i is determined by the number of actions each individual can perform in task i, as suggested above.) Then the advantage of approaching a task socially will be $b-c$, which will increase up to a crucial value of n (which will depend on k_i, K and m_i.) and then decrease rapidly. So we can expect a stratification of social groupings, each no larger than it needs to be for some purpose.

There are two predictions here, which should hold across primate species. First that the number of individuals cooperating on any task will be the minimum required for the task, so that we will find groupings of a number of discrete sizes, depending on the kinds of tasks and challenges that the species faces. And second that cooperation between individuals will be shaped by specific devices for coalition formation, which will tend to be specific to a small range of tasks. The second prediction reinforces the first, in that the coalition-forming devices will tend to form alliances between small numbers of individuals. Since such devices evolve slowly, they will typically lag behind the general capacity for strategic thinking. So there may be more cooperation among smaller numbers, though with less flexibility in its aims.

And this is what we find. As Dunbar writes, introducing the general topic of primate social life:

> ... a local population consists of a large set of animals who have relationships of varying intensity with each other. Certain individuals will associate with and interact frequently with each other. Certain individuals will associate and interact frequently, others only rarely, and in yet other cases the relationships may be purely antagonistic. ... Where the sets of relationships of several individuals coincide and are reciprocated, a formal group emerges as a stable cohesive unit.

Such groupings can be created at many different levels within the same population at the same time. [In the gelada] the patterns of association between individual animals reveal a hierarchically organised series of clusters at five levels. [coalition, unit, team, band, community, population].

An analogous series of hierarchically structured grouping patterns has been described for the hamadryas baboon. ... In this case, definable groupings occur in the form of one-male units, clans, bands, and sleeping troops. These groupings are similar to those of the gelada only in respect to the functional significance of one-male units and bands. Otherwise, they differ significantly from those of the galada in size, composition, structure, dynamics and function. (Dunbar, 1988, p 12.)

The specific pattern of groupings found in any species will be a response to the kinds of tasks and problems that individuals in that species face, and the kinds of cognitive resource that they can bring to them. Dunbar later lists four very general categories of reasons for group formation: protection against predators, defence of resources, foraging efficiency and improved care-giving opportunities. (See Dunbar p 106.) Different species will have different strategies and face different problems in each of these areas, and the optimal social groupings will thus vary from species to species. Remember though the main thrust of the argument of this paper: there is an optimal way of solving strategic problems, namely by identifying equilibria, and there is a unique cognitive strategy for doing this, namely by ascribing preferences to individuals and working out their possible reasoning. If there is anything to this, we should find a mutual pressure between increasing cognitive capacity and the development of routines for thinking through social problems in terms that are more and more explicitly along the lines of this optimal strategy.

One might apply this theme in a definitely speculative way to the comparison of human and chimpanzee social life. With chimpanzees the pattern of social grouping seems to be that individuals form bands which are parts of larger tribes. Bands are defined in part in terms of matrilinear descent: most members of a band are some sort of mother's-side cousins. Beneath the level of the band there is a formation of constantly shifting groups and parties, which have very short-term and task-specific nature. In human societies on the other hand we have stable families, smaller in size than the next grouping upwards (which tend to be larger than chimpanzee bands).

Thus for long-term groupings the smallest human unit is smaller and the next unit larger than with chimpanzees.

We can explain this if we assume that the larger and the smaller grouping serve different purposes and are made possible by the solution of different strategic problems. Assume that the larger grouping is a general-purpose one requiring coordinated action. The paradigm strategic problem at this level would thus be one in which most matters is that everyone perform the same one of some subset of the available actions (travel in some direction; act aggressively, evasively, or appeasingly to some threat; follow this would-be leader or that one.) Assume that the smaller grouping serves different purposes, being focused on a few specific cooperative activities (such as the provision of food and the raising of young). The paradigm strategic problem at this level will be prisoners' dilemma-like, and it will typically be solved by forming some sort of coalition. If we make these assumptions, we can conclude that the greater cognitive power of humans will lead to greater group sizes at the larger level and smaller at the lower level.

The larger groups will be larger because the problem they present is typically one of keeping track of the benefits or losses to a number of individuals of a number of actions. The best coordination - the equilibrium for this kind of situation - is a pattern of action in which everyone gains if everyone conforms to it, and everyone loses if substantial numbers deviate. The limits to the number of agents with whom one can coordinate are thus defined by the limits on the number of outcomes that one can distinguish. In effect, by short-term memory. The more that can be distinguished and kept in mind, the greater the number of agents that can act in concert.

The smaller groups will be smaller because the problem they present is typically one of anticipating the patterns of motivation of others. For example if one fears the possible betrayal of another in a forthcoming confrontation with a third then the conflicting effects of loyalty, desire for future alliance, and the benefits of betrayal, all have to be weighed against one another. If one can imagine and ascribe more in cases like this, solutions become harder rather than simpler. As a result though it is true that there are many-agent situations which a more intelligent agent can handle although a less intelligent agent can only handle the analogous situations for fewer agents, it is also true that there are simple problems where intelligence complicates coordination. There are many situations in which two agents can achieve a mutually profitable coordination as long as neither realises a subtle motive that the

other might attribute to them: only creatures that can entertain quite complex thoughts will find their attempts at coordination blocked by such considerations.

Another factor tends in the same direction. With increased cognitive power a smaller group can sometimes do what needs a larger group given less intelligence. The crucial example is the formation of coalitions. As we saw in the last section the presence of an additional agent can be exploited in a variety of ways to bind a coalition. The extra presence is just a device, and the more one can see ahead, communicate, or make promises, the less need one has of it. So we can expect that in some situations more cognitive capacity will mean that larger groups are not necessary.

The other cause of the smaller size of the smaller grouping is the different functions of the smaller grouping in the two species. In chimpanzees the coalitions are directed at short term food gathering and mating aims. In humans the aims are longer-term, since they involve raising vulnerable human children through their longer development. These longer term aims require more delicate balancing of one short-term aim against another, and thus more thought. A larger grouping of people might perform these tasks more effectively, but it would require a social cognition that is beyond us. So our more demanding core social needs keep our smallest grouping fairly small.

(I am omitting an absolutely crucial factor: the interaction of strategic thinking and theory of mind. Some patterns of strategic thinking are only possible if one can articulate the thinking the other agents may be employing. For example problems of pure coordination, such as those I associated with the larger groupings, are less demanding of psychological cognition than problems of coalition formation, which I associated with the smaller groupings. Coordination requires that one think about outcomes and preferences, while coalition requires that one think about patterns of reasoning, which requires much more psychological sophistication. So rather than talk of human strategic thinking as being both facilitated and complicated by an amorphous cognitive capacity, it would be more helpful to consider how the development of theory of mind facilitates the transition between the 'hard wired' and the 'ideally rational' poles of the previous section. Strategic thinking and theory of mind are two sides of the same coin, I believe, and will argue in my forthcoming *Folk psychology as ethics*. However in this volume Carruthers and Chamberlain, chapter 1, and Dunbar, chapter 4, argue for related conclusions.)

This is speculation: neither the phenomena explained nor the premises used to explain it inspire total confidence. More important is the general primate strategy of producing cooperation through coalitions and of basing coalition-formation on bonding processes linked to very specific tasks. We humans clearly do live our lives between social groups of several different sizes, serving different purposes and relying to different degrees on general strategic thinking and specific coalition-inducing factors. There are good reasons for expecting that there will be such a stratification of group activity in any creature that approaches strategic problems in the general way that we and other primates do. And there are good reasons for believing that the optimal sizes of these groups will be sensitive to the cognitive capacities of the animals concerned. How much fixity is there to the sizes of human groups and the mixtures of cognition that we can successfully bring to them? Important as the question is, we really do not know.

Byrne, R. 1994. *The thinking ape*. Oxford, Oxford University Press.

de Waal, F. 1996. *Good natured.* Cambridge, Mass., Harvard University Press.

Dunbar, R. (1988) *Primate social systems*. London, Chapman and Hall.

Dunbar, R. 1996. *Grooming, gossip, and the evolution of language*. Cambridge Mass., Harvard university Press.

Gould, S.J. 1989. *Wonderful life: the Burgess shale and the nature of history.* New York, Norton.

Luce, R. D. and Raiffa, H. 1957. *Games and decisions*. New York, Wiley.

Myerson, R. B. 1991. *Game theory: analysis of conflict.* Cambridge, Ma., Harvard University Press.

Skyrms, B. 1996. *The evolution of the social contract.* Cambridge, Cambridge University Press.

Smuts, B. and others 1986. *Primate societies.* Chicago, Chicago University Press.

Sterelney, K. 1995. Understanding life: recent work in philosophy of biology' *British Journal for the Philosophy of science* 46, 155-184.

Sugden, R. 1986. *The economics of rights, co-operation, and welfare*. Oxford, Blackwell.