Psychology for Cooperators

Adam Morton

in Christopher Morris, ed. *Practical Rationality and Preferences: essays for David Gauthier* Cambridge University Press, 2001, 153-172.

1. Cooperation and psychology Assume that human beings must cooperate to survive, and more so to flourish. Activities which require a variety of links between our actions are essential to the full range of human life. It follows that humans need to be able to describe and categorise actions and to attribute to one another motives: belief, desire, character. They need a psychology.

Cooperation without psychology is possible for other species, with hard-wired social routines that tell them when to share, when to defer, and when to punish. We are innately social, but we do not have a fixed repertoire of social acts with fixed instructions when to perform them. Instead we have inescapable desires for company, affection, and attention from others, and an inbuilt tendency to think out courses of action in terms of the relations we and others have to common features of the environment. That is our evolutionary niche: to operate in groups but to think our way through the problems groups face. (For psychological and evolutionary evidence for this diagnosis see Chapters 8 and 9 of Byrne [1995], and the first three chapters of Baron-Cohen [1995].) Each person thinks what to do, but must do so strategically, taking account of the decision making of others. Strategic thinking is impossible without concepts to represent the paths of reasoning that lead from motives to acts and outcomes. (It need not use the concepts of "reasoning", "motive", "act", "outcome" and their friends, but it must use concepts which represent reasoning, motive, act, and outcome.) So it needs psychology.

People learn psychology, largely as children from older people and largely in the form of the doctrines, habits, and cognitive tricks of their cultures. We now have reason to believe that human beings are adapted to pick up such folk psychologies, and it is a natural conjecture that there is some relation of mutual support between the psychological conceptions current in a culture and its patterns of cooperation and interaction. (Teasing out this relation is a focus of my current research.) Actual human cultures clearly sustain an imperfect degree of mutual cooperation between their members, and there is most likely a very imperfect fit between their norms and their folk psychologies. The question this paper asks is: what would constitute a perfect fit? What conception of action, motive, and outcome might be used by idealised rational agents maximizing their individual good by thoughtful cooperation?

2. Conventional psychology Each agent possesses a folk psychology, a selection from the space of possible psychologies, which she uses to interpret, explain, and predict the actions of others and of herself. If a number of interacting people have the same folk psychology, and no one of them could not do better by possessing a different one, given that the others possess the psychology in question, then the psychology is theirs by convention. I will call this the conventional psychology of the people in question. More precisely, a conventional psychology for a set S of people consists of a scheme P for supplying explanations for actions, such that for each person p in S the two marks of a convention apply. (a) if most other people in S subscribe to P then it is in p's interest also to subscribe to P. (b) if p subscribes to P then the more other people in S subscribe to P the more it is in p's interest also to subscribe.

(a) and (b) make adherence to P a kind of convention, in the spirit of Lewis [1968] and Sugden [1986]. (b) is important in making the situation conventional, as it requires that it be in each subscriber's interest that others subscribe. A mere equilibrium needs only an (a)-like condition, that it be in each person's interest that she subscribe, if others do. One might choose a stronger definition, in which "most" was replaced by "all". But while that would fit the definition into the tradition of game theory better, it does not seem realistic. One might also choose a weaker definition, without (b). Then the psychology would occupy an equilibrium position which was not a convention. One reason for such a weaker definition might be that it might be in some person's interest that others have a different psychology, for example one that did not give them the resources to see what the person was planning. But I shall assume that the dynamics of social life will force the psychology adopted by a coherent social group to be a convention. In effect, if there is a psychology which it is in most people's interest that most other people subscribe to, then they will find a way of making most people subscribe to it.

The definition leaves basic things unspecified, also. It does not define adherence to P, nor specify how P produces explanations of actions, let alone what is to count as an explanation. And it does not ask what is to count as being in a person's interest. (In particular, it neither asserts nor denies that adherence to P may change what is in a person's interest.)

These unspecified factors are going to be left unspecified in this paper. However it is essential to say more about the use to which explanations of actions are put. The relevant uses center on the capacity of a person to think about the motivation of another. Suppose for example that two people find themselves in a situation described by a pay off matrix such as the one below. I will call it Domcord

I, because one agent is reasoning by dominance, choosing acts which the other has reason to coordinate with.

		B	
		b_1	b_2
A	a_1	2,0	0,1
	a_2	0,0	2,2

Domcord I

Here B can think through his actions entirely in terms of the possible outcomes. And in terms of them b_2 is the obvious choice. (Throughout this paper agent A will be female and agent B male, pairs of utilities will be in the order (for A, for B).) A, on the other hand, has no dominant choice. Her choice is undetermined until she takes account of B's situation. Then it will be clear to her that B will choose b_2 and that she therefore should choose a_2. The use of a psychology lies in the difference between B's situation and A's. To make her choice A has to think in terms of the reasoning that may lie behind B's choice: she has to see her not just as choosing options but as choosing them for reasons. (Many situations - most real situations - are thoroughly strategic, in that each interacting agent has to take account of each other's reasoning, and in fact very often of the reasoning by which others take account of their own and others' reasoning. And so on.) The most basic requirement on a conventional psychology is that it allow agents in situations like A's in Domcord I to go through reasoning like A's.

The means first of all that outcomes and actions must be identified. There are no intrinsic pausing points in the ongoing flux of events, except perhaps the death of each agent. So in thinking of choices in the terms used above we are already assuming that we can impose a conceptualisation on the world that marks some possible effects of our actions as salient. This is not yet psychological. Psychology enters when an agent thinks of another agent as carving the world into a specific set of outcomes and actions. Very often agents think of others as using the same conceptualisation that they do. (Presumably this is the default for real humans.) Many coordination problems would be very hard to solve were this not the case. As David Gauthier has observed (see also Sugden [1995]), it is very often the case that

> We must .convert our representation of the situation into one with but one best equilibrium. This restriction in our conception of what we may do, far from being disadvantageous, is what makes successful coordination possible. (Gauthier [1975], p. 210)

The situation can become very complicated when there is a large number of acts and outcomes and each agent has to think which outcomes are the objects of consideration by the other. The if the other is not aware of an outcome then in a coordinative situation there is no point considering it oneself. (You and I are both approaching the intersection as the light turns yellow for me. My obvious options are to rush through the intersection or to stop. But I could also do a U-turn, reverse back up the road, or get out and do a dance on the car's roof. If you consider all the things I might do you'll never decide what you should do.) More subtly, even when another agent is in some sense aware that an option is open to her it is important to know how she thinks of it. Is it thought of as one action or a set of actions, is it thought of in terms which make it salient, or make it likely that she will think it is salient to you?

On the other hand it is worth noting that there are situations in which it is in an agent's interest not to reason in terms of the outcomes in terms of which she understands the other person's reasoning. Consider for example the situation below.

			B	
		b_1	b_2	b_3
	a_1	22	20	10
A	a_2	04	33	11
	a_3	12	11	11

If we suppose that both agents are considering all three options then A knows that B will choose b_1, and she will therefore herself choose a_1. But suppose instead that the option b_1 is not in B's list of things he might do, and A knows this. Then B will choose b_2. Knowing this A will choose a_2, so that the resulting situation (a_2,b_2) will be better for both than the result (a_1,b_1) which will result when B knows she can choose a1. (Sound familiar? The issues return in sections **4** and **5**.) In practical terms, this can be described by saying that it is sometimes best to pretend you don't know an option exists. More abstractly, this suggests that one conventional psychology might sometimes be more in the interests of its adherents than another when it fails to provide a description of some actions. A less rich conceptualisation might sometimes have advantages.

A conventional psychology, then, will provide characterisations of acts and outcomes, and will allow agents to relate one another to the same act and outcome-types when it is in their interest to. This is only the very beginning of any psychology, though. Situations like Domcord 1 (above) require participants also to consider the patterns of reasoning that others may be following. We - people like this book's contributors and readers - naturally think of this in terms of specific kinds of interaction between preference orderings and degrees of belief. But of course not all people think through their strategic situations in these terms. It is not obvious that it is sensible for all people so to think, that it forms part of a conventional psychology for them. And there is the worrying thought that there might be better ways of thinking out strategies. There might be better ways of describing them and thinking out what may pass through ones own and other people's heads, that are even less related to the folk psychology of our culture or the psychologising tendencies of humankind than game theory is. The aim of the next section is to show that there are constraints on the kinds of conventional psychology that extremely rational agents will use. They are unlikely to be as exotic as this paragraph might suggest.

3. Game theory backwards Assume that we have agents who are optimally adapted to their social and physical environment. That is, each agent has a consistent set of preferences and in each situation there is a best, or co-equal best, preference-maximizing choice for each agent, which that agent will choose. Many of these situations are strategic, in that the best choice for an agent will be a function not just of the agent's preferences and the physical facts but also of the choices of other agents, which may themselves be functions of the choices of the first agent. Assume that agents' choices are determined by facts about situations, and that agents can represent choices and outcomes, and can carry out quite complex conditional thinking ("if A or B happens then as long as C does not happens the result will be D" etc).

Assume also that agents can think about possible situations and what it would be best to do in them.

But do not assume that the agents determine the best choice in terms of the preferences or reasoning of other agents. Simply assume that they possess some way of determining the optimal choice in each situation. It might consist of a mystical infallible oracle, or it might consist of an ability to pick up cues from the behavior of others that contingently correlate with optimal choices in situations involving them. Or it might consist in some knowledge of the nervous system that cannot be expressed in terms of preference and decision. The aim is to show that even such oracle-guided agents will have the capacity to think in terms of the preferences and reasoning of others.

Consider an agent A confronted with a situation S in which a number of acts a_i are available to A and a number of acts b_j are available to another agent B. The agent represents one act a* as choice-worthy. A's choice may reveal nothing about the preferences or choice of B. This will be so when a* may be determined by considerations of dominance, that is, when for each b_j that B may choose a* is the best choice for A. In this case B's preferences are irrelevant to A. Thinking - as we may, though not as A does - in terms of the situation as a game in normal form we can determine a* in terms simply of the pay-offs to A of the given outcomes, without considering those to B.

But this is not usually the case. Usually if you know what the optimal actions are then basic facts about the preferences of the agents are determined. Given a strategic situation- which we can think of in terms of the preferences of the agents, but which they represent in some other way - agents have a way of knowing what is the best action for each participant. Then very often agents can deduce what one another's preferences must be.

The simplest typical cases are those like Domcord I, whose matrix was given above. There the optimal choice for A depends on B's choice, which is itself determined by dominance. Thinking strategically, A must take account of the reasoning that B's preferences will lead to. Thinking in terms of her oracle, A simply gets as a datum that her best action is a_2. But she can reason from this datum, with a little further help from the oracle. First of all, she can know some complex facts about her situation that we would express by saying that her own choice is not dictated by dominance. And in fact she can grasp facts that amount to representing some of her preferences. Let us see why this is so.

A can consider a variant on the actual situation in which b_1 is available to B, but both of a_1 and a_2 remain available to A. (She can ask the oracle: suppose that I was in a situation like this except that... .) This reduced situation would have had the two outcomes resulting from the product of $\{a_1, a_2\}$ and $\{b_2\}$. And the oracle would tell her that the right choice in this sub-situation is a_2. Similarly it would tell her that the right choice for her, if b_1 but not b_2 had been available to B, would be a_1. But the fact that the right choice is different in these two sub-situations can be taken to represent the fact that A's choice in the whole situation is not dictated by dominance. (For us, it is that fact.) And the fact that the right choice in the first of them is a_2 can be taken to represent the fact that (a_2,b_2) is preferred by A to (a_1,b_2), just as the fact that the right choice in the second is a_2 can represent the fact that (a_1,b_1) is preferred by A to (a_2,b_1). So she can learn something that is in effect a translation into the language of optimal choice of what we would express in terms of her preferences.

By pushing the same reasoning a bit further A can know similar facts about B's preferences. Since A does not have a dominant choice and yet there is a single best action for A this must be because one slice of possibilities is eliminated by B's choice, and thus (this being a game of 2 agents with 2 acts each) there must be a dominant solution for B. This is either b_1 or b_2. If it were b_1 then in order to have chosen b_2 A would have to have preferred (a_2,b_2) to (a_1,b_2), which is not the case. Therefore B's dominant solution is b_2, and (extending the reasoning) B prefers (a_1,b_2) to (a_1,b_1) and (a_2,b_2) to (a_2,b_1). Or that is how we would put it, at any rate. A simply realises these complex facts about B's actual and possible choices, which are extensionally the same as B's preferences.

Several things are worth noting about the procedure so far. First of all, there is a symmetry between ascription of preferences to oneself and to others: one starts with situations and the optimal actions they determine and then one figures out what constraints this puts on the states one can ascribe to all parties involved.

Second, these considerations provide a defining condition for preference: an agent prefers one outcome to another if given a choice between an act producing the one outcome but not the other, and an act producing the other but not the first, the first act is the best for her. This is a very constrained definition: it says that if some conditions are met then an agent prefers the one act to the other. Most often the conditions are not met. The more complicated preference-eliciting procedures below can produce more complex defining conditions.

Third, this reasoning applied to these situations only results in a partial determination of the agents' preferences. Thus in Domcord I all that is required of B's preferences is that he prefers (a_1,b_2) to (a_1,b_1) and (a_2,b_2) to (a_2,b_1). The orderings 'in the other direction' between (a_1,b_1) and (a_2,b_1), and between (a_1,b_2) and (a_2,b_2), are irrelevant to B's choice of action. There are two distinct aspects to this underdetermination. First, some preferences, such as B's ordering of (a_1,b_1) and (a_2,b_1), are simply left unspecified. But, also, even when the ordering of outcomes is determined, there is incomplete information about their relative degrees of desirability. Thus in Domcord I and surprisingly many other situations we can characterise agents' preferences in terms of the two-unit scale 1,0, or "want/don't want". In terms of this, we and they can know all they need to about their best choices. And as a result, reading back to preferences from best choices in many situations the finest grid we can impose on them is "want/don't want".

The two underdeterminations are closely linked. The greater the number of preference-comparisons that can be made between outcomes, the finer the grid that can be imposed on their relative degrees of preference. But even starting from a comparatively simple situation such as Domcord I we can deduce somewhat more about the agents' preferences if we allow ourselves more elaborate reasoning. Suppose a situation in which act b_2 (the same b_2) is available, but only if a1 is not, and in which b_1 is available, but only if a_2 is not. (Another way of describing it: B has b_1 and b_2 available, and A has the acts (a_2 if b_2) and (a_1 if b_1) available. Or. equivalently, B has available b_3 and b_4. If B performs b_3 then b_1 becomes available to B and a_1 to A, and if B performs b_4 then b_2 to B and a_2 to A.) If in this situation b_2 is the best choice for B - as it is - then (a_1,b_2) is preferred by B to (a_2,b_1).

The underdeterminations are not surprising. After all, Domcord I is a very simple situation, and so it can be expected to yield only limited information about the agents' preferences, even over the actions found in it. So one might expect that by considering more complex strategic situations we should be able to impose a finer triangulation on preferences. This is indeed the case.

Consider what happens when we add another option for each of the participants, to get Domcord II.

			B	
		b_1	b_2	b_3
	a_1	3,0	0,2	0,1
A				
	a_2	0,0	2,1	2,2

Domcord II

This situation is to be seen through B's eyes. Thinking in terms of preferences and reasoning, B's decision is made as follows. b_1 is dominated by (b_2 or b_3). But the choice between b_2 and b_3 cannot be made in terms of B's preferences alone. On the other hand A's preferences make her coordinate with B: if she expects B to choose b_1 A will choose a_1, and if she expects B to choose (b_2 or b_3) then A will choose a_2. A knows that B will not choose b_1, and thus A will choose a_2. Knowing this, B's preferences are coordinative, leading him to choose b_3.

In deciding what to do B has to consider A's reasoning about his reasoning. Some of the effect of this reasoning about reasoning about reasoning can be recaptured by thinking backwards, from choice to preference. Suppose, again, that B has the use of an oracle which announces that b_3 is the best choice for him. Then knowing that b_3 is his choice B can know that (b_2 or b_3) dominates b_1, by reasoning like that used above. B knows that if A's best choice were a_1, b_3 wouldn't be his best choice, and so B knows that A's best choice is a_2. B can know that this choice is not dominant for A and so it must be motivated by coordination. So A must prefer (a_1,b_1) to (a_1,b_2) and (a_2,b_2) to.(a_2,b_1).

So B can ascribe 'horizontal' preferences to A, in effect preferences between the results of B's actions. Coordinative aspects of situations elicit preferences between results of the other person's actions, and dominance aspects elicit preferences between results of ones own actions. Note also that the reasoning that B applied to Domcord II just above reveals some of A's beliefs as well as her preferences, since A's choice of a_2 reveals her belief that B will choose b_2 or b_3, and thus that neither (a_1,b_1) nor (a_2,b_1) will occur. In fact, this can be taken as giving the practical content for B of A's belief about these outcomes.

It is not hard to construct a Domcord III in which domination is embedded within coordination within domination (just as in Domcord II coordination is embedded within domination), and in which an agent's choice of action depends on another level of embedding of motives. And in this situation even more of the structure of the agents' beliefs and preferences is revealed. And so on, with increasingly complex situations and - unfortunately - increasingly complex reasoning inverting the usual game-theoretic considerations to deduce from what acts are optimal for agents what their preferences are. The conclusion is that given a sufficiently rich variety of strategic situations in which an agent has knowledge of her best action, the preference and belief structure of that agent and others interacting with her are determined. If an ideally rational agent equipped had an oracle which told her the optimal actions in all possible strategic situations involving herself and a given other, then she could deduce all the facts about the other’s preferences which would be relevant to determining those optimal choices.

I believe this is a very significant conclusion, with consequences in the philosophy of mind. (Eliminative materialists should take note of it, for it says that if you can make good choices in strategic situations then you can ascribe beliefs, desires, and thinking; and thus have the core of folk psychology.) One intriguing aspect of the analysis is that attribution of preferences to self and other are interdependent: you have to conceive of others as having preferences in order to attribute them to yourself. Another is that attributions of more fine-grained preferences to self and other depend on considerations about embedded reasoning, about one person's thinking about another person's thinking about the first person's thinking, and so on. One might think that attributions of even very finely grained preferences were conceptually more primitive than such involuted thinking. But the patterns of reasoning described in this section suggest the possibility that the attribution of rich preferences may depend on that of complex reasoning about reasoning.

These are suggestive possibilities in the philosophy of mind. But for the argument of this paper the important conclusion must be that agents making optimal choices are in a position to ascribe beliefs and preferences to one another. The minimal objects of these ascribed states are acts of interacting agents and the consequences of combinations of those actions. Thus we can expect that a conventional psychology that gives its subscribers the resources to make good decisions will also allow them to ascribe beliefs, desires, and reasoning to one another.

4 Harvestless in Humeston The aim has been to see what kind of a psychology would best aid people negotiate the problems of strategic interaction. A standard understanding of 'best' has led us by an interesting route to a fairly unsurprising bottom level of such a psychology. It will describe actions in terms which people can generally share, and it will allow people to attribute to one another beliefs about actions and preferences between actions and outcomes. That is obviously only the very bottom level; any workable set of psychological ideas will be much richer, and one would expect it to facilitate the interactions of its subscribers in more complex and subtle ways. In particular it would be natural to consider whether a shared psychology can bring advantages in connection with those interactions in which the pursuit of individual self-interest seems self-defeating. The obvious test cases are situations like the prisoner's dilemma where a non-equilibrium outcome is intuitively something the agents ought to be able to aim for.

Questions about how agents can cooperate for mutual advantage are often framed in terms of rationality. Then we ask whether it is, for example, ever rational to cooperate in a prisoner's dilemma. I have not used the R word in this paper until this paragraph. I have spoken of what is in agents' interests, and of how they can get more of what they want. I am suspicious of theories that put too much weight on what is or is not rational. I doubt that the concept of rationality can bear the weight. My aim in the rest of this paper is to show how a conventional psychology can ease the route to cooperation. But psychology is not needed if logic will do the job. David Gauthier has famously argued that as long as agents are equipped with the psychological concepts of the previous section they ought to be able to see that it is rational of them to act cooperatively, under suitable circumstances. (See Gauthier [1986], especially chapters V and VI.) I agree, but I think the claim is best understood as one about conventional psychology rather than about rationality. I must explain why this is so.

The simplest characterisation of rational action is as action which will on average give optimal results. So if it is sometimes rational to cooperate then it should follow that cooperators will sometimes do better than non-cooperators. There are two crucial ambiguities here. Optimal by what measure? And averaging over what range of possibilities, with what weighting? Standard choice theory answers both these questions by relying on the agent's own states: optimality is measured in terms of satisfaction of the agent's preferences, and the averaging is over all states to which the agent assigns a degree of belief, weighted by those degrees of belief. These are clearly not the only ways of resolving the ambiguities, and clearly they raise problems of their own. The degrees of belief are normally inaccurate; the preferences may be bizarre. (So why should we define 'optimal' in terms of an averaging weighted by false probabilities? Or giving points to outcomes that will make the agent miserable.) Even more threateningly, both preferences and beliefs are for anything like a real person defined on only a tiny subset of the range of possible future outcomes. So it is almost inevitable that the expected value of an act, in terms of the agents' own preferences and beliefs may be utterly different from the average outcome for the agent. (And it is almost inevitable that on many occasions agents doing the 'irrational' thing will fare better than agents doing what the theory calls rational.)

But these problems fade into insignificance when we move to strategic choice. For there the crucial factor affecting a choice is the choices made by others. And an agent has no degrees of belief about these. Degrees of belief about other agent's choices cannot be data for strategic choice for two reasons. First, the agent's reasoning centers on forming conclusions about what the other agents may do. Once you get to that point the decision becomes a standard non-strategic one. And second, there is a symmetry between interacting agents. Each is trying to conclude what the other may choose: so if you could start with probabilities about their choice they could start with probabilities about your choice, and knowing this you would have probabilities about your choice, and could spare yourself the effort of reasoning.

So the definition of optimality for strategic choices is deeply problematic. It is hard even to understand what is meant by the choice that will on average give the best results. Realising this should help to prevent us being too deferential to game theoretical solutions to real life problems. But it should also prevent us from thinking that we can solve practical problems simply by deciding what we are to count as rational.

To illustrate both these points, consider an example derived from Gauthier [1994] derived from Hume. It is not a prisoner's dilemma but a situation that I, at any rate, find rather more disturbing and thought-provoking. Analysing it while trying to avoid pointless disputes about what might or might not count as 'rational' we find ourselves led to conclusions about what psychologies agents might profitably use to describe one another.

Dave and Tom are graduate students at Pitttsburgh. Jobs in philosophy are scarce and so they leave to become farmers in Humeston, Iowa . (It exists, not far from Princeton and Pleasantville.) They buy neighbouring farms so they can talk philosophy on winter evenings. One year Dave plants X and Tom plants Y. X needs to be harvested a month earlier than Y, and each crop is best harvested by two people. So Tom proposes that they each help bring in the other's harvest. Dave likes the idea, especially since this will be his last harvest: as a result of their winter discussion he has published an article on dynamic choice which has attracted so much attention that he has been offered a tenured position at the U of Hawaii. Knowing that there will be no more Dave & Tom harvests leads Tom to some sober reflection. His motive for helping Dave will be the expectation that Dave will in turn help him. But what motive will Dave have for helping him, since Tom will no longer have the sanction of withdrawing his help? It seems that Dave has no reason to help Tom. But if that is so Tom has no motive to help Dave. So they will each bring in their harvest alone.

Tom's conclusion is uncomfortable not just because they seem headed for more difficult harvests. That is the practical problem, but there is also an intellectual difficulty. Dave seems to lack a motive for helping Tom next month. So he will not, and so Tom will not help him this month, and so he will be worse off than if he had had a motive to help. This does not seem right: surely the fact that he would be better off if he had a motive gives him a motive. (This reasoning presupposes that the cost of helping is less than the benefit from being helped. For one set of utilities bearing this out see Gauthier [1994], p 714.) In response to this thought Gauthier presents a very subtle analysis of the conditions under which it would be rational for someone in a situation like Tom's to commit himself to an action like that of helping Dave. Gauthier summarises his discussion by saying.

> ...[S]ometimes my life will go better if I am able to commit myself to an action even though, when or if I perform it, I expect that my life will not thenceforth go as well as it would were I to perform some alternative action. Nevertheless, it is rational to make such a commitment provided that in so doing I act in a way that I expect will lead to my life going better than I reasonably believe that it would have gone had I not made any commitment. As a rational agent I.shall be able to offer and honor assurances when it is advantageous for me to do so. (Gauthier [1994], p 707)

Suppose that this is right. Suppose that it is rational for Dave to commit himself to helping Tom and then actually to help on the day. How much of the puzzle does this remove? Suppose that you are Tom and it is the night before Dave should arrive to help you, in return for your help of a month before. Suppose that you are convinced, from your studies in Pittsburgh, that it was rational of him to promise and will be rational of him to keep his promise. Does this allow you to sleep easily, confident that he will be there in the morning? I must confess that if I were Tom I would pass a sleepless night, worrying what my philosophical friend would decide to do. And my sleep would not be helped (much) by the conviction that it would be rational for Dave to be there. This conviction would not help me as long as I could imagine Dave calculating what course of action it would pay for him to take, and then considering taking it even if it were irrational.

Or imagine that you are Dave, wondering whether to turn up for Tom's harvesting. You have nothing to gain from it; nevertheless you believe that it would be rational to help, given your earlier commitment. Then the thought strikes you: sometimes one can gain from doing an irrational action. You remember your aunt Florence who set fire to the office of a salesman who had sold her a defective truck. She didn't get her money back on the truck; in fact she had to pay a substantial fine. But the fact that no one could predict her occasional uncalculating fits of self-expression meant that few people tried to cheat her. Perhaps not helping with Tom's harvest would be a profitable irrational act. If rationality were a matter of maximization then the cases where irrationality pays should be fairly exceptional. But Pittsburgh taught you that rationality is more complex than that. So you waver;

you spend the night in internal debate. It would be irrational not to help Tom, but perhaps it would be the most profitable, most sensible, or simply the best thing to do.

The general problem is that an analysis that cuts the link between what is rational and what will generally turn out best for an agent also cuts the link between thinking that something is rational and seeing a reason to do it. And as a result it cuts the link between thinking that it would be rational for someone else to do something and expecting that if they reflect enough they are likely to do it. These are vital links. They are central to the use of the concept of rationality. In particular, they are central to its psychological use, in anticipating the reasoning and actions of others. That *psychological* aspect is vital here. Agents need to have information about one another that gives them reason to be confident in the cooperative or uncooperative aspects of their future actions.

What information could Tom and Dave have about one another that would give them the confidence they need? That must depend on what the facts are about Tom and Dave. Suppose that their minds are such that there is some act that each can perform at one time which will make it impossible for them not to perform a given action at a later time or under specified conditions. Call this "strong binding". (See Kavka [1983], Bratman [1986], McLennen [1990].) Then their problem is just knowing that strong binding has taken place. To know this they will first of all need to have the concept of binding: they will need a conventional psychology that describes it adequately. Then they will need to have seen adequate evidence of binding. A conventional psychology will be essential here too. Strong binding could take man forms. They could be capable of getting into a state of mind in which the consequences of not doing the act were unbearable. (Remorse.) Or they could be capable of forming a kind of intention which had automatic priority at a later time. Or they could be capable of getting into a state in which preferences at an earlier time were the basis for actions at a later time. (Will.) Or they could have beliefs about the rationality of acts subsequent to an act of binding, beliefs which were so strong that no amount of perceived self interest could overcome them. (Faith.) Creatures that were capable of strong binding would have advantages over creatures that did not. Promises would be easier for them.

They might not be capable of strong binding. But they might still *believe* that they had the capacity. Conventional psychology would play a larger role then. (And the label 'conventional' would be more appropriate.) Suppose that they have the concept of binding and believe that binding is effective, and in fact when an agent is bound there is a good chance, though not a certainty, that the action in question will be performed. Call this "weak binding." Then many forms of cooperation will be possible for them. They will trust each other's promises. Sometimes these attempted cooperations will fail, and conventional psychology will come in again, to give case by case reasons for the failure, which do not undermine the belief that binding works. Suppose on the other hand that though they have the concept of binding and believe that it is effective it in fact quite often fails. Then some profitable cooperation will still be possible. Tom and Dave can trust each other's promises and Dave's harvest will get gathered. Then Tom's harvest may or may not get gathered, depending on the details of the case. Often it will be, and even if it is not Dave's harvest has at any rate been gathered.

Finally, they might not even be capable of believing in binding. Each sees himself and the other as making decisions in a moment to moment way influenced by the information and preferences of the moment. These beliefs may be true, but they will make it harder to achieve some mutually beneficial cooperation. (See Frank

[1988].) The remedies are well-known, though. As each case in which assurance is vital comes along they can try to set up a penalty for non-cooperation. (Dave can plant marijuana among Tom's crop and they can jointly send an anonyomous letter to the FBI in Washington which, the US postal service being what it is, will arrive in a month. Both will then be in deep trouble unless Dave has come over with his tractor to help harvest both the crop and the evidence.) Or they can join with others to set up a system of commercial law, with courts to enforce contracts, thus providing a single device for guaranteeing a wide range of mutually beneficial arrangements. All these remedies have costs, but the more intelligent Tom and Dave are the smaller they can make the costs. Since it is in everyone's interest to have such remedies, it is in everyone's interest to make the costs minimal.

In each of these four cases there is information that Tom and Dave can have which will ease their problem. Which would they be best advised to try for? That depends on the facts about them. If they are capable of strong binding then they should use it when appropriate. That is, instead of simply harvesting or planning to harvest they should first bind themselves: with these additional option the situation will not have the structure of the original game. But of course they may not be capable of this. Perhaps ideally rational agents would be capable of strong binding. Edward McClennen can be read as arguing for this (see McClennen [1990]), and with a little imagination so can Gauthier. But it does not follow that humans should try for strong binding, if humans will fail at it. (In this connection see Jackson and Pargetter [1986].) Given their limitations, it is possible that humans should try something they can actually do. So in order to come to a conclusion about rationality we need a psychological premise.

Rational agents might also find themselves in the last case. Their very rationality might show them options and show them how others might see these options, which would block both binding and belief in it. In that case they have the corresponding remedies: setting up penalties or binding contracts. These can be quite demanding of their thinking powers, though. They have to see non-cooperation traps coming and head them off, and they have to devise suitable low-cost remedies for them. Rational agents of arbitrarily great intelligence would have little difficulty here. They would be caught in the occasional trap by unpredictable random events, but in most cases they would have seen a problem coming and have devised a solution to it at minimal expense. But we humans have very limited intelligence. When we adopt this strategy we seem to oscillate between expensive substitutes for foresight and very fallible substitutes for ingenuity. We go either for overbearing institutions, which enforce cooperation in a large range of cases but at a high cost in resources and loss of liberty, or for happy go lucky improvisation, which can produce low cost solutions but which allows many uncomfortable situations to come up on us unexpectedly.

5. **Psychology to the rescue?** If we had one kind of ideal rationality we would be able to bind ourselves to cooperative actions. If we had another we would be able to devise cooperation-enforcing transformations of our situations, at minimal cost. In either case we would need enough psychology to understand the solution adopted. We are not ideally rational. We waver in our resolutions, and we do not see problems until they are upon us. (So the facts described in Chernaik [1986] are very relevant to

moral philosophy, and indeed to the practical design of institutions. See also Slote [1989].)

That may be the end of the story. There may be only two rational solutions to one of the fundamental problems of human life, and we may not be very well equipped for either of them. However in this final section I shall argue for a third solution, a more intrinsically psychological one. It corresponds to the second and third cases described above, in which agents are not capable of strong binding but do understand what it would be like to be bound. I think that, for better or worse, this is the solution adopted by most human cultures.

It is easy to think of characteristics people could have, whose presence would bind them to present or future cooperative behavior. There could be states of character, innate or acquired, which made it impossible for their bearers to renege on a promise or act on a narrow construal of their own interest. There could be kinds of pride or self-respect which would be unbearably damaged by letting down another person. There are important differences between these and other similar characteristics, but they share the feature that if a person had such a characteristic her cooperation could be relied on, even when in the particular occasion cooperation was not in her interest. Of course it does not follow that people can or do have characteristics with these features. A variant on these familiar ideas can be got from Gauthier. There could be a characteristic C such that if a person had C then if she believed that someone else had C then she would choose an option with a particular cooperative feature in some specific range of strategic interactions with that other person. (The wording is meant to avoid the circularity problem described by Smith [1991].) Or there could be a characteristic which ensures that a person will follow through with certain commitments. If another person has performed an action A on the assumption that she will later perform an action B, and she expects that her life will go better if she performs B than it would have had the other person not performed A, then she will perform B. A culture could have a name for these two characteristics. It might even refer to one or both of them as “rationality”. (The term would then be on the border between the psychological and the ideological. *Rough* connections with Gauthier [1977] and chapter X of Gauthier [1986].) Ascriptions of them could then play a large role in people’s decisions whether or not to enter into various interactions.

In some circumstances if all interacting people believe that each of them has the characteristic then cooperation will be assured, whether or not the beliefs are true. In others the result will be one-sided cooperation (which may or may not produce more total benefit than no cooperation at all, depending on the details.) There is something worrying about this, though, even when the result is cooperation. Could rationality result, after consideration of accurate evidence, in false beliefs, however useful?

The beliefs may not be false. There are two reasons for this. The first is that it is in each person’s interest that others believe that she has the characteristic. But it may be hard to fake. It may be linked to contingent features of human psychology which are very difficult to acquire except by going through a certain developmental process. That is, the easiest way to get people to think you have the characteristic may be actually have something much like it. (The easiest way to have the result of rational dispassionate dealings with people may be actually to like them, for example. See Bertram [1995].) I said “something like it” because the result of the process one has to go through for the characteristic to be plausibly applied to one is unlikely all by itself to have the required effect. The effects of self-attribution are also likely to be needed. For example if the characteristic is the simple sense of shame then cooperation may need not only a tendency to disquiet at broken promises but also the anticipation of the shame one expects will follow treachery. That anticipation is itself unpleasant, but it results in part from the belief in the conventional psychology. And if the characteristic is the Gauthier-style commitment-fulfilling disposition described just above then a person will keep her commitments not just because she has acquired the frame of mind needed to convince others of her trustworthiness but also because she expects herself to keep them. Believing that she has the characteristic in question she will not make the contingency plans necessary to follow up broken promises or failed expectations. Very occasionally she may find an opportunity for improvised unreliability that can be taken on the spot, surprising herself with what she takes to be out of character behavior. But much more often in order to take advantage of another you have to plan in advance, which you will not do if you do not believe yourself capable of it.

There is other reason why the beliefs engendered by a suitable conventional psychology may not be false. For there is another, subtler, way in which a conventional psychology can induce mutually beneficial choices. The psychology supplies the vocabulary used to describe actions and strategic situations. As I pointed out in section 1, echoing Gauthier [1975], it can make a large difference to ways agents approach a strategic situation how the acts are described. It can also matter how the situation is described. Any strategic situation, occurring at a particular moment between particular people, is an instance infinitely many situation-types. Which types a vocabulary for predicting acts and ascribing motives will make salient to those people at that moment, can influence the way they think through their possible actions. The influence can be crucial

A situation that happens to be a prisoner's dilemma, for example, will also be many other things. It will be an instance of a non-zero-sum game, a game with one equilibrium which is not Pareto-optimal, and a situation in which each person will hope the other is not paying attention. It may also be an instance of a situation in which one person wants to get an advantage from another, or a situation in which each desperately hopes they can trust the other. The psychological vocabulary of the people involved may not have any simple concept picking out all and only prisoner's dilemmas. For if we classify social interactions in ways which would be natural for people in almost any society - as buying and selling, building, reporting, hunting, or whatever - we find that most of them involve ranges of situations, many of which have PDs as special cases within the range. The result is that many PD-shaped situation-instances occur often, but PD-shaped situation-types are rarely described.

I imagine this is a very plausible claim. But to reinforce it consider two types of type of situation. The first might be called cumulative cooperation. Here there is some shared good which can be created by cooperative action. It may be produced in proportion to the number of cooperative acts performed or in some more complex way (in particular, it may be a non-linear function of input: twice as much cooperation may produce more than twice as much of the good.) There is a cost to individuals for cooperative action which is compensated for if enough others cooperate. (The cost may simply be physical tiredness.) In any such action the pay-off matrix in the 2-person case will have the form

$$\begin{array}{lll}
 & C & D \\
C & g(2) - c, g(2) - c & g(1) - c, g(1) \\
D & g(1), g(1)-c & g(0), g(0)
\end{array}$$

where $g(n)$ is the amount of the good produced when n cooperative acts are produced and c is the cost to the individual of performing a cooperative act. This will not in general be a PD. It will only be a PD when $g(2) - c < g(1)$, $g(1) - c < g(0)$ and $g(2) -c > g(0)$. On any particular occasion it will usually be quite hard for agents to know whether the situation instance is a PD, although they are conceptualising it as, in effect, cumulative cooperation, and know the form of $g(n)$ and the size of c.

Another large class of types of situation might be called foul-dealing. (Taking the term from Pettit [1986], but generalising to cover cases that Pettit did not intend to.) Here there is an aggressive action which one person can direct at another in order to gain some benefit, or can refrain from. If two people both act aggressively towards one another then the effect may be that both gain the benefit, or that both are harmed, or that both are protected by their aggression, or something in between. The pay off matrix will be like this:

	C	D
C	0,0	0-a, b
D	b, 0-a	0-sa+tb, 0-sa+tb

Here a is the harm done by an aggressive action, b is the benefit to be gained by aggression, s is the proportion by which harm is reduced by counter-aggression, and t is the proportion by which benefit is reduced by ones victim's counter-aggression. These situations, too, will be PDs only in special cases. Assuming that all these factors are always positive, they will be PDs when $a > sa - tb$, and $tb < sa$.

In real life there are other types of situations whose instances include simple or generalized PDs. (Especially when one moves beyond two agents and two options.) The consequence should be obvious. Suppose that people operate with a conventional psychology which prompts them to classify situations as buying and selling, building, reporting, hunting, and so on, and then at a higher level as cumulative cooperation, foul dealing, and so on. The concept of a prisoner's dilemma lies in between these two levels of description. Suppose that they do not have any simple characterisation of it. Then when they find themselves in a PD they will usually think of it as e.g. a house-building situation of a cumulatively cooperative sort. And thought of that way, the cooperative option will usually be the one each will both expect the other person to choose and choose herself. Or, to put it differently, each will know that the other is committed to a cooperative action because they know that the other is engaged in a an activity for which cooperation is the reasonable and profitable choice.

Rational beggars can't be choosers. We do have some resources both to bind ourselves to future actions and to anticipate cooperation-traps so we can devise low-cost transformations of them. But these resources have definite limits, and to that extent we have to find ways of solving our problems with more primitive means. The two uses of conventional psychology in this section do not make extravagant demands on our rationality. In fact, they both exploit our cognitive limitations. The first builds on the fact - if it is one - that the easiest way for a limited agent to give the appearance of a cooperative disposition is actually to have one. And the second builds on the fact - if it is one - that the classifications of situations that fit most easily into the decision-making heuristics that mere humans must use assimilate many situations where non-cooperation would be advantageous to ones where it would not be. As a result, a sufficiently intelligent human will always be able to find some situations in which by thinking around the concepts used by others she can take advantage of them. But there will never be too many such situations. (Rarely as many as hopeful intelligent humans persuade themselves they can find.) One reason is that no person is that much smarter than those around her. Another is that too original a way of thinking will break a person's link with the conventional psychology she needs to anticipate the actions of others. You can't exploit people unless you know how they are thinking, and you can't do this unless you can use the same descriptive and explanatory concepts that they do.

Bibliography

Baron-Cohen, Simon [1995] *Mindblindness* MIT Press.

Bertram, Christopher [1995] Self-effacing Hobbesianism *Proceedings of the Aristotelian Society*

Bratman, Michael [1986] *Intention, Plans, and Practical Reason* Harvard U.P

Byrne, Richard [1995] *The thinking ape* Oxford University Press

Frank, Robert [1988] *Passions within reason* Norton

Cherniak, Christopher [1986] *Minimal rationality*, MIT Press

Gauthier, David [1975] 'Coordination', *Dialogue* 14, 195-221

[1977] 'The social contract as ideology', *Philosophy and Public Affairs*, 6

[1986] *Morals by Agreement* Oxford University Press

[1994] 'Assure and Threaten' *Ethics* 104, 690 - 721.

Jackson, Frank and Robert Pargetter [1986] ‘Oughts, Options, and Actualism’ *Phil. Rev.* 95, 233-255

Kavka, Gregory [1983] 'The Toxin Puzzle', *Analysis* 43, 33-36

Lewis, David [1968] *Convention* Harvard University Press

McClennen, Edward [1990] *Rationality and dynamic choice* Cambridge University Press

Pettit, Philip [1986] 'Free riding and foul dealing' *Journal of Philosophy* 83, no 7, 361-380

Slote, Michael [1989] *Beyond optimizing : a study of rational choice* Cambridge, Mass. : Harvard University Press

Smith, Holly [1991] ‘Deriving morality from rationality’, in Peter Danielson, ed. *Contractarianism and rational choice* Cambridge U. P. pp 229-253.

Sugden, Robert [1986] The economics of rights, co-operation, and welfare Blackwell.

[1995] 'A Theory of Focal Points' *The Economic Journal* 105, 533 - 550