Folk Psychology is not a Predictive Device

ADAM MORTON

1. The prediction thesis

Human beings are special, as a species, not in having minds but in having the concept of mind. We have evolved, biologically and culturally, to be able to attribute states of mind, to be able to have and use folk psychology. And we are this way because we are social creatures; we have a very particular kind of sociality, in fact, which makes most of our activities cooperative while forcing us to manage them by a mixture of reasoning, social shaping, and imagination, without very many innate social routines. The human condition is to need answers about others' behavior, and to have to get these answers largely by various forms of thinking.[1]

The aim of this paper is to challenge what might seem like an obvious consequence of this fact. It might seem obvious that the central question about another person's behavior is "what will this person do?", in other words that folk psychology focuses on *predictions*. One might think that the purpose that drives a person's gathering and organising information about others is to know in advance what others will do. This is what I shall call the "prediction thesis. I doubt it. In this paper I shall not argue directly against the prediction thesis, but simply argue that it is not supported by plausible construals of the claim that the role of folk psychology is to allow human social life. Moreover I do not want to challenge some obvious truths which closely resemble the prediction thesis. To make things precise, formulate the prediction thesis as follows:

> *The prediction thesis*: The information about others that people normally use in making decisions about their own actions consists to a large extent of predictions that these others will perform specific actions.

It is important not to confuse the prediction thesis with a weaker claim, which I take to be incontestable, namely

[1] This picture of human nature, familiar enough from Aristotle onwards, has received a new impetus and some new twists recently through work in evolutionary biology. See the essays in Byrne and Whiten (1988).

Mind, Vol. 105 . 417 . January 1996

> *The prediction fact*: People in everyday life have significant knowledge of what other people can and will do.

In undermining the prediction thesis while not contesting the prediction fact I am claiming that we do not have the reasons we might suppose for thinking that it is by predicting action that we achieve many of the purposes for which we need to know about one another. More specifically, I am claiming that there are ways in which we can achieve the purposes of social life without having to predict what one another will do. Whether or not we *do* often make folk psychological predictions, we do not *have* to. Moreover, there are good reasons for preferring non-predictive ways of managing our relations with others in many situations.

In many contemporary discussions something like the prediction thesis is taken for granted. Often it seems too obvious to need careful statement. Sometimes the thesis does emerge fairly explicitly. For example Dennett writes

> Do people actually use this strategy [belief–desire attribution and prediction on the basis of rationality]? Yes, all the time. There may someday be other strategies for attributing belief and desire and predicting behavior, but this is the only one we all know now. And when does it work? It works with people almost all the time .... (Dennett 1987, p. 21)

Claims like this have a long history. Here is Hume on the topic:

> Were a man, whom I know to be honest and opulent, and with whom I live in intimate friendship, to come into my house, where I am surrounded with my servants, I rest assured that he is not to stab me before he leaves it in order to rob me of my silver standish; ... I know with certainty that he is not to put his hand into the fire and hold it there till it be consumed: And this event, I think I can foretell with the same assurance, as that, if he throw himself out at the window, and meet with no obstruction, he will not remain a moment suspended in the air. (Hume, *An Enquiry Concerning Human Understanding,* §VIII, Part I)

The prediction thesis is easy to confuse with the prediction fact in part because we underestimate the extent to which the choices for which we need information about other people are strategic; that is, concern outcomes which are the result of the actions of a number of agents, each trying to take account of the plans and reasoning of the others. And it embodies a mistaken picture of how we do and should make such decisions. Or so I shall argue. The underlying issues are related to some hard questions in decision theory. My method will be first to describe (in §2 and §3 below) the situations in which we need to come to terms with what others will do, in such a way that the opportunities for deciding without predicting begin to emerge. Then (in §4) I shall give a particular simplified model of how people can operate without explicit predictions. These parts

can be seen as blocking the argument "sociality therefore prediction" by showing that prediction is not the *only* way in which individuals can coordinate their actions. Then (in §5) I describe problems that arise when coordination is made to depend on prediction, in order to cast doubt on the claim that prediction is the *best* basis for coordination. At the end of the paper I argue that, if my account is right, issues about patterns of explanation in folk psychology are closely related to issues in ethics.

2. *Five cases*

Here are five very ordinary cases, illustrating typical needs people have for information about others and typical ways in which people use this information. In each case one person forms an expectation about what another may do; but each expectation differs in some way from a paradigmatic prediction.

First case: pleasing. Arthur wants to please Zenaida. Perhaps just because he likes her, or perhaps because he is terrified of her and would like her to treat him more gently. He knows she spends hours tying bassoon reeds, and so he commissions a carpenter to make her a special reed-tying stand, with storage space for cane, thread, and tools, and a special third-hand device for holding the reed while she ties it. Does Arthur know how Zenaida will react? Well, he doesn't know where she will put it or how often she will use it or whether she will give him a present on his birthday. But, if he is right about her, he knows that she will be pleased.

Second case, pure coordination. Yolanthe and Bruno have arranged to meet downtown. They were supposed to meet at the train station, but train services have been disrupted by a wildcat strike and so Yolanthe takes a bus into town. She thinks he may have taken a bus, too, or he may have been given a ride, or he may have given up and stayed at home. So there she is at the bus station, wondering how to meet him. She remembers that he likes high places, places from which you can see a long way, and she knows that he is aware of her liking for optical curiosities. So a good place for both of them to go would be the camera obscura overlooking the gorge. So that is where Yolanthe goes, but she won't be the slightest bit surprised if Bruno is not there.

Third case: fall-back plans. Xantippe has advertised a rare stamp in a collectors' magazine. Carlos phones her to buy it. They agree that she will send him the stamp and he will simultaneously send her the asking price. As she is addressing the envelope Xantippe reflects that she has no assurance that Carlos will really send the full agreed amount. Nothing has been put in writing. She posts it anyway, but contacts other collectors and deal-

ers to find out if Carlos has a bad reputation. If he does then she will be prepared to take steps to recover the stamp. She does not know whether or not Carlos will act honorably, but she has acted on the assumption that he will, and begun to prepare a course of action in case he does not.

Fourth case: faute de mieux. Wilma has agreed to meet the notoriously untrustworthy Dashiel. He has promised to be at a particular restaurant at 7 sharp. His promises are rarely kept, and so Wilma does not think that there is even a 50% chance that he will be there. She can think of twelve other places he might be, and does not feel confident that he is more likely to be at the restaurant than at any of these alternatives. But what can she do? In the absence of more definite information she takes him at his word and heads for the restaurant.

Fifth case: negotiation in action. Vanda and Elvis are at a party. It is a boring party and Elvis would like to leave. But he has just met Vanda and would like to know more about her. He has the distinct impression she would like to know more about him too. She is at the other side of a crowded room, talking to a doubtless very boring man. Elvis goes and gets his coat and stands in the doorway slowly putting it on. She sees him and points to her half empty glass. He hangs around putting his coat on very very slowly and half-heartedly conversing with people. Vanda appears with her coat but arm in arm with the boring man. As they pass Elvis she slips a note into his coat pocket. This is a complete surprise to him. But it is the satisfactory result of a course of action he set in motion as a result of his thoughts about her state of mind.

3. The general pattern: entanglement

The five examples of the last section illustrate some basic points about the contexts in which people form expectations about what one another will do. Two points are particularly fundamental.

First, much of our thinking about other people is directed at intrinsically social ends. We want to produce, or to avoid, situations which are defined in terms of what other people want, believe, and feel. In the simplest cases we want just to please or annoy, as with Arthur and Zenaida. But you can't please people unless you know what they want; you can't placate them unless you know what could enrage them; you can't disarm them unless you know their fears. So often we think about people in psychological terms in order to achieve ends defined in terms of what they want, think, or feel.

Second, many of the decisions a person makes are directed at outcomes which depend not just on that person's actions plus the way the world is,

but on those actions plus the way the world is *plus the decisions of other people*. And those other people's decisions are directed at outcomes which depend in part on what the first person decides. So each decision-maker has typically to take account of several other decision-makers, and of how each of these may take account of each other's taking account of their deciding, and so on. In the jargon of decision theory, very many of our choices are not parametric but strategic. One aspect of this is a potential instability of predictions: a tentative prediction that another will do one action may lead to the realization that the other's awareness that one may predict that action may lead them to do a different action, and so on, in some cases without a stable end to the process. (In §5 below I connect these points with work by Skyrms and others on the unstable dynamics of decision.)

Both of these points concern the entanglement of each person's decisions with those of others. Other factors lead to more entanglement. One is the making of joint decisions. Very often two or more people have related aims, which are best achieved by concerted action. And then not only do they act in concert, they decide in concert, by dividing up the job of choosing an action in various subtle ways. In the simplest case I indicate to you what I want and you indicate to me what I should do. The Elvis and Vanda case above is slightly more complex. Most real cases involve linguistic communication. Then a high degree of mutual knowledge can result in intricately entangled decision-making. But the converse is also true: action makes communication. Often one chooses an action because it opens up a line of action to other people. One way in which it does this is to reveal to them that this line of action exists and can have various consequences. Often, revealing this involves revealing facts about your own beliefs, desires, or tendencies. And, the other way round, often one acts in order to discover other people's beliefs, desires, and tendencies. There is no very definite line between ordinary social action and explicitly communicative behavior.

Another profound entanglement between different agents concerns the objects of our desires. Very often we do not want or fear outcomes defined simply in terms of what we individually receive or loose. For the *motives* of other people whose actions affect us also matter, and this can result in enormously complex decisions. A gain of $10,000 received as a reward is a different outcome from $10,000 given as a sign of contempt or from $10,000 given as an alternative to giving more. Suppose that *A* and *B* are deciding what claims to make for a sum of money to be shared between them. Each of them thinks that they have a claim to most of the money. Suppose that *A* considers giving *B* an equal share as a way of expressing her contempt for his greed. *A* is greedy herself but enjoys feeling superior even more than she likes money. But then *A* considers that *B* may be hav-

ing the same thought, so that an equal division of the sum may be for him a sign of contempt for her. This immediately rules out that outcome for her. (These issues have been discussed by Hammond, McLennen, and Nozick. See §5 below.)

This is the natural situation, the home environment, of folk psychology. Folk psychology developed and flourishes among agents whose interests and decision-making processes are entangled with one another's. (When dealing with an individual whose concerns are quite separate from their own, a stranger or a loner, people will often try to create entanglement, by eliciting either complicity or enmity.) The central problem of an individual agent faced with all the possible actions of other people is that of achieving a socially defined end by entering into a decision-making process which takes account of what others think and want and which is congruent with the decision-making of those others. The socially defined end can be as simple as pleasing a particular person or as complex as achieving a fair solution to a dispute; it need not be a warm and cooperative outcome: paranoia is as other-directed as love. It may be achieved by forming contingency plans, by making a decision to trust or mistrust, by explicit negotiation, or in many other ways, all of which require the agent to think about the beliefs, desires, and reasoning of others, but which do not necessarily require that action be based on a prediction of their actions.

Distinguish between *predictions* and *expectations*. An agent may base a decision in part on information about the likely actions of other agents. The agent may seek out this information before beginning to decide what to do. These are predictions. On the other hand it is also possible that in the course of making a decision the agent acquires opinions about what others may do, as a result of the same processes that result in the decision. These are expectations. They are results from, rather than inputs, into decisions. The decisions could have been made without them.[2]

The expectations that result from an agent's decision rarely have as their content that a particular other person will perform a particular action defined in terms of that person's purely individual desires. Instead they are expectations that the other or others will perform actions related in some way to the goal the agent has focused on herself. When a frightened four year old runs towards a parent, the child does not predict whether the parent will stand still, come towards her, or move toward a safe refuge where he can join him. The child's expectation is just that the parent will do something that corresponds to her fear and need. Trust is not specific prediction.

[2] There is a symmetry here between the first and third person cases. In deciding what to do you don't predict *your own* action, even though after deciding you know what you are going to do. You do know many things about your future actions, but these are roughly symmetric with your knowledge of other people's future actions.

4. One way to choose an action

In the examples of §2 agents choose actions in ways that depend on their knowledge of other people's minds but which then use this information directly as input to the agents' own decisions. The examples did not describe how the agents got from this information, plus their own beliefs and desires, to their decisions. Indeed it is usually not obvious how they do so. Our psychology is not that transparent, and questions of how we ought rationally and morally to get from information to action are rarely very easy even in simple cases. But it is not too hard to describe idealised procedures, by which agents could make this direct transition. Here is one.

Imagine a community of people facing an endless series of Prisoner's Dilemma interactions, in which the payoffs, the nature of the actions, and the identities of the participants, vary. Imagine that most social interactions in the community have the structure of (2-person) Prisoners Dilemmas, and that the interactions do not come in series: each is entered into with no anticipation of the effect its results may have on later interactions. Imagine moreover that the folk psychology of the culture contains two mutually exclusive terms *A* and *B*, which describe passing states of mind and are ascribed on the basis of a person's general demeanour. These terms fit into the simple decision-rule: if the other is in an *A*-state, cooperate; if in a *B*-state, defect.[3]

The effect of this will be that when two *A* people interact—each of them classifying the other as *A*—each will do reasonably well. When two *B* people interact they will do fairly badly. And when an *A* person and a *B* person interact—the one classified by the other as an *A* classifying the other as a *B* and vice versa—the *A* person will do very well and the *B* person very badly. But each person will have no problem deciding what to do; at any rate the decision will be as easy or hard as identifying which actions lead to which payoffs and categorising the state of mind of the other.

These decisions can generate expectations. If someone believes that, for example, another person is an *A* person and chooses to cooperate with that other, then she may expect the other to cooperate also. It might be argued that the expectations would rationalise the actions. (It is definitely controversial whether they would.[4]) But the expectations are not necessary for the actions chosen to be successful ones. Suppose that the expectations are false, so that for example although people in the culture

[3] For discussions of the prisoner's dilemma see Gauthier (1986, Ch. 1), and Hargreaves Heap, Hollis, Lyons, Sugden and Weale (1992, Ch. 9).

[4] For the rationality of cooperation and expectation of cooperation in such situations see Gauthier (1986, Ch. 6). For arguments the other way see Danielson (1991) and Smith (1991).

generally believe that *A* people usually cooperate and *B* people usually defect, people's tendencies to cooperate and defect are determined entirely by the decision-rule just stated. (So a person perceived by another as an *A* person will not cooperate if she perceives that other as a *B* person.) Then the people still may do well by acting according to the decision-rule. They will do well if they often read each other's state as *A*. (In such a society the capacity to produce the signs of *A*, a firm handshake and an authentic accent perhaps, would be ingrained early in each individual's life.)

A and *B* in this example are typical folk psychological terms in that they tend to be associated with simplistic behavioral assumptions which tend to remain unrefuted and to support profitable modes of interaction and decision-making as long as they are applied to people who subscribe to the folk psychology from which they stem. But the decision-rule in the example is absurdly simple and inflexible, leaving no room for anything like our actual complex motives, and finding no space for anything like thinking out what one oneself or another should do. A step towards real life, still in the form of a rule-governed procedure, might go as follows. Consider a number of agents in a strategic situation, in which it will be very hard for them to predict each other's actions, and even harder for them to turn these predictions into mutually desirable outcomes. But suppose that each of them has an idea of the beliefs and desires of each of the others, and in terms of this many of the essential features of the strategic situation facing them all will be clear, and known to all. (For example various kinds of equilibrium, Pareto-optimal outcomes, and potential conventions might be clear to all of them, if not under those labels.) And suppose that there stands out of the strategic situation one or more salient "standard" outcomes. (For the moment, leave unanswered the question what makes an outcome standard.) Then one plausible way for each person to think through their own decision is as follows.

The mutual strategy

(a) Find the standard solutions, from what you know of the beliefs and desires of all the people concerned.

(b) Isolate those actions of your own which could lead to standard solutions. Eliminate the rest.

(c) For each of these actions consider the possibility that you might do it while each of the others might perform an action leading to the same standard solution.

(d) Eliminate from the list any of those actions such that from what you know of the other people's motives and characters you cannot expect them to perform the required complementary action.[5]

[5] This expectation could be based on social imagination, inductive knowledge of a particular person's behavior, or simply on a hunch.

(d+) Return to (c), if time and interest permit.

(e) Perform one of the remaining actions.

Three features of this strategy are worth considering. Most fundamentally, it is a *stratified* method: rather than attaching a single comparative value permanently to each option it goes through them in a cyclic manner, eliminating options, but employing different considerations on different cycles.[6] Second, it is potentially a *shareable* method: at points (c), (d) and (e) the various cooperating agents can share information about which possible actions are still under active consideration for them. Often in a joint project agents can provide information about their choices for later actions simply by performing earlier actions. (I tell you I am considering taking the table up the stairs legs-upward by the way I choose to grip it while we are taking it through the door.) Third, it is a *converging* method: if different interacting agents have congruent notions of what would constitute a standard solution then if they all follow the method they will generally arrive at a solution that is standard for them all. (But full convergence may need more time than it is reasonable to give it.)

What is a standard solution? It can be anything that is easy to determine from knowledge of the states of mind of the people concerned. A norm of action can say: count this sort of thing as an acceptable outcome. Or the ethos of a culture can say: these features of people's desires and the relations between them are particularly significant. Or one person's attitude to another—love or loathing—can specify what would for that person count as satisfactory in their interactions with that other. It thus does not have to follow that if each person focuses on what they take to be a standard solution the result will be standard for either of them. What is important is just that congruence of norm, ethos, or attitude will tend to mutual standardness, and that as long as one grasps the norm, ethos, attitude, or whatever, of the other their action will be intelligible, even if unwelcome.

My suggestion is that there is a basic folk psychological strategy of thinking through situations involving other people in ways that resemble this general pattern. (The resemblance can be broad; even in the absence of something like use of standard solutions there can be stratified, shareable, and converging decision methods.[7] In so doing we have a manageable way of taking account of what others think, want, and feel in a way

[6] Such decision procedures are studied in Morton (1991). Very similar procedures are described in Bratman (1987, Ch. 3). The relation of different cooperating people's subplans (hence shareability) is discussed in Bratman (1992).

[7] The mutual strategy as stated could not be the complete decision-making routine for any rational agent. It leaves out really basic things such as a caution against actions which one should only undertake if one is absolutely certain that everyone else will play their part. Of course, it is just an example of the right sort of routine.

that can lead to mutually profitable coordination of their actions, and which provides both explanations and expectations about one's own and other people's actions. To this extent, it is at the heart of folk psychology. Handling information about others in order to produce coordinated action is what our ideas about mind are *for*.

5. Manageability

Whether or not some strategy such as the one I have been describing is fundamental to folk psychology, its very existence shows that decisions that take account of other people's actions do not have to do so by predicting them. This leaves open the possibility, though, that prediction is the *best* way of doing the job. By deducing specific predictions about what others might do from general beliefs about them we may be able to have a flexible and powerful means of gaining cooperation and avoiding treachery. Prediction may be able to give us good results in a wider range of cases than any other method, or so it might be argued. In this section I describe disadvantages of predictive strategies. The disadvantages have a common theme: our motives are complex, and entangled with one another's, and as a result explicit predictions of what we will do are very hard to get, often too hard to bother with. Life is short and our intelligence is limited. I shall appeal to holism, complexity, and entanglement.

Holism, as expounded by Davidson and others is the fact that it is not single belief–desire pairs but whole systems of belief and desire that rationalise actions.[8] Any belief or any desire can be rationally consistent with any action, given suitable other beliefs and desires. A desire for a quiet life can be rationally consistent with joining a group of itinerant revolutionaries if you believe that the only safe place in the next few months of chaos will be among the troublemakers. A belief that smoking causes cancer and a desire not to get cancer can be rationally consistent with smoking ten packs a day if you want to seem sophisticated even more than you want not to get cancer, and think that smoking ten packs a day will make you seem sophisticated.

That was rational consistency; all the more so for simple practical prediction. If you know that someone wants to get out of the building and thinks that the front door is the only way out, you cannot predict that she will go through the front door unless you can rule out the possibility that

[8] See Davidson (1984), and Heal (1989, Ch. 5, §5.3). Another side of holism, very relevant to the larger context of this paper, but which I shall not deal with here, is explored by McDowell (1994).

she, for example, thinks that opening the front door will set off a bomb, or thinks that the skylight is the front door. And even if you can rule out such complicating other beliefs and desires, for real life prediction you still have to know that the person is going to act in (what you take to be) the most sensible way to achieve their ends. Perhaps in her haste to get to the front door she will panic and run in the opposite direction.

The conclusion is that in order to predict actions you have to know more than a finite list of beliefs and desires. You have to know general features of the person's motivations that allow you to exclude whole classes of states that would be inconsistent with the beliefs and desires you attribute leading to the predicted action. In fact there is no way that one can know enough to make the connection between ascribed states and predicted action completely deductive. In practice some of the gap is filled by imagination: you suppose you had the beliefs and desires (and fears and hatreds) in the situation of the other and consider what other states and what actions would intuitively fit with them.[9] But a large part of the gap is also filled by restricting the choice of possible actions. Given the situation of the person and others they are interacting with, you consider only actions that your folk psychology tells you someone with that kind of motive in that kind of (social) situation would consider. You use normal solutions to contain holism; individual rationality leads to predictions only within the social bounds on intelligibility.

Suppose that holism presented no problems. Suppose that you could take a finite set of beliefs and desires, attribute them to someone, and consider only the actions that were rationalised by those actions. Then we would be in the land of game theory and rational choice where all might seem clear and simple. It might seem that when problems are thus posed there is a well-studied set of procedures describing reasonable ways of anticipating the actions of others. But in fact there are many problems. I shall focus on problems about complexity, about whether these are procedures that human beings with their limited capacities can actually use.

Rosemary knows the preferences and degrees of belief of Basil, over all the relevant consequences of a set of actions between which Basil is choosing. And Rosemary knows her own. So Rosemary should be able to predict what Basil will do, then see what, given Basil's choice, will be best for herself, and make a rational choice. But wait: Basil's decisions may depend on what Basil expects Rosemary to do. So to predict Basil's actions Rosemary may have to go through the reasoning Basil would perform to make his decision. (Rosemary thinks: he thinks I will try to get that parking space, so instead of waiting until it is safe to enter it he will rush in now.) So much

[9] For descriptions of kinds of imagination that might be suitable here see Goldman (1992), Gordon (1992), Morton (1980, Ch. 3), and Morton (1994).

is manageable. But when going through Basil's reasoning Rosemary may find that he will reason about her reasoning. So to predict his actions she may have to go through the reasoning she attributes him attributing to her. (She thinks: he thinks I want to get to that parking space before he does and will therefore rush into it, so he will rush first to prevent me.)

To predict Basil's actions she may have to know what he thinks she thinks he wants. The situation can easily get much more complicated. In quite simple situations one can need knowledge of knowledge of knowledge of preferences: knowledge cubed. It is not hard to find quite simple 2-person *n*-act situations in which each person needs knowledge to the *n*th. These situations are not artificial ones: their structure is that of situations that people face every day of their lives.

Its very hard to think in the face of such complexity. In particular, it is hard to see what step by step procedures will take you from the given data to a sensible decision about what to do in the light of what the other(s) may do. Mainstream game theory, dedicated to situations like this, does *not* give such a step by step procedure. Contrary to some descriptions of the theory, it does not have a coherent *dynamic* component, which describes the reasoning agents can follow. Rather, it focuses on various *equilibrium* concepts. That is, given a description of a situation it describes combinations of actions which are stable in various ways: once such an equilibrium is achieved the agents will not have reason to wish that they had acted otherwise. (The best known such concept is that of a Nash equilibrium, but there are others, with various advantages over it.) There is thus an important project, of deriving a dynamic theory from the equilibrium one. But this is difficult. Some dynamics assume more complex processes than any finite creature could instantiate. Others can produce chaotic trajectories through the space of possible situations. All of them are extremely sensitive to small differences in the degrees of belief and desire attributed to the various agents, in a way that does not augur well for modelling common sense ways of thinking.[10]

But we do face these situations all the time. How do we think them through? By not trying to apply dynamic principles. We do not try to deduce what the other person will do and then think out our own best response to it. Rather, we focus on common-sense equilibria. We restrict our attention to outcomes which we expect the other person will find it natural to consider,

[10] For the difficulty of justifying the use of equilibrium assumptions see Kreps (1990), and Rabinowicz (1992). For problems of modelling game theoretic processes with finite resources see Binmore and Shin (1994), and Shin and Williamson (1994). For chaos in deliberation see Skyrms (1990) and (1992). A position generally rather like the one in this paper has been proposed by Stalnaker (1994). Charles Newman's dissertation in progress suggests that these issues also connect with "framing" phenomena such as those discussed in Kahneman and Tversky (1979).

and then we apply everything we know about the person to anticipate which of these she may go for. Same conclusion: explanation by calculation only works against a background of social constraints on intelligibility.

Take it then that although there are good ways of thinking through what someone will do if you know all their relevant beliefs and desires and if no other states of mind are relevant and if the person acts in accord with a game-theoretic conception of rationality, these are ways that real people will find extremely hard to apply. But there is worse to come. There are situations in which even this ideal but impractical way of thinking cannot be got to apply. These are situations in which people's preferences are entangled, in the sense that what each one cares about is not simply the physical outcomes of the combination of their actions with others' but also the relation between these outcomes and what they and others care about. In the example in §3 above people cared about whether each other's actions resulted from contempt or gratitude. In other situations people care about whether their own and others' motives are based on opportunity or fairness, on respect or calculation. The difficulty this presents for models of rational decision is that as agents proceed through a decision process they gain information about the reasoning and motives of themselves and others. They learn more about what motives the resulting actions would be likely to be based on, and so their preferences over the possible actions change. In effect, when preferences and outcomes are entangled in this way, making a decision changes the information on which the decision is based.

It is far from obvious what is the best way of making a decision when information or preferences are changing during deliberation.[11] It is not obvious how to get the standard machinery to apply. For it assumes a set of outcomes towards which different agents can have different attitudes. We can describe cases like this either by saying that there are outcomes, but the agents have fluctuating attitudes to them, or that there are no outcomes independent of the deliberating processes, since an agent's attitude to "*o* chosen through process *p*" can be different from the attitude to "*o* chosen through process *p*'". So it is not obvious what effective means people have for predicting other people's actions in such situations, even though the majority of decisions faced by people living among other people involve some degree of entanglement between preferences and outcomes.

But in fact there are natural and common-sensical ways of making decisions in the face of changing information and preferences. For example there is the following procedure.

[11] Most work on dynamic choice is influenced by Peter Hammond. See Hammond (1976), McLennen (1990, Chs. 6, 7), and Ch. 2, Nozick (1993, pp. 55–6).

The dynamic strategy

(a) Determine a lowest acceptable outcome, from what you know of the possibilities of the situation.

(b) Isolate those options which, evaluated according to your starting beliefs and desires, are above the threshold. Eliminate the rest.

(c) Reconsider your beliefs and desires in the light of any new developments (in particular in the light of the fact that you have arrived at the ranking you did in (b).

(d) Eliminate from the list any options that rank very low under the new ranking.

(d+) Return to (c), if time and interest permit.

(e) Perform one of the remaining actions.

The similarity between this strategy and the mutual strategy of the previous section should be clear. Both allow agents to cut through complexity by ruling out options in a reasoned manner, starting from a non-arbitrary but possibly very imperfect first stab. In the mutual strategy one begins from combinations of actions which will be salient to all concerned by virtue of common socialization, and in the dynamic strategy one begins from outcomes which will have ascertainable desirability by virtue of ignoring the complications of what one might learn by reflection on one's own deliberation. From these starting points a stratified procedure converges on a final decision, taking it as far as time and patience allow.

Suppose that someone applies the dynamic strategy to working out what they will do in an entangled situation. Suppose moreover that there are one or more "standard solutions" defined in terms of relatively uncomplicated aspects of the situation and the attitudes to it of the people involved, such that its value to the decision-maker is above the lowest acceptable outcome. Applying the dynamic strategy, they will find themselves running through the steps of the mutual strategy. The standard solutions and others above the threshold will be evaluated relative to one another in a sequence of cycles in which increasingly complex information is used. The procedure does not have to terminate; it can just stop.

The central conclusion to draw from the ways in which entanglement can be tamed is the same as from holism and complexity: the standard ways we have of getting predictions of one another's behavior, in situations in which we have to compete or cooperate, are ones that work best against a background of option-limiting procedures which are by themselves aimed at getting coordination without prediction.[12]

[12] Combining individual means–ends reasoning with more collective forms of rationality is a delicate business. Trust tends to be eroded, paranoia spreads. This is a frequent theme in literature, for example in Ben Jonson's *Volpone*.

6. Morality

There are very hard questions here, and no philosopher can see a path all the way through them. It is generally agreed that our standard model of rational choice runs into problems with a long list of topics. The list includes: coordination and trust, resolute choice, changing preferences and beliefs, and risk. My own diagnosis is that the root cause is the complexity of the kinds of thinking required for strategic choice. There simply is no simple step by step thinking that will get you through all strategic problems, (just as there is no simple step by step proof procedure that will get you through second order logic). As a result, what common sense provides us with is a fairly good set of heuristics that put us on the route to acceptable solutions to strategic problems. Augmented with psychological intuition, craft, and phronesis, they often enough give non-grotesque results. When we approach non-strategic choice we use these same heuristics. For example we think of decision with changing preferences and beliefs as like a negotiation between present and future selves. But in this domain they give results rather different from the orthodoxies of decision theory, which no doubt have their roots in other departments of common sense.[13]

Whether or not this diagnosis is right, it is clear that if folk psychology is centrally concerned with structuring human interaction, then it must run directly into the complexity of the ways we think out what to do. This puts moral considerations at the heart of folk psychology. Not necessarily profound or radical moral considerations, but simply solutions to problems about how to get along with others for mutual benefit. A person's considerations about what others may do occur normally in the context of decisions about what she herself should do, and these decisions are usually thought out in terms of procedures which tend to coordinate the actions of people who use them.

If people make their decisions in very different ways—if they subscribe to different folk psychologies—then there is unlikely to be much mutually beneficial coordination. There is likely to be some element in a culture's ideas about mind and motive, then, of arbitrary but vital norm. That is, we can expect to find classes of action which people sharing the same formation will assume that one another will or will not do, in the absence of evidence to the contrary. (One obvious example is a norm to do what you said you will, other things being equal.) Such a norm governing people's default assumptions about one another will reduce the variety of possible actions they need initially consider to manageable proportions, leaving

[13] Hurley (1994) hints at a similar diagnosis.

mainly combinations of actions whose effects would not be generally disastrous.[14] We can also expect to find that there is an overall norm of intelligibility: people whose motives are too inscrutable are likely be treated as dangerous or untrustworthy.[15]

Both of these are probable results of the fact that we understand people in order to live with them. They are both special cases of a more general fact, that folk psychology is a part of *collective* rationality. Call an act-type collectively rational for a group of people in a given situation if when most people perform it most people will be better off if any single individual performs it than if that individual does not.[16] So if a single person makes her decisions by applying a norm specifying what each other person is likely to do, the result will usually be happy as long as all or most of the others whose actions affect the outcomes of the chosen actions apply the same norm. Acting by the norm is collectively rational, as acting by a rather different norm also might be.

Norms of action are only a small part of collective rationality. Lying behind them are norms of conceptualization. The value of thinking of people in terms of beliefs, desires, and personality, or in terms of social position, need, and the dictates of the gods, is conditional on other people thinking of them in the same terms. The descriptions and the explanations have to mesh. And they have to mesh with the rest of what people believe. So the psychological concepts, the metaphysical beliefs, and the moral norms that people acquire when they are being socialized into a culture must pull in the same general direction. They must add up to a coherent ethos, which permeates situations in which people find one another's actions intelligible in a way that lets them know how to deal with one another.[17] That is, the ethos that shapes a folk psychology must define a version of collective rationality meeting a number of conditions. One condition is that when people act in accordance with it the results will be in their long-term good. Another is that collective-rational acts of individuals should be intelligible to other individuals, intelligible in a way that allows individuals to give motives for one another's actions. Another is that *not* acting in accordance with collective rationality must be in the long run self-defeating: in a community sharing a conception of rational action, deviators from the equilibrium must eventually and on average do less well. There must be other conditions.

[14] For a discussion of norms of general benefit to those who collectively subscribe to them see Gibbard (1990, Chs. 11–3).

[15] Is this an abstract rationale for xenophobia?

[16] The natural classifications of actions will often cut across the concept of "same action" that this formulation demands, so it might be better to speak not of act types but of functions assigning actions to agents in situations.

[17] See MacIntyre (1984, Ch. 14).

Individuals whose actions are collectively rational may be capable of seeing the larger rationale that makes their way of interacting worth following. Or they may not be. A full explanation of the benefits it brings may simply need a long term average of preferences satisfied and unsatisfied. Or it may need to refer to what is objectively good for the people concerned, perhaps just minimally in leading to satisfaction of the needs that shape their desires.[18] In neither case need it form part of the explanations that individuals give when they give motives for past or future actions. Similarly, the fact that a community shares a conception of collective rationality does not entail that members of the community understand when they and others will conform to it and when they will deviate from it: to say what is rational is not to say when people are likely to be irrational, or even when it might be in their interest to be. To that extent the expectation that the other person will do the collectively rational thing—the tendency to trust that she will—need not take the form of a confidence that she will. (If she does not, you may be disappointed but not surprised.)

To know that an outcome would be collectively rational is not to know whether it would be rational for an individual. To know that an outcome would be rational, collectively or for an individual, is not to know what rational processes could lead to it. To know that a process is rational is not to know when a person might or not follow it. But it is often collectively rational for individuals to act as if others were collectively rational. Expectation is not prediction.[19]

Department of Philosophy
University of Bristol
9 Woodland Road
Bristol, BS8 1TB
UK

ADAM MORTON

REFERENCES

Binmore, Ken and Shin, Hyon Song 1992: "Algorithmic knowledge and game theory", in Biccieri and Dalla Chiara (eds.), *Knowledge, Belief, and Strategic Interaction*. Cambridge: Cambridge University Press.

Bratman, Michael 1987: *Intention, Plans, and Practical Reason*. Cambridge, Mass.: Harvard University Press.

——1992: "Shared cooperative activity". *Philosophical Review,* 102, pp. 327–42.

[18] See Railton (1986).

[19] Robert Black, Susanna Braund, John Broom, David Hirschmann, Charles Newman, Keith Wigglesworth, and the philosophy seminar at Bristol gave me invaluable help with this paper.

Byrne, Richard and Whiten, Andrew (eds.) 1988: *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford: Oxford University Press.

Danielson, Peter 1991: "Closing the Compliance Dilemma", in Vallentyne, P. (ed.) *Contractarianism and Rational Choice*. Cambridge: Cambridge University Press.

Davidson, Donald 1984: "Belief and the Basis of Meaning", in *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press. pp. 141–54.

Dennett, Daniel C 1987: *The Intentional Stance*. Cambridge: MIT Press.

Gauthier, David 1986: *Morals by Agreement*. Oxford: Oxford University Press

Gibbard, Allan 1990: *Wise Choices, Apt Feelings*. Oxford: Oxford University Press.

Goldman, Alvin 1992: "In Defence of the Simulation Theory". *Mind and Language*, 7, pp. 104–19.

Gordon, Robert 1995: "The Simulation Theory: Objections and Misconceptions, in M. Davies and T. Stone (eds.) *Folk Psychology: The Theory of Mind Debate*. Oxford: Basil Blackwell.

Hammond, Peter 1976: "Changing Tastes and Coherent Dynamic Choice". *Review of Economic Studies*, 43, pp. 159–73.

Hargreaves Heap, S., Hollis, M., Lyons, B., Sugden, R. and Weale, A. 1992: *The Theory of Choice*. Oxford: Basil Blackwell.

Heal, Jane 1987: *Fact and Meaning*. Oxford: Basil Blackwell

Hurley, Susan 1994: "A New Take from Nozick on Newcombe's Problem and Prisoner's Dilemma". *Analysis*, 54, 2, pp. 65–72.

Kahneman, D. and Tversky, A. 1979: "Prospect Theory: An Analysis of Decision Under Uncertainty". *Econometrica*, 47, pp. 263–91.

Kreps, David M 1990: *Game Theory and Economic Modelling*. Cambridge: Cambridge University Press.

MacIntyre, Alasdair 1984: *After Virtue*. second edition. London: Duckworth.

McLennen, Edward 1990: *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.

Morton, Adam 1980: *Frames of Mind*. Oxford: Oxford University Press.

——1991: *Disasters and Dilemmas*. Oxford: Basil Blackwell.

——1994: "Game Theory and Knowledge by Simulation". *Ratio*, 7, pp. 14–25.

Nozick, Robert 1993: *The Nature of Rationality*. Princeton: Princeton University Press.

Rabinowicz, Wlodzimierz 1992: "Tortuous Labyrinth: Noncooperative Normal-form Games Between Hyperrational Players", in Christina

Biccieri and Maria Luisa Dalla Chiara (eds.) *Knowledge, Belief, and Strategic Interaction*. Cambridge: Cambridge University Press.

Railton, Peter 1986: "Moral Realism". *Philosophical Review,* 95, pp. 163–207.

Shin, Hyon Song and Williamson, Timothy 1994: "Representing the knowledge of Turing Machines". *Theory and Decision,* 37.1, pp. 125–46.

Skyrms, Brian 1990: "Chaos in Game Dynamics". *Irvine Research Unit in Mathematical Behavioral Sciences*, technical report no MBS 91–18

1992: "Equilibrium and the Dynamics of Rational Deliberation", in Biccieri and Dalla Chiara (eds.) *Knowledge, Belief, and Strategic Interaction*. Cambridge: Cambridge University Press.

Smith, Holly 1991:"Deriving Morality from Rationality", in Vallentyne, P. (ed.) *Contractarianism and Rational Choice*. Cambridge: Cambridge University Press.

Stalnaker, Robert 1994: "On the Evolution of Solution Concepts". *Theory and Decision,* 37, 1, pp. 49–76.