The Importance of Being Understood

Folk psychology as ethics

Adam Morton

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

The Importance of Being Understood

'A very enjoyable and provocative book on an important topic, in Morton's inimitable style.'

Peter Goldie, *King's College London*

'A very valuable addition to the literature. Morton brings together material from game theory, philosophy and psychology in genuinely new and illuminating ways.'

Gregory Currie, *University of Nottingham*

The Importance of Being Understood is an innovative and thought-provoking exploration of the links between the way we think about each other's mental states and the fundamentally cooperative nature of everyday life.

Adam Morton begins with a consideration of 'folk psychology', the tendency to attribute emotions, desires, beliefs and thoughts to human minds. He takes the view that it is precisely this tendency that enables us to understand, predict and explain the actions of others, which in turn helps us to decide on our own course of action. This reflection suggests, claims Morton, that certain types of cooperative activity are dependent on everyday psychological understanding and, conversely, that we act in such a way as to make our actions easily intelligible to others so that we can benefit from being understood.

Using examples of cooperative activities such as car driving and playing tennis, Adam Morton analyses the concepts of belief and simulation, the idea of explanation by motive and the causal force of psychological explanation. In addition to argument and analysis, Morton also includes more speculative explorations of topics such as moral progress, and presents a new point of view on how and why cultures differ.

The Importance of Being Understood forges new links between ethics and the philosophy of mind, and will be of interest to anyone in either field, as well as to developmental psychologists.

Adam Morton is Professor of Philosophy at the University of Oklahoma. He is author of *Frames of Mind* (1980), *Disasters and Dilemmas* (1990), *Philosophy in Practice* (1996) and *A Guide Through the Theory of Knowledge* (1997).

International Library of Philosophy

Recent titles in the ILP:

The Facts of Causation
D. H. Mellor

The Conceptual Roots of Mathematics
J. R. Lucas

Stream of Consciousness
Barry Dainton

Knowledge and Reference in Empirical Science
Jody Azzouni

Reason Without Freedom
David Owens

The Price of Doubt
N. M. L. Nathan

Matters of Mind
Scott Sturgeon

Logic, Form and Grammar
Peter Long

The Metaphysicians of Meaning
Gideon Makin

Logical Investigations, Vols I & II
Edmund Husserl

Truth Without Objectivity
Max Kölbel

Departing from Frege
Mark Sainsbury

The Importance of Being Understood
Adam Morton

The Importance of Being Understood

Folk psychology as ethics

Adam Morton

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2003
by Routledge
11 New Fetter Lane, London EC4P 4EE

Simultaneously published in the USA and Canada
by Routledge
29 West 35th Street, New York, NY 10001

Routledge is an imprint of the Taylor & Francis Group

This edition published in the Taylor & Francis e-Library, 2005.

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book has been requested

ISBN 0-203-99493-0 Master e-book ISBN

ISBN 0–415–27242–4 (hbk)
ISBN 0–415–27243–2 (pbk)

Contents

Foreword

Where we're going

The central idea of this book should be intuitively to attract anyone who has driven in Rome, Paris, or Manhattan. You can rarely predict what any other driver is going to do. To that extent all folk psychology and all social contracts are useless. What you can do is notice what outcomes everyone would rather avoid, and then cautiously and very visibly to act in a way that makes those outcomes less likely. When you do, then often enough you create a situation in which the other drivers are constrained by their own interest (in not having their clean cars dented or bloodstained) to go along with that course of action. When you make yourself intelligible you gain a predictive hold over the behavior of others.

We are a mind-attributing species. We very readily think of one another in terms of emotions, desires, beliefs, thinking. We are also a cooperative species. Human life has always required that we share projects and work together, doing our best to avoid the possibilities for exploitation that this presents. This is the not inherently very promising niche that humans have chosen and flourished in, that of social animals whose sociality is flexible because of its relative poverty of innate social routines.

It is a not very daring suggestion that these two characteristics are linked. One important reason for attributing states of mind is to generate the expectations that make cooperation possible and allow us to anticipate and forestall exploitation, fake cooperation or cheating. We do not usually explain one another's actions out of pure scientific curiosity: there are reasons why we want to know why people act as they do, and these are generally practical social reasons. In this way *folk psychology*, the body of knowledge, opinions and routines that we use in conceptualizing and anticipating one another's actions, functions as a background to cooperative activity. Indeed one might take it as a background to ethics, taken very broadly as the enterprise of helping one another to the best lives. (Whether or not rights, duties and obligations are actors in this particular story.)

The aim of this book is to explore one family of connections between mind-ascription and cooperative activity. I explore a variety of ways in which the everyday explanation or prediction of actions is aided by the fact that the person whose actions are explained shares the routines of explanation or prediction in question. When this is the case then the presence of the routine is one of the reasons for the behavior that it explains, and *this* will rarely occur unless the behavior is shaped in a direction that tends towards mutually profitable interaction. Or, more generally, the presence in a culture of ways of understanding action can create a situation in which patterns of shared action are encouraged, which themselves help make the understanding appropriate. Beneficial circularities. And in the course of looking for examples of these phenomena I investigate more general interdependencies of modes of everyday psychological understanding and types of cooperative activity.

There is a big central idea here. We make ourselves intelligible, in order to have the benefits of being understood. It should not be exaggerated, though. I am not claiming that the function of facilitating cooperative activity by itself creates or determines any easily-named aspect of folk psychology. The thread that I am following runs through much of the fabric of everyday understanding, but it may not bear the full weight at any point. As a result, my exposition consists of analyses of a number of topics – explanation by motive, the concept of belief, the causal force of psychological explanation, simulation – in which I argue for claims from which conclusions relevant to my themes follow.

I am sure that the thread extends further than I have traced it. And I can give reasons for suspecting or conjecturing that it is found in several places where I do not have definite arguments for its presence. I don't think that philosophy should consist only of definite arguments for clear conclusions, though. We also need suggestive considerations that let us see possibilities that we can then slowly try to deal with. Every philosopher strikes the balance between 'clearly true but possibly boring' and 'clearly interesting but possibly false' in a different way. My strategy is to include both, but to label them as such. So the form of this book is that after five chapters in which argument and analysis are taken as far as I can take them, there is a conclusion. And then there are four shorter items, which I do not call chapters but explorations, which make suggestions about how the points I have made could be pushed rather further. They are speculative, and they help themselves to premises where they need to. Some readers may find some of them more interesting than the more careful chapters that precede them. But others may prefer to ignore them. You can decide which type of reader you are.

Writing this book has had an effect on my attitude to other people. When I began it I thought of people primarily as bearers of individual beliefs and desires, who have to negotiate their way around one another to get what they want. As I thought more about the project of understanding another person I came to think of people as searchers for basic goods, some of them intrinsically social goods. And as I thought even more I found myself thinking of people as balancing between goods they cannot clearly articulate and an equally obscure regret for the equally invisible goods they have missed. These were by-products rather than conclusions of my thinking. But as my attitude to people changed in these directions, to my surprise, I found myself liking them more.

Drafts of many of these chapters were presented to audiences, who helped me see what I really wanted to say, and why what I had produced did not say it. I cannot remember all the people at Adelaide, Armidale, Auckland, Brisbane, Bristol, Canberra, Christchurch, Dunedin, Keele, Leeds, Perth, Rutgers, Sheffield, Uppsala, Wollongong and Yale who gave useful encouragement and abuse. I have also had invaluable comments on drafts or in conversation from Jonathan Berg, Chris Bertram, Paul Bloom, Susanna Morton Braund, Tad Brennan, John Broome, Martin Conway, Jonathan Dancy, Martin Davies, Julien Deonna, Naomi Eilan, Norman Freeman, Christopher Gauker, Laurence Goldstein, Russell Goodman, Alison Gopnik, Robert Gordon, Paul Harris, Jane Heal, Andy McLennan, Laurie Paul, Josef Perner, Ashley Piggins, Peter Railton, Karolien Reiffe, Georges Rey, Stephen Stich, Rosemary Varley, Henry Wellman and Timothy Williamson. Some of these people would be amazed to know the impact a stray remark has had on my thinking. I was helped in the very early stages of the project by a grant from the Leverhulme Foundation. And at the last stage of the project I was helped by extremely pertinent comments on the whole manuscript from José Bermúdez, Greg Currie and Peter Goldie. Last of all I would like to acknowledge the influence of two good friends and good people, David Hirschmann and Martin Hollis, who died during the period I was working on this book.

Chapter 1

Microethics

> When this common sense of interest is mutually expressed, and is known to both, it produces a suitable resolution and behaviour. And this may properly enough be called a convention or agreement betwixt us, though without the interposition of a promise; since the actions of each of us have a reference to those of the other, and are performed on the supposition that something is to be performed on the other part.
>
> Hume, *Treatise*, III, ii, p. 2

> One does not predict where the other will go, since the other will go where he predicts the first will go, which is wherever the first predicts the second to predict the first to go, and so on ad infinitum. Not 'what would I do if I were she?' but 'what would I do if I were she wondering what she would do if she were I wondering what I would do if I were she'.
>
> Thomas Schelling 1960, p. 54

Microethics

This chapter shows how cooperation and wariness can do some of the work of psychology. First some examples.

You are driving along a narrow one-lane road, with a ditch on one side and a stone wall on the other. Another car is approaching from the other direction. Both cars slow down slightly, realizing that there is not room for both. There is an easily visible widening in the road, nearer you than the other car. You speed up to get to it, the other car speeds up to pass by you while you are in it, and both pass smoothly by each other.

You are walking towards a closed door, with your arms full of groceries. Another person is also approaching the door, slightly ahead of you. He accelerates his pace slightly. This generates an expectation in you. He has either seen the problem you face and intends to solve it by opening the door for you, or he sees that you might expect him to open the door and is rushing to get through before the issue arises.

You are playing tennis – doubles. You are near the net while your partner meets the ball. You can see one of your opponents preparing to reply, in a way that you haven't a hope of intercepting from where you are. You move out of the way, on the assumption that your partner is moving into position to meet the ball and return it through where you had been standing.

In all of these examples you form an expectation about what another person will or might do, which is linked to possibilities of cooperation or controperation (to coin an equally vague opposite). The patterns of reasoning that might support the expectation can run in both directions between considerations about cooperation and considerations about the other person's motivation. In the first case there is an obvious solution to the impending impasse for the two cars, and each acts as if the other has decided to implement it. So there is a potential inference from cooperation to psychology. In the second the assumption that you want the door opened could suggest to the other person either that he should get in a position to help or that he should get in a position to evade the request. So there are potential inferences from psychology to cooperation. In the third you can be understood as reasoning from the solution to the imminent problem to the acts it requires, for yourself and your partner. But you can also be understood as reasoning from your partner's apprehension of the problem to her decision to act appropriately (to your own decision to act so as to complement what you expect her to do).

There is no conflict here. We do not have to decide whether psychology is generally inferred from cooperation and controperation or whether the inferences generally run the other way. In most real cases there are many simultaneous thoughts, and many interlocked transitions between them, so that the overall pattern can be very mixed. Typically one derives solutions to problems of cooperation from attributions of states of mind *and* the other way round, in a complex interdependent pattern. But this fact is itself very significant. It opens up new ways of thinking about our everyday understanding of one another, of what is often called folk psychology. ('Folk psychology' is a very loaded term. See warnings and prejudices in the endnote.)[1] It suggests that sometimes and in some ways we understand because we can cooperate rather than the other way round.

Let me coin the term *microethics* for the collection of ways of thinking we have in everyday life for finding our way through frequently occurring situations in which the stakes are low but there are potential conflicts of interest between individuals. (I'll include also the thinking that smoothes out conflicts and transition stages of a single individual's history). Children learn a lot of microethics in the first few years of life, at the same time as they are picking up mental concepts. At this point I need make no claims of continuity or discontinuity between microethics and full scale explicitly conceptualized moral thinking. There may well be conflicts between the hardly conscious routines a person uses to minimize conflicts with others in everyday life and some of her most deeply held principles about duty, obligation and the good. And I need not assume that there is a fundamental unity to the different elements of microethics. Some components may have very different origins or nature to others. It is pretty clear that some components are shaped almost completely by inherited adaptations to social life, and some other components are the effects of particular patterns of social life and of particular beliefs and values. (See the introduction and essays in Barkow *et al.* 1992, and the essays in Whiten and Byrne 1997, especially those by Schmitt and Grammer, and Gigerenzer.) And the interaction between components of these two kinds must be extremely intricate.

Microethics and the attribution of states of mind are intimately connected in everyday life. It is rarely obvious on any given occasion which microethical lemmas are derived from which psychological ones, and vice versa. Often we can see how an attribution of a state of mind *could* be based on a solution to a microethical problem. For example in the first case above the other driver could be taken as reasoning: a collision would be bad, a deadlock would be bad, the only easy way of avoiding both is for the car approaching me to speed into the passing place, so I expect the driver to intend to achieve that. In this case the psychological state attributed is an intention, but it could as easily have been a desire, a process of reasoning, or a belief.

The derivation of rudimentary psychological attributions from microethical premises is the theme of this chapter. I aim to make visible a layer of thinking which can serve many of the functions of attributing beliefs, desires and other states of mind via the solution of problems arising when agents interact. The existence of such a layer should not be surprising, when one considers that small children, very early in their acquisition of adult concepts engage in activities requiring the kind of delicate coordination found in the examples above. And other intelligent social animals face and solve similar problems. (In fact, the patterns of thinking described in this chapter are, I would say, the common property of two-year-old children and adult dogs.) One thing that I must make clear, therefore, is the ways in which microethical thinking mimics or imitates or substitutes for the attribution of states of mind. That is the task of the next section of the chapter. In the rest of the chapter I shall argue for two conclusions. First, that microethics is often a good basis for psychology, in particular that inferences from solutions to problems of cooperation to attributions of states of mind are often sensible ways of reasoning given the limited data, time and thinking power available. (Similarly, the states of mind that people are interested in finding in one another are typically states whose presence makes a difference to the possibility of worthwhile interaction.) In arguing this I shall have to make the ideas of microethics and of a solution to a cooperation problem clearer. And, second, I shall argue that to some extent we *have* to base our attributions of states of mind in part on microethics. In particular, we have to classify the actions that we explain by attributing states of mind in microethical terms. These two conclusions are different. The first conclusion is just about possibility, since these are not the only sensible ways of reasonng about the topics concerned. The second conclusion is about necessity:

although the dependence on microethics that it claims is weaker than that described in the first conclusion, it is one to which there are far fewer and less likely alternatives.

Strategic problems

To see the special intimate link between microethics and the attribution of states of mind we must consider what is special about many-person interactions. We must consider *strategicality*.

Strategic situations are those in which the outcome for each of a number of interacting agents depends on the actions of the others. If all the people in such a situation are deliberately choosing their actions, each person's decision will have to accommodate its connections with those of others. Notice that I have not characterized strategicality in terms of agents' preferences or the reasoning they may follow. I am after something more fundamental. In a strategic situation the actions of the interacting agents are responsive to some factors in the environment, which are themselves affected by the agents' actions. Each agent is such that in the absence of other agents they would act so as to bring about some end result or maximize some quantity. But in the presence of the other agents the choice of the end to bring about or the degree to which the quantity is maximized is a more complicated matter. For it is the joint effect of all the agents' actions, itself caused by the presence in each agent of processes parallel to those which select actions for each other, that is efficacious. (The wording is meant to prevent the impression that the outcome-yet-to-be teleologically magics the actions into being. See Chapter 4 "Strategic attractors'.)

The simplest examples are at one extreme in species evolution and at another in deliberate economic action. In the first, whether a trait developed by one species maximizes reproductive fitness depends on the traits developed by others, and vice versa. And in the second, whether an act brings in enough money to maximize the utility of a particular agent depends on the choices made by others, and vice versa. But most of the strategic situations that interest me fall between these two extremes, being neither the result of random mutations nor explicitly calculated self-conscious choice.

Some terminology. I will refer to the processes that determine a person's acts as choices, hoping not to give the impression that a choice has to be calculated or self-conscious. If you are walking to the bus stop as the bus comes around the corner, and then just find yourself running, this will be for me a choice. And I shall refer to strategic situations rather than to games, when, as usually, I want to avoid the impression that the interacting people are to be thought of as idealized, explicitly deliberating, or paradigmatically rational. Though I will not be fully consistent in the usage I shall usually use 'person' in the sense of 'flesh and blood human being' and 'agent' in the sense of 'idealized actor as construed by some theory of action'.

Two constraints on strategic choice are particularly important. The first is that the basic material shaping a person's decision will not include definite probabilities of each other person's making one choice rather than another. For the other people's choices are themselves dependent on the as yet unmade choice of the person in question. (If you could know in advance what the others are going to do then you could know what you yourself were going to do, and you wouldn't have to bother deciding.) As a decision progresses each person can come to conclusions about what others are likely to do, just as she comes to conclusions about what she may do. But these are products of, rather than inputs to, the decision. (This is a feature exploited to great effect by Brian Skyrms' exploration of decision-dynamics. See Skyrms 1992.)

The other important constraint is that a person cannot simply choose the action with the best outcome. For an action may have a wonderful consequence *if* others make suitable choices, under circumstances which on reflection show that the others would be very foolish to make those choices. And if they don't, its consequences may be disastrous.

Procedures that choose a person's actions, whether by explicit calculation or not, will have to produce rather than rely on estimates of the likelihood of acts performed by other agents. And simply going for the best is potentially disastrous. The result is that the idea of a best, or most rational, choice becomes rather problematic. A choice is not better than alternatives when its expectation is greater, that is, when on average it will produce better results. For this average can't be given a sensible value without weighing in the likelihood that the other people concerned will do one or another of the acts open to them. The most that can be said is that it is appropriate for one person to choose her actions by a procedure that is congruent with those of the people she is interacting with, in that the resulting combination of actions will give her better results, on average, than she would have got by choosing actions according to any other method. And if the procedures involved also produce estimates of the likelihood of particular people choosing particular actions, it is appropriate for such a stable combination of action-choosing procedures to be such that the actions chosen do give the best results according to these likelihoods.

This fact will appear in some form on any approach that recognizes the strategic character of many-person interactions. In game theory, the formal theory of strategic choice, it results in a variety of *solution concepts*. A solution concept is a criterion for satisfactory choice, such that if a person's choice conforms to the criterion, *and so do those of other interacting people*, then the result will be reasonably good for that agent. There are many solution concepts. Two well known ones are von Neumann and Morgenstern's classic minimax criterion – choose the action which is least bad whatever the other chooses – and the now orthodox Nash equilibrium – choose an action which is part of a combination in which no one can do better by unilaterally choosing differently. There are much more subtle candidates (an accessible treatment is found in Hargeaves-Heap *et al.* 1992. For a deep and rigorous treatment see Myerson 1991). No solution concept is intuitively right in all cases. For the sake of exposition I shall work (in this section only, mind you) with the standard idea that the Nash equilibrium is at any rate a necessary condition for a solution to a strategic situation.

Now we can finally say more clearly why making a strategic decision will always be an implicitly psychology-involving process. Consider an example, a pretty simple one, but just complicated enough to bring out the essential factors. We have two agents, call them ‘Cooperative doormat’ and ‘Selfish pig’. Each has two actions available to them; call them A and B. There are thus four possible outcomes resulting from the combinations of each agent’s two actions. These can be represented in the standard matrix, as follows below, in which a pair such as (4,0) means that the outcome is satisfactory for A at level 4 of some scale and satisfactory for B at level 0 of some scale. It is very important to understand that the scale need not measure how much A wants the outcome in any intuitively familiar sense. All that is assumed is that the scale measures the weight the outcome has with respect to some behavior-choosing process that manages the kind of entanglement typical of strategic choice. (And, in case it is necessary to say this too, only the ordinal quality of the numbers matters, not their cardinal values, and as a result there is no implicit comparison between the degrees to which the outcome is satisfactory for the two agents.)

	Pig	
Doormat	A	B
A	(1,1)	(0,0)
B	(0,2)	(4,0)

There is only one equilibrium of the situation: Doormat chooses A, and Pig chooses A also. If this is what they choose then neither will do better by reconsidering alone. Consider how the two characters could think their way to this equilibrium, if it is by thinking that they tackled the situation. From Pig’s point of view this is a quite simple bit of conditional reasoning. If Doormat chooses A, Pig is better off choosing A rather than B and if Doormat chooses B, Pig is also better off choosing A rather than B, so in either case Pig should choose A. But now consider the situation from Doormat’s point of view. Doormat is best off choosing A if Pig does, and best off choosing B if Pig does. So to know that A is the best choice Doormat has to rehearse Pig’s reasoning. Then Doormat can see what Pig will choose, and as a function of that, what she herself should choose.

(A story to go with the structure. Each of them can go to a film or a party, but Pig strongly prefers the film – her favorite – to being anywhere Doormat might get drunk and maudlin. And Doormat strongly prefers Pig’s company.

So Pig thinks: film whatever. And Doormat thinks: she'll go for the film so that's what I'll have to go for.)

In this example it is Doormat's choice that is essentially strategic. Pig does not have to consider Doormat's reasoning, and in fact she can think of Doormat's actions as impersonal events in the world, like the weather, without considering the reasoning that may lead to them. Doormat on the other hand has to consider what reasoning may occur to Pig. In many equally simple situations each interacting agent has to consider the reasoning that may occur to each other agent. One important such situation is that of pure coordination in which each agent, like Doormat alone in the example above, has most to gain by doing whatever the other does. These situations are represented by matrices in which the agents have highest utilities in the outcomes of their both choosing the same actions – along the diagonal of the matrix in the usual layout. To choose an action in a situation of pure coordination an agent has to find the action that the other agents will choose, given that each of them is looking for the action that the agent herself will choose (will choose in looking to coincide with their choices, that is.)

The example was quite simple in that Doormat had only to consider Pig's likely reasoning about the consequences of her actions. There are examples that are only slightly harder to describe in which agents have to consider one another's reasoning about one another's reasoning about the consequences of their actions. And there are examples that are only slightly harder than these in which agents have to consider one another's reasoning about one another's reasoning about one another's reasoning about the consequences. And so on. (The examples are discussed in the appendix to this chapter.) The complexity is intrinsically linked to psychological attribution: it consists in having to consider the reasoning that the other person may be entertaining.

So the basic connection between strategic decision and psychology is made. In a strategic situation each person's act has to take account of what determines each other's. Since the discussion above was standard game theory the determining was described as explicit inferential reasoning. But the story is the same however the choices are made. As long as the processes that result in each agent's choice give good results in strategic situations they will have to take account of the processes that lie behind the choices that other interacting agents are making. They will have to represent, or mimic, or otherwise accommodate the structure of, whatever determines the actions with which the chosen action will have to combine.

In fact the psychological link is even deeper than this suggests. For there is a way around the complexity. Instead of considering what one another might be reasoning the agents can simply look for the equilibria, and choose actions that lead to them. To do this they have to be assured that each is operating in this way. In game theory this is taken as an assumption of mutual knowledge of rationality. Given such an assumption interacting agents can know what the outcome of reasoning about one another's motives would be, without actually doing it. But in fact the potential psychological implications of this way of deciding are deeper and more interesting.

The crucial fact is that equilibrium is a consistency property. It holds when a set of choices (or a set of probabilities of choices) by a number of interacting agents, and a set of preferences, are mutually sustainable. (Any solution concept will be a consistency property.) It thus allows one to infer information about any one element given the others. So it is customary, given equilibrium, to infer acts from preferences. But information about preferences can also be inferred from acts. In the appendix to this chapter I show how in many cases given the acts performed by a set of interacting rational agents and an equilibrium assumption we can constrain the agents' preferences, sometimes to the point of determining them uniquely. Other fillings-in are also possible. From the acts chosen and the preferences of the other person or persons you can infer your own preferences. The preferences you thus ascribe to yourself may or may not be consistent with what you have introspective or other reasons for taking your preferences to be. If not, something has to give. (In fact some of what you take to be introspection may be inference from your attributions to others. Introspection is not very good at telling us when we introspect. This is a theme of Exploration I.)

The credentials of basic strategic reasoning as a fundamental attribution device are now in place. It does not require the reasoner to have any prior idea of rationality, of attribution, or even of their own states. ('ought to do A' thoughts are not necessary, nor even 'she thinks that p', nor even 'I think that q'.) The attributer is faced with a problem of deciding what to do – how to pass along this road without a collision, win this tennis game, have a satisfactory evening out with a selfish partner – which he solves as well as he can. Then in the course of so doing he comes up with both a conclusion about what he will do and one about what the other interacting person or persons will do. This is so whether the decision is made by following chains of reasoning about reasoning, or by mentally mimicking the reasoning of others or by finding equilibria. It can thus operate as part of the background against which sophisticated thinking involving ideas of rationality, of evidence for attributions, and inferences from self to other, can operate. In fact, any really basic way in which individuals take account of what lies behind one another's actions will have to be a procedure for making strategic choice. For the situations in which most actions are made by social creatures are overwhelmingly strategic, and thus an individual's determination of what she is to do and her determination of what others will do will nearly always be inseparable. So just as most choice is strategic, most attribution will have to go hand in hand with strategic choice. The really deep and difficult question is what the psychologically manageable ways of achieving it are.

Some questions not to beg

(This section is a methodological digression. The main argument resumes in the section 'The flesh and blood version'.) There are questions that it is important not to beg here. These are issues where the most helpful function for a philosopher is not to argue for one or another possibility but to argue that the range of possibilities is very open, and that it cannot be argued down to a smaller range.

The questions can be put in terms of a contrast between extreme views. The first extreme is the view that the vocabulary that naïve and sophisticated people use to describe one another, and the ways of thinking that are needed to back up this vocabulary, are reasonably accurate descriptions of basic aspects of human psychology. People have beliefs, desires, preferences, hopes, plans and memories, on this account, which are among the important causes of their behavior, and would have them whether or not they were described in these terms. I suspect that something like this position is taken for granted by most unreflective people.

At the other extreme to this there is the view that these concepts apply to us just because we apply them to us. According to this position the presence in a community of psychological thinking, the fact that people interpret one another's behavior in terms of belief, desire, intention and the like, creates a pressure to behave in ways that can be so interpreted. The pressure affects us from a very early age, so that by the time we are capable of reflecting on the way we think the need to be understood, it has in fact shaped our thinking so that it is as if it were intrinsically describable by everyday mental vocabulary.

It is not easy to find examples of philosophers who are all the way out along this extreme. We should exclude philosophers who think that *all* concepts are only applicable because we apply them. (There are only rabbits, on this attitude, because people have the concept 'rabbit'.) And we should exclude eliminativists, who think that everyday mental terms do not apply to pre-social humans because they do not apply to any humans. But there are philosophers who are in neither of these camps, and who hold a position near to the extreme I have just described. Ian Hacking and Martin Kusch can be read as subscribing to such views. (See Hacking 1995 and Kusch 1999, also McGeer 2001. I think some works in feminist epistemology give arguments pushing in this direction. See Longino 1999.)

Proponents of the first, more realist, attitude can have more straightforward views on truth or falsity in folk psychology. For if there are definite states of people that our attributions describe well enough for it to be definite that those are the states that they refer to, then our attributions could, though roughly right, be definitely false. Our concepts are near enough to the facts that our assertions can be true or false of them, either by suggesting relations between properties which hold or fail, or by suggesting that a definite kind of event has a definite kind of cause which it does or does not have. So from this point of view Stich, Fodor, and the early Davidson, for example, for all their differences all have the realist attitude. (Dennett's position is harder to classify).[2] On the second, more conventionalist, attitude, it is much harder to see how a folk psychology could be false of people whose life it mediates. And its truth is in a way hollow. It is as if we counted it as a deep truth about cats that they are called by their real name 'cats'. So from this point of view the later Wittgenstein has a conventionalist attitude.

The truth is surely somewhere in between. But it is unlikely to be expressed simply by picking the right mid-way point. Some mental concepts, and some of the ways of thinking that support them, will no doubt be best described along the lines of the first extreme, and some others along the lines of the second. Most others will be scattered at different points in between. I have no intention of defending any extreme or compromise position on this issue. In fact, my aim is to urge us to leave the issue as open as we can. There is simply no way it can be settled given the resources at our disposal at the moment.[3]

In this book I shall avoid assumptions that force us to any particular point along this spectrum. I shall consider it an argument against a philosophical position that it seems to close these open issues. (But I expect that any reader whose prior instincts were to subscribe to the first extreme position will end the book with more respect for the possibility that the truth may lie some small way in the direction of the second.)

This is the first of the questions that should not be begged. Call it the contrast between realism and conventionalism with respect to common sense mental concepts. (One might expect someone who thinks that folk psychology is prior to microethics to be a realist, and someone who thinks that microethics is prior to folk psychology to be a conventionalist.) The second question concerns the causal importance of the facts that we mention when explaining or predicting someone's actions. Consider an example. A person phones her lawyer and makes an appointment to draft a will. When asked why she did it she says 'I have several plane trips in the next few months, and I have a child now, so it's important to me that everything be in order, just in case.' Assume that this explanation is honest, accurate, and satisfies all the conditions we could reasonably ask a commonsense explanation of an action to satisfy. Does it follow that if the factors she cited had not applied she would not have made the appointment? Or that these factors would not apply in possible situations in which she does not make the appointment? Or that they do apply in most situations in which she makes the appointment? It is far from obvious that the answer to any of these questions is Yes.

Again there are two extreme positions. At one extreme is the view that the factors we normally cite are the main causes of our actions, and express the major factors that shape what we do. (I won't be more precise about 'main', 'major', and 'shape'. The extreme view is obviously a bundle of extreme views.) Jonathan Dancy may be expressing some such view when he begins a book as follows.

> When someone does something, there will (normally) be some considerations in the light of which he acted – the reasons for which he did what he did . . . Intentional, deliberate, purposeful action is always done for a reason . . . So normally there will be, for each action, the reasons in the light of which the agent did that action, which we can think of as what persuaded him to do it.
>
> Dancy 2000, p. 1, taken somewhat out of context in that Dancy is operating with his own very non-Humean conception of a reason. See also Davidson 1969.

What makes these remarks not be the platitudes they might seem is the combination of the repeated 'the' and the assertion that the reasons 'persuade' the agent to act. The definite article suggests that the reasons are unique or complete, and the element of persuasion suggests that they are what the agent would cite in explaining the action. In combination, these contrast the position with an opposite extreme position. According to that opposing view, in any agent at any time there are enormously many competing and coinciding causal factors that send the agent in many different directions. Some of these factors can be expressed in commonsense terms, and many cannot. What determines which of them affects an agent's actions at any given time is partly determined by chance, and to a large extent by factors of which we have no knowledge. (It's not that the ordinarily cited motives are not *among* the causes. It's just that we take a small sample from an enormous population.) Consider again the person making the appointment to draft her will. Suppose that she hesitates after dialing the lawyer's number, and hangs up the phone. She explains to a friend 'it's not quite the moment; in a week's time I'll be a lot clearer about what I want.' It may be true that she will understand her situation better in a week, and it may have been true in the first scenario, when she did make the appointment. The suggestion is that all we can truly say about her may be compatible with an enormous range of actions, all of which could be made sense of by judicious selection from the rich body of competing motives that any adult person has. There may also be actions that she will *not* do in any alternative possibility: but there may be no way of discerning these from a knowledge of what she wants, thinks and feels.

Call the first of these two extreme positions 'optimism' with respect to everyday reasons, and the second 'pessimism'. There are reasons for skepticism about extreme optimism. In the first place, there is the wealth of data in the past 30 years indicating that people have very unreliable judgments about their motives. They know very little about which factors shape their actions, or which of the factors they are aware of actually play a role in leading them to act. (Much of the immense literature establishing this fact is summed up in Ross and Nisbett 1991. Pears 1984 provides a stimulating perspective on the first wave of this work.) This suggests that the causal stories we tell in everyday life are at most a small part of the whole picture. And, complementing this, there is the testimony of fiction. In the novels of Henry James and Proust, and the plays of Sartre, there is a sustained attempt to portray people as moved by a confusing wealth of factors, of which they and those they interact with are to a large extent unaware. Some of the most expressive and observant attempts to fit actions into intelligible patterns conclude that the patterns make most sense when it is made clear that they are part of a larger more mysterious totality.

These considerations are far from conclusive. There are also arguments against extreme pessimism. We do have at least some faint predictive hold on one another's actions, on the basis of the attributions we normally make. People usually have some idea what they, and others they know, may do. (I return to this in the last section of this chapter.) And the fact that we have the concepts of the lame excuse and the implausible rationalization suggests that we are not capable of welding absolutely any motive to absolutely any action. These also are clearly not conclusive. Again we have a spectrum of possible positions, in which the extremes are not particularly attractive, and we lack the evidence to choose an intermediate position. I shall take care not to assume either optimism or pessimism about everyday reasons, and I shall consider it a defect in a philosophical position that it begs the question of where between them the truth is to be found.

The open-ness of the two unbeggables will recur throughout the book. At this point in this chapter they serve to warn against simplistic ideas about what may go on in real people when they negotiate strategic situations. Early in the previous section there was a warning: do not take the psychology that seems to be implicit in game theory too literally. Take it as a way of describing some relations between situations and actions, in terms of a particular kind of idealized and cognitively unlimited agent. Don't take these agents as descriptions of human beings. This warning may seem unnecessary. Of course real people make mistakes in reasoning and cannot handle more than a limited amount of information. Of course real people change their preferences in the course of deliberation. But the real point of the warning is to make us keep in mind the possibility that the ways real people find their way through strategic situations may be *much* more different from the models of economists and philosophers than this. We may often not reason about one another's reasoning; we may often not head towards situations that can be defined in anything like preference rankings. And so too with ordinary explanation in terms of preferences, desires, beliefs, beliefs about beliefs, chains of reasoning and the like. This vocabulary cannot be completely ungrounded in the facts. But we should be very wary of simple pictures of where in psychic reality its roots attach.

The flesh and blood version

The claim is not that fundamental social processes are churning out perfect examples of game-theoretic reasoning. Such a claim would not be completely unreasonable. For a species that survives by means of negotiating strategic situations must do so by means that give good results much of the time and good results are usually not going to be too different from the outcomes that ideally rational agents would gravitate towards. (Game-theoretic equilibria are much more plausible as predictions of what will happen when an interaction is repeated many times. For then when an agent can improve their result by choosing differently they will next time round. Hence the power of game theory as applied to evolution. See Morton 2000). But this does not mean that we can model social situations in terms of simple game theoretical matrices, compute the orthodox solutions to them, and then trust that human agents are approximating to the same computations.

Putting aside doubts about whether game theoretic orthodoxy does in fact always say even what perfectly rational agents would do, there remain deep issues of principle. The main problems concern complexity. Actual human social situations usually involve a large number of potentially concerned people, even if only a small number play a major role. They usually present a large number of options to each agent, in fact quite routinely infinitely many. (Stop, walk, run slowly, run fast, run as fast as you possibly can, run fast but not as fast as you possibly can.) Moreover, the consequences of the combinations of the different agents' actions usually ramify into the indefinite future. And they usually involve many aspects about which agents do not share the same information. In other words, considered simply as problems in game theory they are daunting to the point of impossibility. (Existing game theory can handle complexity, if not very plausibly, but we are at the edge of our understanding when dealing with choice given non-shared or false information. For very different takes on this see Myerson 1991 and Williams 1995.)

Consider chess. From the point of view of game theory there is nothing particularly remarkable about it. There is a best strategy, for black or for white. We just happen not to have computed this strategy yet, stymied by inessential combinatorial obstacles. But from the point of view of actual human chess players the situation is completely different. Their problem is not to find the ultimate winning strategy but to beat particular opponents, who also do not know the ultimate winning strategy.

Yet we do think our way through strategic problems, a hundred times a day, with good enough effect. We clearly do not do this by aiming for complete uniformly correct solutions. They're beyond us. Instead, we embody a collection of loosely linked procedures that provide good enough approximate solutions quickly enough to be of use. And since the real problems we face are very complicated, we apply these adequate heuristics to problems that would take superhuman power to get complete and accurate solutions for. While game theory aims at complete and accurate solutions to simple problems, the strategic reasoning that comes naturally to human beings aims at tolerably incomplete and bearably approximate solutions to hard problems.

These strategic heuristics are the essence of microethics, I believe, and form an essential part of the background for our grasp of mind. The comparison with game theory brings out essential features of the problems they have to solve, and also some very abstract features of the solutions. Some of these features are shared by examples that invite comparison with different high-order intellectual accomplishments, notably ethics of the large-scale richly conceptualized kind. Consider another very everyday example.

Someone is helping you move a table through a narrow doorway. As the door approaches he shifts his grip on the legs in a way that would make sense if the plan was to stick the legs round the corner to the right. You shift your grip accordingly. When you get to the door he shifts again, the other way round, and begins to rotate the table as if the legs were going round to the left. You have to either re-shift your grip or overcome his turning by turning the legs hard in the other direction.

The behavior of your helper is annoying. It is a microethical analog of a broken promise. He generated an expectation which you acted in accordance with and then did not fulfill the expectation. Every child learning to do things with others learns about generating expectations and then sticking to them, and about picking up intentions and then sharing and relying on them. (See Velleman 2000.) Of course the principles the child picks up are subtle, context-linked, and shaped by cultural factors. In some societies a parent or other authority figure helping a child with a task may be able to shift intentions without generating resentment. Indeed in the story as I told it, it may make a difference that it is the helper rather than the person to whom the task belongs who is shifting their intentions.

Many similar everyday examples could show microethical analogs of abuse of trust, of violations of dignity, and injustice. There is a rough division into examples that suggest misperformance of a shared project, and those that suggest ineligibility or misappropriation of the role of participant. (Bungled cooperation and controperation.) And of course there are corresponding virtues and successes. I have been using examples where the communication between the participants is non-verbal, in order to help make the case that some of the sources of microethics are very basic to the way we think. In any specific case it would be foolhardy to have too definite opinions about the source of the intuition or the competence, but it requires a lot less courage to claim that there is a human inheritance of capacities for regulating cooperative activity. The cheater-detection mechanisms postulated by evolutionary psychologists are a good example, both in their general plausibility and their speculativeness as applied to any particular case. (See Cosmides and Tooby in Barkow *et al.* 1992).

The microethical intuitions in the table-moving case and similar examples center on expectations that people form of one another's actions. They are thus closely linked to prediction and explanation. (Explanation is the topic of Chapter 4.) But it is important to see that the expectations are not prior to the moral or strategic intuitions. You do not predict that your helper will turn the table to the right and then conclude that this is what he should do. Instead, you analyze the situation in a way that could be over-intellectualized as follows. ‘The important thing is that we both turn the table the same way when the time comes. There may well be some advantage to one way of turning. He is acting as if for a right turn, which at any rate resolves our coordination problem and may indicate the better choice. So that is what we should do. So I’ll move my hands accordingly. Then that is the agreed course. So he will stick to it.’ A judgment about the right course of action, the right joint course of action, is prior to the expectation about what he will do. Of course this is just one of several variant intellectualizations, shaped by a particular contemporary conception of moral thinking. (And that conception is itself shaped by conclusions about strategic reasoning.) But any plausible way of articulating the reasoning concerned will make it center on a solution to a coordination problem, from which the decision about what you should do and the expectation about what the other person will do are both derived. (That is why surprise and disapproval are so mingled when he doesn’t act as expected.)

The general pattern is this. One finds oneself in a situation in which the actions of several people, oneself included, are delicately inter-dependent. One characterizes the situation. (Possible characterizations, projecting downwards from larger-scale ethics are: coordination problem between cooperators, dangerous encounter with a hostile/manipulative/unfathomable other, conflict of interest with a potential ally, test of authority, situation of potential mutual benefit, and so on.) One then finds a solution or good outcome to the situation, drawing on innate social capacities, one’s stock of previously solved situations, general criteria defining good outcomes to situations of that character, and one’s abilities at discursive consequence-uncovering reasoning. From this solution one derives a decision about what one shall do, and expectations about the actions of others. (Evidence for the existence of innate tendencies to cooperative behavior are found in Turiel 1983, Dunn 1988.)

There are many ways of filling out this general pattern. One is of course to find an analog of the game-theoretical equilibrium (bearing in mind that in many cases this will lead to rather than be based on attributions of desires to the agents involved, as argued above.) Another is to grasp intuitively a morally good outcome, or to think it out explicitly. Of course this will not lead to an expectation about others, and perhaps not to a decision on one’s own part, unless one has characterized the situation as an interaction of cooperative and morally sensible people. (But if one has, the expectation and the decision may be unhesitant and natural.) Another is the opposite of what conventional language would describe as moral. One may characterize the situation as that of a confrontation with an enemy, and go for a pre-emptive strike – giving oneself as much benefit at the expense of the other as possible. One possibility is that one then acts as if the enemy also realizes that this is the thing for one to do, so that one’s aggression is one side of an equilibrium, paired with the enemy’s best defense against it. It is worth noting that many such good solutions faced with an enemy would be taken as the right and honorable course of action in many cultures, past and present. (The enemy's reaction if one did not pursue one's advantage would not be moral admiration but contempt for one's stupidity.)

We sometimes make joint decisions and expectations in ways that fit this pattern, surely. And just as surely we sometimes do not. Sometimes, for example, we follow an essentially non-strategic pattern of thinking. First we attribute beliefs and desires to the other people concerned, in ways that have nothing to do with the situation at hand (sometimes because they have told us what they think and want). Then we introspect, and come to conclusions about our preferences over the available outcomes. And then, finally, we explicitly think out what we should do and, independently, what to expect the others to do. There is no real point or interest in asking how often we reason in which way. The important, and hard, questions concern priority and fundamentality. Which ways of thinking about other people can make which other ways possible? Which ways are required as background or preparation for which other ways? I believe that microethical thinking of the kind described is part of the essential human background for the ascription of states of mind. We with the resources of our species would find it next to impossible to have and ascribe the concepts we do without it. The much more restrained conclusion that this chapter is moving towards concerns only the ascription of action-descriptions and intentions. Microethics affords a natural and efficient basis for forming expectations about people's actions, and we have to base *some* of our ascriptions on it.

Pre-emptive objection; intention-clarifying reply

Objector

I can see where you're going. You want to argue that some forms of strategic reasoning are more basic than the concepts of folk psychology. Eventually you are going to tell us that it is because we are capable of managing strategic situations that we can progress to applying belief, desire, preference and other psychological concepts to one another.

But there is a fundamental and insuperable problem here. In order to carry out any form of strategic reasoning, at any rate any that anyone has ever described, you have to know the preferences of all people involved. As a result the reasoning cannot be really primitive. It has to presuppose a prior ascription of preferences.

Defender

This is to misunderstand the way the modeling works. We start with a strategic situation, in which there is a certain interdependence between the choices of a number of agents. We then suppose two things. First that the situation can be described in terms of something that is broadly similar to preferences for the agents concerned, over a common set of outcomes. And second that the agents are equipped with cognitive routines which will come up with choices of actions which will be in some ways, about which there is a lot more to be said, satisfactory, given their ranking of the outcomes. These routines may not do anything that can be described as ascribing preferences.

There are many ways this can happen. Suppose that in fact the agents do have preferences and that what makes an outcome satisfactory is a matter of its relation to these preferences. Then one possibility is that there is evidence that indicates what the agents' preferences are likely to be, and which is accessible to the decision-making cognition. The cognitive process can work directly from the evidence, however, without having to detour through the ascription of preferences. (Just as a program that sends an animal out hunting on some nights and not others may be sensitive to information that is evidence for the relative positions of sun, moon and earth. It then chooses an action when these three are related so that the moon reflects a lot of light to the earth. But the animal is not attributing relations of non-occlusion to heavenly bodies.) Another possibility is that the cognition is sensitive to factors that affect the well-being of the agents concerned, which tends to be well-correlated with their preferences. It may automatically reason in a way that has the effect of weighting drinking more highly than thirst, a safe place to sleep more highly than a night in the open, or joy more highly than despair. It may be a specific routine for solving some particular class of strategic situation, and be sensitive to some specific kinds of good. For these situations, in which these goods are at stake, it may calculate the expected choices of interacting agents, which then regularly correlate with satisfactory outcomes.

Or, alternatively, the processes that determine agents' choices could be directly sensitive to their happiness, or flourishing, or some other basic aspect of the good for them. If we are preserving our neutrality on the first unbeggable question, we should not rule out the possibility that in some or many cases the major determinant of action is not what the person wants but what is in some respect in their interest. I explore this possibility in Chapter 4.

Efficiency and inevitability

Consider an agent in a complex strategic situation, one in which there is a high degree of dependency between the actions of the interacting agents. Important features of the outcomes for each agent depend on the choices made by each other, important features of which in turn depend on the choices of others. (In the simplest case one agent's action might be determined independently of the choices of others, which themselves depend on it. In more complex cases some agent's choice might be narrowed down to a small list independently of the choices of others. In yet more complex cases each agent might have a wide range of actions, choice within which cannot be made without reference to the choices of others. But do not think the complex cases are rare.) In the section 'Strategic problems' I contrasted two general approaches to such problems. On the one hand, each agent can think out the motives and reasoning of each agent, and try to predict what decisions will result, as part of thinking out what would be best for her herself to do. Or, on the other hand, each agent can try to define some equilibrium outcome, focussing on the properties that a situation would have if each person were reacting to the motives of each other but not directly representing the motives, and extract both expectations and decisions from it. Call these the *motive-based* and *solution-based* approaches. In the section 'The flesh and blood version' I argued that this contrast doesn't depend on the technical ideas and simplifying assumptions of game theory.

There are two vital disadvantages to the first approach, expectation via motives. The first is its complexity. In even quite simple situations each agent will have to consider the thoughts that each other will have about the thoughts each other will have about the thoughts of each other. (And less simple situations can be much less simple.) Thinking in terms of these multiply embedded propositions is inevitably demanding of time and cognitive power, and the risk of error is inevitable.

Complexity considerations are taken further in Chapter 5 and its appendix. For the moment, notice the close link between the complexity of motives that have to be taken into account and the number of interacting agents. In a two-person situation with a relatively low degree of dependency between the agents' actions, each agent may have to consider what each other thinks about her motives. But in a three-person situation with the same degree of dependency each agent will have to consider what the first other agent thinks about what the second other agent thinks about her motives. The more people the greater the likely embedding of motives, and the greater the potential complexity. (See Kinderman, Dunbar and Bentall 1998.)

The other disadvantage of the motive-based approach is that it often doesn't give an answer. The simplest example is a situation like the following. There are two acts available to each of two agents, and the consequences for both agents will be bad if they do not choose the same act, but the consequences of both choosing one of them are better than those of both choosing the other. (We can escape through the woods or along the river bed. If we take different routes we are done for, but we'll be safer at night in the woods. But we must now run from our separate cells, hoping to meet in the woods or by the river.) No amount of practical reasoning will deliver each agent a prediction that the other will choose the act with the better result. (An easy way of seeing this. The conclusion 'she will do A' can never be reached unless the conclusion 'I will do A' has already been reached. For it would be irrational to choose A if the other were not also going to. But the conclusion 'I will do A' cannot be reached unless the conclusion 'she will do A' has already been reached.) Yet human agents put in such a situation will have no difficulty deciding to do the better act in the expectation that the other person will too. (You have just got out of your cell. You know the facts about the woods and the river bed and you know that your accomplice does too. Is there any question about which way you should run?)

Solution-based approaches have neither of these problems. Strategic situations can still be quite complex – again see Chapter 5, and especially the appendix to that chapter – but will always be simpler than on a motive-based approach. All one has to do is to identify the appropriate satisfactory solution and then see which acts by each participant will lead to it. (This does *not* mean: use an apriori notion of a good solution and apply it to the situation. As should be clear from sections 'The flesh and blood version' and 'Pre-emptive objection; intention-clarifying reply' it means: find the class of solutions towards which the motivation of the people involved is motivating them in this particular case. See the remarks towards the end of Wiggins 1998 in this connection.) And even if it is defined in terms of the desires or preferences of the people concerned it can be got by simply comparing them, rather than by following out complex patterns of reasoning. Moreover, finding satisfactory solutions lends itself to simplifying heuristics, shortcuts. This fact too is developed further in Chapter 5, but it is easy to see that, for example, an approximation to a game theoretical equilibrium can be found in many situations simply by asking of an outcome 'will everyone concerned see that this is something they would accept, if it came about?' And it is clear that this sort of heuristic applies to more informal kinds of satisfactory solutions too. For example if two people are dividing some cash they found on the street, each knows that though each would like to get as much as he can, neither would protest if each ended up with half.

Solution-based approaches do not have the second problem either: they more often produce answers. When motive-based accounts come up with predictions that are in harmony with decisions that are based on the same information, the result describes a satisfactory outcome. (When prediction and decision are not in harmony, something is deeply unsatisfactory.) So whenever a motive-based approach gives a prediction, a solution-based approach will too. But the converse is not true. For example in coordination problems like the one described above, a combination of actions that produces optimal results for all concerned is easily picked out. (And it is easily defined, to a first approximation, as an equilibrium in which each agent is better off than in any other equilibrium.) So a technique of deciding-and-anticipating that fixes on it will give a right answer here, though no purely motive-based account will. (You have escaped from your cell and must run to the woods or the river bed. You think 'we'll be better off in the woods', and so you head there, expecting to meet your accomplice there. The procedure that comes up with this as a satisfactory outcome has taken into account your shared preferences and information in presenting this as the kind of thing to go for.)

If one was designing a creature along the general lines of a human being, needing to make quick judgments about what other such creatures were likely to do, one would surely build in some basic capacities for solution-based expectations. One would do this if only as a basis for a minimal cooperation and coordination between individuals, which would later allow more finely tunable and conceptually more demanding developments. (One would build in a canny yet naive sociability like that of a two-year-old, in order that the later grasp of the concepts of belief, preference and reasoning could safely develop. And one would build in some degree of *un*predictability, an idea that is taken further in Chapter 5.) This does not prove that actual humans are designed in this sensible way. Nevertheless, one can make a strong case for the claim that *some* aspects of our concept of mind have to be founded on solution-based thinking.

The argument focusses on the classification of actions. I assume that a large number of the situations in which we have to anticipate the actions of others are strategic, and often to a considerable degree. I also assume that we have limited time and intellectual resources to bring to them. Consider now the ways in which actions are classified. Suppose that they turn entirely on bodily motions and people-independent consequences (bending one's thumb, setting the forest alight). Then when facing any situation in which another person's choices are crucial one would have an indefinitely wide range of labels to apply to the actions of the other people, and which they could apply to one's own actions. The open minded approach would be to take all of the labels which designate actions that a person could perform to pick out actions to be considered. Thus one might consider whether a person would bend a thumb, bend his right thumb, bend his left thumb, bring his thumb towards one of his fingers, bring his thumb towards the third finger of his right hand, hold a match between the thumb and third finger of his right hand, strike a match, make a match light, use a lighted match to start a fire, light a fire, start a fire which causes the woods to light, set the woods on fire, and many many more. And, worse, one might consider whether the other person expected one to expect him to do any of these. And so in solving the strategic situation one would be trying to think out a situation in which each agent had to choose between an enormous number of acts, and had to define each other person's attitude to each person's performing each of an enormous number of acts. The simplest situation would be utterly unmanageable.

We obviously do not allow situations to be this unmanageable. We apply filters to the range of action-descriptions applicable to any situation. (We ask: will he light the woods, or not?) We pass very quickly and automatically from the full conceivable list of possible actions to a short list of candidates to be given serious consideration. The first conclusion then is that we have to apply filters to the number of actions under consideration.

Not any filter will do, though. The range of strategic situations is so great, in terms of the kinds of actions involved, the number of people involved, and the consequences of the combinations of the actions of the interacting people, as to rule out many filters. Classifying actions in terms of some salient human consequence – whether they would lead to a death or not, whether they would cause someone to look ridiculous, or the like – will not do. For any finite list of such consequences there will be situations in which the important differences between actions do not occur on the list. And classifying actions in terms of some psychologically straightforward feature such as the order in which they occur to one, just in order to produce a manageably short list, will fail for similar reasons. The features in question will not consistently separate those actions between which choice is crucial. In fact, the features of actions that a usable filter must fasten on to are pretty abstract. They must in particular be independent of the number of interacting people, in that one can expect strategic interactions with anywhere from two to a dozen others. (Why these numbers? There are reasons to believe that we evolved in contexts in which our primary capacities were directed at interactions within groups of specific limited sizes. See Dunbar 1996.)

There is only one way, as far as I can see, of meeting the constraints on an acceptable filter. Actions must be classified in terms of their characteristics as elements of strategic situations. For example, in simple game theoretic terms we could classify them as solutions to coordination problems, as Nash (and other kinds of) equilibria, as Pareto optima, and so on. Of course these are both too technical and too simple, as I argued in the previous section. They leave out vital factors like the degree of shared knowledge and the extent to which other agents understand the situation. But they are the right *kinds* of characteristics. Even more appropriate characteristics are reflected in the everyday language of action. We speak of actions as helpful or obstructive, friendly or hostile, defensive or aggressive, and so on. These are labels that can be used to classify actions in an indefinitely wide range of situations, involving any number of people. So given an arbitrary strategic situation, a person can ask whether she should help, hinder, defend, attack and so on. (See the longer, structured, list below.) By seeing which of these labels are relevant to the situation she can see what kind of a situation it is. And then by deciding between the ones that are relevant she can ensure, in many cases, that she is choosing in a way that does in fact distinguish the vital from the irrelevant aspects.[4]

This argument may seem to be another description of how one would ideally design an intelligent social creature. But in fact it describes how any intelligent social creature operating with limited cognitive resources must approach strategic problems. (It describes one small aspect of how any limited creature must operate.) Let me rephrase the argument in schematic form, to make this explicit.

Any limited creature can consider only finitely many options.
In any choice situation there are infinitely many options.

Therefore any limited creature must possess ways of reducing the number of options it considers.
Any way of reducing the number of options must apply to an indefinite range of situations.
For every characteristic of actions that concerns only their intrinsic character as actions there are infinitely many situations in which this characteristic makes inappropriate distinctions.
Therefore any limited creature must reduce the number of options in accordance with a classification that relates actions to other actions.
The relevant other actions are those of other agents, which are chosen in accordance with the classification in question.
Therefore any limited creature must reduce the number of options in accordance with a classification that relates actions to aspects of strategic situations.

The argument shows that any limited creature engaged in an indefinite range of social interactions will need a particular aspect of the situation-characterizing devices which earlier in this section I argued a well designed social organism should have. As befits an apriori argument, it leaves the details of implementation unspecified. So we cannot deduce from it exactly what strategic characteristics of action we can expect to find filtering choices. Nor should we be able to; the task depends on the fine-grained psychology of the creature in question. We humans do it in ways that reflect our inherited focus on sharing food, cooperative hunting, group rivalry, and contested authority, to name just the most obvious themes. As a result, we think of actions as falling under headings such as the following:

chase, catch, strike, see, flee
quarrel, threaten, frighten, help
attack, protect
take, grab, share
lead, follow
prevent, cooperate, guide, obstruct
show, hide.

These are extracted from the ordinary English vocabulary of action. But I am confident that all languages have near synonyms of almost all of these. (More generally, all languages will describe actions in terms of the fulfillment of intentions, as described from the point of view of others affected by them.) And all three-year-old children in all cultures have mastered concepts like these – before they have mastered the concepts of belief, preference and inference – and use them as the fundamental basis for their grasp of other people. The most primitive psychology rests (in part) on microethics.

Joint attention and mutual knowledge

The argument has been that some action-classifications must be related to strategic action. Psychological thinking requires that actions be conceptualized. They are what is predicted and explained, and used as evidence for attributions. So in this way the capacity for strategic action helps set the stage for psychological thinking. There is psychological evidence for another link. The evidence is found in the literature on joint attention.

From the age of about eight months children begin to follow the gaze of a person who has first made eye contact with them. They will look in the direction that person looks. This capacity becomes more frequent with age, and soon after they begin to look in the direction of the person's gaze and then back at the person to check that they have correctly identified the object of attention. Later, typically between 9 and 12 months, children begin to gather information about other people's attention while performing activities, for example by pausing at suitable moments to check whether the other is attending to them, to some other relevant object, or is no longer sharing attention. The progress of increasingly skilled monitoring of the attention of others is correlated with the beginnings of imitative learning and referential language. A causal hypothesis offered by some psychologists is that joint engagement (gaze-following) is a necessary precursor of communicative gestures which themselves make more complex attention following and imitative learning possible. All of these are then precursors of the acquisition of referential language. See Carpenter, Nagell and Tomasello 1988. (For a helpful philosophical commentary on developments in joint attention see Bermúdez 1999.)

The behaviors and capacities that develop in this sequence are common to a number of crucial human attributes. They are implicated in the child's first attempts at linguistic communication, at cooperative activity and at psychological attribution. I shall say no more about language, though it is clearly a vital element in the combination. The details of the route to psychological attribution are murky and controversial, but the reason why it seems so suggestive are obvious. When the child begins to think of the world as overlayed with the changing attention of others she begins to think of others as being in states that can be indexed by the things in the world to which they are directed. She begins to grasp intentionality: human beings can be understood by connecting them with objects and situations. She is thus on the way to a precursor of the concept of belief, since she can think of a person's attention as being directed towards or away from a particular fact. (She is near to the concept of ignorance, though still a long way from the concept of error. She is nearest to what Perner 1991 calls 'prelief'.)

It is significant that the attention whose direction at some objects and not at others the child learns to appreciate is shared. The child learns to separate out the other's attention from a triadic structure in which her own attention, that of the other, and the location of the object are all involved. And the contexts in which the learning takes place are ones in which the child and others are acting cooperatively, coordinatively, or antagonistically (this last usually in pretence). Playing games, following demonstrations, helping in practical activities: in all of these the child is acting on objects, in a way which requires the child to take account of the directedness of others to them. Referential states of both self and others can thus be appreciated together. (It would be very enlightening to explore the developmental effects of one-sided gaze following: where the child is encouraged to follow the gaze of others but where the child's own gaze is not followed in social and communicative contexts. One could not ethically create such extended situations: do they occur naturally?).

At this point it would be nice to be able to refute an implausible hypothesis. According to that hypothesis the *only* shared activity required to get on the psychologizing path is that of shared seeing. From shared seeing one could derive an appreciation of one person seeing and then of perception, then via the concept of inference that of belief. (Baron-Cohen 1995 seems to encourage such a hypothesis.)

I take this hypothesis to be a version of the radical realism about commonsense psychology that in the section 'Questions not to beg' I resolved neither to believe nor to reject. It does seem to me very implausible, though not really refutable. One problem is the inferential relation that is to connect perception and evidence. No epistemologist has ever made sense of a sufficiently powerful and inclusive relation. Contemporary epistemologists tend to describe the formation of belief as a much more social business, appealing to processes and virtues which allow people to transmit and combine their information. (See Schmitt 1998 and Zagzebski 1996. The issues are similar to those that will arise for practical reasoning in Chapter 2.) Another worry about the view is that a primary function of shared attention must surely be to prompt shared attitudes and responses, for example to fearful objects. It seems incredible that a child should not be able to progress from registering what her mother is perceiving to sharing her mother's fear of that object, without having to go through attributing a belief to her mother.

In any case, to acquire a concept of mind one has to acquire many concepts besides that of belief. One has to understand desire, intention, anger, kindness, fear and many others. Most of these are intentional, in that they relate people to objects. In most of them, the relation involves action directed at the objects. And in most of these, the natural context for the action is that of shared projects, competition, or play. It is hard not to suspect a two-way scaffolding: learning complex interactive routines teaches one how actions are directed at objects, and this understanding then facilitates further such routines.

A deep and interesting concept that arises in this context is that of common knowledge. In many cooperative activities, including some remarkably simple ones, one needs to be aware not only of the facts of the case, but also of which ones the other people concerned are aware, and of which ones they are aware that they are all aware. And so on. But this seems to be impossibly demanding. The standard solution to the problem is to require that the information available to participants in a shared activity be available in a form that allows them all to deduce from the obvious facts of the case as many of this infinite series of facts about awareness as they need to. The standard examples of this work from a perceptual basis. Suppose (an example due to Stephen Schiffer 1972) that there is a lighted candle on the table between us. Each of us knows that it is there, and knows that it is there within easy sight of the other, who has good vision and a functioning brain. So each can figure out that the other knows it is there, and by reflecting on the fact that this reasoning is available can figure out that the other knows that they know it is there. And so on. So we do not need to keep more complex thoughts about iterated knowledge in our heads than we need. But when we need them we can deduce them.

At this point links with joint attention form again. The paradigm situations in which common knowledge is accessible to a pair of people are ones in which some object is perceptually salient to each of them, each can perceive that the object is so salient to the other, and each is psychologically equipped to register these facts and their consequences. Creatures not so equipped would lack a basic skill for shared activity. *The capacity to identify the objects of shared attention is a basic device for finding solutions to strategic problems, which is also a crucial requirement for the acquisition of mind-ascribing and linguistic abilities*. The way the child's use of joint attention mediates her participation in playful and practical activities makes it clear that she is capable of using the fact that something is visible to two or more people as a basis for activities which require that the fact that it is so visible also be available to both. (Heuristics which mimic the effects of mutual knowledge will often not register the fact that e.g. o is visible to a and to b but the fact that o is visible to a cannot be deduced by b. I would conjecture that in many interactions infants act as if there were mutual knowledge even when, as in these situations, there is not.)

Among the cooperative activities that require mutual knowledge is language, as many writers have argued. (See Lewis 1969, Schiffer 1972, Sperber and Wilson 1986.) Without an awareness of what is known to all participants in a conversation (and known to be known, etc.) one cannot distinguish between what is said and what is communicated by the fact of saying. Thus it is very plausible, as several other writers have argued, that a specific capacity for language allows a child to have the effect of patterns of reasoning that are in advance of what she can think through directly. Among these facilitated reasonings is the attribution of states of mind (Harris 1996.) So here is one special purpose strategic reasoning routine which at the very least plays a strongly facilitating role in the acquisition of mind-ascribing capacities.[5]

The importance of being understood

I have been concerned with a very basic level of our grasp of one another, more fundamental than most of what is considered under the heading of ‘folk psychology’. To end the chapter I shall discuss one point of continuity with more sophisticated levels of understanding. It is best described in terms of a contrast between what I have been arguing and a premise of much standard philosophy of mind.

Philosophers often assume that the central point of having a concept of mind is to be able to predict the behavior of others. Human beings, on this assumption, developed the ability to think of one another in mental terms so that they could anticipate threats and pursue coordinated action by forming accurate expectations about one another’s actions. And in the development of each individual human the conceptualization of mind, through maturation and through learning a culture, brings benefits to the individual largely through the ability to know what others will do.

Daniel Dennett is just one of many writers who make this assumption. As he puts it:

> 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.
>
> D. C. Dennett ‘True Believers’, in 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*, section VIII, Part I

My position is very subtly at odds with what Dennett and Hume appear to be claiming. I am not denying that it is of great benefit to know what others will do, and that we often do have this knowledge, largely as a result of attributing states of mind. But I am denying that the power of folk psychology comes primarily from the ability to make bare predictions of future behavior. The point is best seen in terms of examples such as the one prefacing the introduction to this book. Exaggerate it: you are confronted with an environment of lunatic car drivers. They seem completely unpredictable. Perhaps some of their choices are made completely at random; others are driven by motives that you cannot fathom. (Even if you could fathom them they might not tell you what the people were going to do.) But you do know a few things. You know that they do not want to collide with you, if they can help it. (You know that blood spoils the appearance of a freshly painted Alfa Romeo. And the delay subsequent on a collision tends to be undignified and make one late for parties or assignations.) And you know that they will take precautions to avoid a collision. So what you can do is this. You head out slowly into the traffic, making your intentions as clear as possible and acting in a collision-minimizing way. Other drivers will then be able to avoid you, and they will, as long as they can anticipate your actions. What your survival requires is not that you predict their actions but that they be able to predict yours.

(A party of drivers from Düsseldorf or Kansas City is approaching the Piazza Venezia or Times Square. One arrives first and, horrified by what he sees, adopts the tactics above. Then another, and another. Soon half the drivers are being conspicuously predictable. It is now in the locals' interest also to be predictable, so they adapt temporarily to the new style. The last strangers to arrive are amazed at how orderly the traffic is. 'Not as we were warned at all; these people are easy to predict'.)

The point is this: we often cannot make simple predictions of the form 'P will do *a*'. When we cannot we often still can make a special kind of conditional prediction 'P will do *a* if I do *b*'. By acting in particular ways we can create the conditions under which people are predictable. Often a vital element of this consists in giving the person an understanding of your motives and course of action. The result is a degree of mutual understanding, created by practical psychological thought.

This chapter has described only some very simple ways in which we create mutual intelligibility. It has focussed on ways in which a person can act as if others were also going to find their way to a solution to a common problem. The person then acts in a way that will be intelligible to the others if they do in fact find their way to the solution. Then, as in more complex cases, the person forms an expectation about the other person's actions. The expectation is not a simple prediction, in that it is not independent of the choices of the predictor: in arriving at it one is also choosing one's own action. And it is usually conditional: if one does not carry through with one's own chosen action the other person is likely to abandon or backtrack on theirs. Moreover, both in these processes and in more fully developed cases of making intelligible one's action often has the general character of a promise, since one is (often) acting as if there were mutual trust. (Baier 1986 emphasizes the importance of trust as an element of the conditions in which the understanding of mind can develop. See also Holton 1994, Jones 1999, McGeer 2002.)

Acting as if both sides were moving towards a common solution is a simpler business than moving towards the solution in order to bring the other side to see it. The latter process requires an understanding of error, communication and reasoning. It is part of the more general and more normal case in which cooperation and understanding are almost inextricably entwined. We are able to get along with people – or frustrate them – because we can think about what they believe, want, intend and reason. And we can have these thoughts *in part* because we are embedded in routines of cooperation and defense. So there is no easy way of picking apart the creation and the discovery of intelligibility. In subsequent chapters I will not argue for any simple relation between them. I shall discuss the attribution of reasoning, beliefs and intentions, and in each case I shall try to highlight aspects in which the fact that we understand one another is a result of patterns of behavior that favor shared projects. I am not claiming that these aspects are more than a small part of the whole story. What I say about them is quite simple; I summarize it in the middle of the book in the summary, 'Where we've got to', following Chapter 5. But it may seem confusing if you take it as a sketch of a complete account of belief, reasoning, or the attribution of states of mind. So, please, do not. I am trying to trace some strands in a complex pattern. I hope to persuade you that they are an essential part of the whole, that our everyday understanding of ourselves would be a very different business if there was not a two-way shaping between our roles as understanders and understood.

APPENDIX: PREFERENCES FROM CHOICES

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, so 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 such facts about situations, and that agents can represent choices and outcomes, and can carry out quite complex conditional thinking ('if A or B happen then as long as C does not happen 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 be able to recreate 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 is given below. (Domcord 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 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 unavailable 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_1,b_2) is preferred by A to (a_1,b_2), just as the fact that the right choice in the second is a_1 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 realizes 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 characterize 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 a_1 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	a_1	3,0	0,2	0,1
	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 one's 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 dominance is embedded within coordination within dominance (just as in Domcord II coordination is embedded within dominance), 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 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 is 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.

The important conclusion must be that agents making optimal strategic 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. So if a creature possesses resources for making good decisions in strategic contexts and if it is also capable of deductive reasoning, then it will be able to ascribe beliefs and desires. (The primary focus is on the ascription of preferences and reasoning. Beliefs would be a corollary.) The idea of a logic-using creature for whom strategic choice is more basic than psychological ascription may seem alien to us. It may seem that this is the opposite of the human order. I think this is a mistake. What would be very distant from human thinking would be to have an *all-purpose* facility for strategic choice that did not draw on the ascription of states of mind. When we humans want to apply the same way of thinking to a range of strategic situations we have to use the vocabulary of preferences, beliefs, intentions, choices and deliberation. This is our higher level procedure for checking, correcting and comparing choices. But we have many special procedures for making smaller ranges of strategic choice. Many of them are very fallible heuristics, but that does not block the reasoning that I have been describing. (It does raise the likelihood of conflicts between ascriptions based on different local procedures, though. In particular, I do not know how the considerations of this appendix would apply to heuristics for making choices in a series of strategic interactions, such as the famous Tit for Tat procedure.) The actual human order of things, I suggest, is that many individual psychological ascriptions are based on such local heuristics, the presence of which scaffolds a capacity for ascription in general, whose mastery then makes possible the higher level reasoning that when needed corrects the basic heuristics.

Mine is not the only way of getting preferences out of choices. It is worth comparing it to the standard routines of Ramsey, von Neumann and Morgenstern, and Jeffrey (See Ramsey 1931, Chapter 3 of von Neumann and Morgenstern 1944, Chapter 3 of Jeffrey 1983. Jeffrey's discussion is particularly helpful. This appendix is based on a section of Morton 2001b.) These all work with non-strategic ('parametric') choice, and present agents with choices between gambles, from which their preferences and probability assignments can be deduced. There are three contrasts with the reasoning I have described. The first is just that they work with non-strategic choice. The second is that their raw material is choices between gambles rather than the conceptually simpler choice of basic actions. The third is that their output is numerically precise degrees of preference and subjective probability. This third is significant. The task of these non-strategic procedures is to determine, when an agent prefers A to B, to figure out by *how much* A is preferred. The present procedures, on the other hand, aim to establish preference itself. For they operate in the strategic context, in which it is not obvious from an agent's choice between simple options what her preferences are.

Chapter 2

Motives and virtues

> There is greater variety of parts in what we call a character, than there are features in a face: and the morality of that is no more determined by one part, than the beauty or deformity of this is by one single feature: each is to be judged of by all the parts or features, not taken singly, but together. In the inward frame the various passions, appetites, affections, stand in different respects to each other. The principles in our mind may be contradictory, or checks and allays only, or incentives and assistants to each other. And principles, which in their nature have no kind of contrariety or affinity, may yet accidentally be each other's allays or incentives.
>
> Samuel Butler, *Sermon XII*

The place of this chapter

When we move beyond the simple situations discussed in Chapter 1 the relations between psychological understanding and patterns of cooperation and rivalry are much harder to make out. Still, some of the flavor of the simplest circumstances remains. Let me make two claims, which sum up this flavor.

The common focus claim

The concepts we use in everyday psychology have a bias towards ways in which one person's actions impact on the lives of others. So folk psychology aims to explain situations that are of moral significance.

The point about the vocabulary of action made in 'Efficiency and inevitability' in Chapter 1 thus generalizes not to psychological concepts but to psychological *explanations*. We are interested in explaining cooperation, betrayal, trustworthiness and the like; accomplishment, disaster, stagnation and the like. The influence on other everyday concepts will thus be indirect: they have to be able to figure in explanations of the things that matter to us about one another. The things that matter include evils as well as goods.

The good effects claim

When people understand one another in accordance with the expectations that everyday psychology gives them, the results will generally be good. ‘Good’ and ‘understand’ are deliberately vague and naïve here. The claim is a placeholder for a scattering of specific sharper points that will be argued for.

This chapter develops both claims in the context of attributions of practical reasoning. I argue that usually when we explain why someone adopted a certain means to an end, or explain an action in terms of their seeing something as a means to an end, we draw on a stock of concepts that describe desirable and undesirable ways of thinking. Many, but not all of these are fairly described as virtues. Some are vices. The common focus claim is then the claim that our attributions of means–ends reasoning go hand in hand with morally relevant characteristics. In accordance with the good effects claim I then argue that there is an additional effect biasing the characteristics that we frequently and easily refer to towards those that it is in our collective interest to possess and be familiar with.

It will take a while, though, for the connections with these two claims to emerge. In the earlier parts of the chapter I shall argue for a negative claim, that there is no simple pattern connecting beliefs, desires and actions, alone, which we can use to explain and predict what people do. I shall consider standard examples about people who want a beer, know there is one in the refrigerator and walk over to open the fridge door, or want to escape from a sinking car and open the window, or want not to be checkmated and sacrifice their queen. The initial negative claim is that there is no general pattern common to such examples, if we require that it be expressible purely in terms of beliefs and desires. This leaves a gap, which other concepts such as skills, character, and most significantly virtues, have to fill. At that point the positive arguments for claims of common focus and good effects can enter.

The negative bit

The flavor of the negative point can be got from a simple example. Consider the problem posed to small children in a typical ‘false belief’ task. (In these tasks the children have to grapple with a fact that is very hard for them, that people can operate on the basis of erroneous information. They have a really remarkable degree of difficulty with it. See Perner 1991, and the equally remarkable experiments reported in Gopnik 1993. A good exposition is in Byrne 1995. But the point here is not that fact but the kind of task that is used to illustrate it.) An object o is moved from location l_1 where an agent A saw it placed, to a location l_2. Where will A look for it? The answer is supposed to provide an index of the belief that is attributed to A. For if A thinks that o is at l_2 then surely A will look there.

But the inference is *not* obvious. Consider all the reasons why A might look elsewhere, though thinking that o is at l_2. A might think that though o is at l_2, it just might be at l_1, so that it makes sense to search at l_1 first. A might think that though o is at l_2, it would be awful if it were at l_1, so that that possibility is to be excluded first. A might think that though o is at l_2, it would be wonderful if it were at l_1, so that possibility is to be investigated first. And so on.

This example may convince you that considerations of the form 'that is what she wanted and this was a good means to it, so this is what she did' are almost never complete and sufficient grounds for explaining or predicting an action. The interesting questions are then more subtle. How important a role do such considerations play in explanation? How near to adequate are explanations appealing only to them? What are the default assumptions that allow us often to state only these principles even though many others must be tacitly appealed to? (If you are convinced of this, then you may wish to skip to the next section, avoiding the argument by exhaustion of possibilities that fills the remainder of this section.) Consider the following simple claim:

1 Suppose that a person has a desire which we can express by saying that the person wants (desires, would like) p to be true, and believes (thinks, accepts, takes to be true) that if she performs A then p will come true. Then she will perform A.

This is obviously false. No philosopher believes it in this form. I want a million dollars and believe that if I poison my grandmother I will have a million dollars. Knowing this about me doesn't give you the slightest expectation that I will poison my grandmother, unless you also know other more interesting things about me. Perhaps some philosopher believes one of the following variations on the simple claim.

2 Suppose that a person wants p to be true and believes that if she performs A then p will come true. Suppose that she then performs A. Then stating these facts about her will explain why she performed A.

3 Suppose that a person wants p to be true and believes that if she performs A then p will come true, and there is no q such that the person wants q more than she wants p and believes that if she performs A then q will not come true. Then the person will perform A.

4 Suppose that a person wants p to be true and believes that if she performs A then p will come true, and there is no q such that the person wants q more than she wants p and believes that if she performs A then q will not come true. Then if the person is rational she will perform A.

5 Suppose that a person wants p to be true and believes that if she performs A then p will come true, and there is no q such that the person wants q more than she wants p and believes that if she performs A then q will not come true. Suppose moreover that she then performs A. Then stating these facts about her will explain why she performed A.

These are all obviously false. If variations 4 and 5 are false then so are 2 and 3, so let me just discuss 4 and 5.

Variation 4 entails that if I want a million dollars and believe that if I poison my grandmother I will have a million dollars, and don't want anything more than I want a million dollars which is incompatible with my poisoning granny, and I am rational, then I will poison granny. Now that consequence seems crazy to me, since it seems to suggest that if you want something enough then you are rational to go for it, however awful it is. But suppose that there is some way of understanding 'wants more' and 'rational' which makes it less than insane. Consider then the case in which I want money more than anything else in the world, the more of it the better, and someone offers me a choice between a gamble – heads I get two million dollars, tails I get nothing – and a gift of $5,000. Variation 4 entails that I will go for the $5,000, since I believe that if I take it I will get that sum and although I want sums greater than $5,000 more than I want $5,000, I do not believe that if I take the gamble I will get any such sum. But surely it could be rational to go for the gamble rather than the gift. And indeed we can make theories of rationality that back up this intuition: but they are not going to make variation 4 true.

The same examples show that variation 5 is false. If I poison my grandmother for a million dollars you may want a much better explanation of why I did it than the bare fact that I wanted a million dollars very very much. Or you may not, depending on what you know of my character. If I go for the gift of the $5,000 rather than the gamble, citing my desire for money will not make you understand why I acted with such extreme caution. The caution may be explicable, but it will take more than my desires to explain it.

It is clear that variations 6, 7 and 8 below have most of the same problems as variations 4, 5 and 6.

6, 7, 8 Suppose that a person wants p to be true and believes that if she performs A then p will come true, and believes that A is the best means to achieving p. Then she will (will if rational/it can be explained if she) perform A.

A variation on these may look more promising, though:

9 Suppose that a person believes that an action A is the best means to satisfying her desires. Then if she is rational she will perform A.

10 Suppose that a person believes that an action A is the best means to satisfying her desires. Suppose moreover that she goes on to perform A. Then citing this belief will explain why she performed A.

Variation 9 avoids some of the problems of 5. In particular it handles the example of the profitable but mildly risky gamble. It still seems to me insane in that it entails that if I am rational I shall not scratch my nose or listen to music if doing that is not part of the best way to satisfy the totality of my desires. For variation 9 entails that if I am rational I must be doing whatever will best satisfy my desires, according to my beliefs, rather than scratching my nose. So it seems to share with utilitarianism the puritanical conviction that we ought to be doing what is absolutely best at every waking moment. But perhaps there is a conception of 'best means' that avoids this consequence. (Such a conception might be extracted from Chapter 2 of Slote 1989. Slote's position has a lot in common with the later sections of this chapter.) The claim is still obviously false. For suppose that a person has the insane delusion that the best way to satisfy her desires is to climb the Eiffel tower naked. Then variation 9 entails that if she is rational she will climb the Eiffel tower naked. But if she is rational what she will actually do instead is to reconsider her belief about what is the best means. This problem about rationality with variation 9 generates a problem about explanation for variation 10: if the person does perform A then ascribing to her a lunatic belief about the best way of satisfying her desires may not cast any light on why she acted if this belief itself is inexplicable.

Variations 9 and 10 become a lot more plausible if we replace belief with knowledge, getting:

11 Suppose that a person knows that an action A is the best means to satisfying her desires. Then if she is rational she will perform A.

12 Suppose that a person knows that an action A is the best means to satisfying her desires. Suppose moreover that she goes on to perform A. Then citing this knowledge will explain why she performed A.

Variations 11 and 12 are not beyond controversy, for well-known reasons. If a person's desires are sufficiently weird then there can be serious doubts about whether even the best means to them leads to a rational action. The destruction of the world and the scratching of Hume's little finger. In fact it is possible to doubt that there is a coherent conception of means–ends rationality that can function in real life with arbitrarily perverse desires. (This is a greater potential problem for variation 11 than for 12, since even if a person's desires make her rationality doubtful, the relation between her action and her desires may make some sense of why she acted as she did.) And then there is the undefined 'best means', certainly not a very clear idea. But, still, variations 11 and 12 are the most defensible of the lot so far.

Suppose that variations 11 and 12 can be made true, perhaps by some creative but permissible redefining of terms. How widely can we apply them in everyday psychological thinking? Not very widely at all. Their assumptions are satisfied in very few of the situations in which we predict, explain, or judge the rationality of people's actions. The basic reason for this is that people very rarely know what the best means to satisfying all of their desires is. Quite possibly they never know the best means.

Suppose that I want a beer and believe that if I walk to the pub and order a pint my desire will be satisfied. Suppose also that I have done a full day's tiring work, there are likely to be friends in the pub with whom I can talk, and with some of whom I have good reasons to want to talk, and so on. It makes sense for me to walk to the pub, and I do so.

Do I know that walking to the pub is the best way of satisfying my desires? In order to know that I would have to rule out the possibility that by staying an extra ten minutes at work I may solve one of the philosophical or administrative problems that has been bothering me. And I would have to be sure that there is not some other way of getting to the pub, which would result not only in a beer but in philosophical insight and a great monetary profit. I cannot think of any such way, any more than I can think of a reason why if I stay at work any longer I may make some breakthrough rather than just getting tired and frustrated. But I know people who are in various ways more intelligent than me, and I know that I often find their actions very puzzling until they are explained to me. So I am not in a position to say that I know that there is not a much better way of satisfying my desires. But I do want that beer, so I know what to do.

Let me put the point differently, as it may seem that the argument just shows that we rarely know whether we know the best way to satisfy our desires. There nearly always *is* a better way to satisfy our desires, which we are not aware of. A person can usually be sure that she doesn't know the best way of satisfying her desires. If instead of having this beer now I could spend five minutes on the phone and make some foreign-exchange deal which earned me a million dollars, then it would be well worth postponing the drink. And if I had an IQ of 300 two minutes with the past week's newspapers would put me in a position to forecast the mistakes in reasoning of merely normally intelligent foreign exchange traders and profitably out-guess them. Moreover, this is just one of thousands of actions that would satisfy my desires better than walking to the pub for that beer now, if only I was capable of seeing them. Sitting in my office at the end of a long day I cannot think of any of these better actions, but I know they exist, and therefore I know that what I will do is not the best way to satisfy my desires. (The gods look down on us and see how we try to satisfy our desires; they might feel inclined to pity or even to offer help, if they were not laughing so hard.)

None of this denies that we can describe better and worse ways to reason and to make decisions. But it does point to a dilemma. Principles that say what ideally rational agents would do, will not give much of an explanatory hold on what we actually do do. Even principles that say what ideally rational agents who had somehow managed to have our mixed-up beliefs and desires would do, or principles that say how ideally rational agents would advise confused and limited finite agents to proceed, will not help much. On the other hand people do often act in accordance with principles of rationality: sometimes we deliberately follow some formula which gives a rational procedure and more often we simply reason unreflectively in a way that is in fact not completely unrelated to some optimal procedure. So we need to include patterns of rational deliberation among the ways in which we expect people to act. In fact, we need to give them a fairly important place. Rationality seems both vital and irrelevant to what we actually do.

Consider now some plausible everyday explanations. I want a beer and I would like to chat with my friends, who will probably be in the pub around the corner. I have finished as much work as I can do today and I can think of nothing else that I need to do. I tell you all this. A few minutes later you see me in the pub and you are not surprised. Someone asks you why I am here and you say 'he wanted a beer and a chat after a long day.'

There are two offices next to mine. In one of them there is an idiot and in the other a genius. After talking to me you talk to the idiot. (Not in all ways an idiot, simply practically challenged, let us say.) He explains that he would like a beer and needs to talk to some people about how to get the light bulb in his office changed. And he has had a really profitable but tiring day typing an encyclopedia article into his word-processor so he can read it later at home. The nearest beer is in the department refrigerator, and the nearest people to talk to will be at the bus stop, so he feels torn, unsure what to do. Later, in the pub, you see him and you are puzzled. How did it occur to him to come here?

After talking to the idiot you had a word with the genius. Again there was a desire for a beer and for conversation. She explains to you other things that are on her mind too. There is this matter of the value of the ruble, about which a few phone calls might pay off handsomely. And then there is something you don't really understand about a bus route home which starts off in the opposite direction but actually would get her there more quickly, including a twenty minute stop near a pub where the television producers hang out. Later, in the pub around the corner, you see her. Again you are puzzled. You wish you understood why she has come here.

In all three cases your attitudes are shaped by your knowledge of what might be a profitable, rational, or sensible way for someone to get what they want. In different ways, though. In the first case your thinking, to a first approximation, is 'he has these desires, and if I had these desires then, other things being equal, here is what I would take to be a sensible way to satisfy them'. In the second case, the idiot, your thinking is, just as roughly, 'here is the reasoning I would go through, but I'm not at all sure he will get through more than the first couple of steps of it.' And in the third case, the genius, you think something like 'here is the reasoning I would go through, but I'm sure that she will think of many things that I would not, so her conclusion may well be very different'.

In the first case your reasoning might be even shorter than this. You may just think 'here is what I would do'. And in the third case there is a longer and more informative line of reasoning open to you: you may work out slowly the kinds of considerations that could have affected her choice of drinking place, taking more care than you would with your own decisions. This will obviously work best if she guides you, leading you through the pros and cons and derivations. The result will be a slow-motion assimilation of the third case to the first. It is not at all clear what the analogous procedure for the second case is.

To sum this up: when we explain people's actions by appealing to what would be the best way of achieving their desires we typically lay out a variety of actions that they could have performed, together with considerations about ways in which these actions would have led to the satisfaction of the person's desires. Then we choose the pair of action and justifying considerations that best fits what we know of the person. The examples of 'what we know of the person' in this section have concerned intelligence, but as we will see that is far from the only relevant factor.

(Who has asserted what this section denies? It is not easy to find absolutely pure examples, where the claims concern just explanation and prediction, rather than also the causation of actions. But the impure examples are legion. After Aristotle, Hobbes, Hume and Kant – 'who wills the end wills the means' – good examples are essay 2 of Davidson 1980 and Smith 1994. For a survey of related positions see Lennon 1990. There are more subtle connections with Baier 1985, who argues for a 'non calculating' understanding of practical reason. Also with the fascinating final pages of essay 14 of Davidson 1980, which can be read as arguing that we never learn anything about how people connect motives and beliefs to actions: we learn about different people's beliefs and wants, and we fit them into an unvarying apriori framework. If only it was obvious that he was really claiming this, it would be exactly what I am arguing against.)

Virtues

There is no general formula connecting belief, desire, action and no other factors, which both fits a large proportion of motivated actions and which we can use in everyday explanation and prediction. There may be general relations tying belief and desire alone to action, but they are not likely to be expressible in common sense terms. (If any exist, they are likely to be expressed in terms of precise degrees of confidence and preference rather than in terms of beliefs and desires. And they are likely to drive a large wedge between what it is rational for someone to do and what they can actually be expected to do.) The variety of ways in which people act on their motives is just too great. When we try to squeeze the variety into one container we find ourselves relying on ambiguous flexible notions like that of the best means to an end. But the idea of a best means is treacherous for two important reasons. The first is that there are many factors that make something a good means, and there are no obvious ways of comparing most of them. (Some moral philosophy may propose ways of comparing them; some way might actually be right; but no way is built into common-sense explanatory practice.) The second is that there is very often no best means to an end, at any rate no best means available to mere human reason.

I will return to the second reason later (in ‘Heal’s problem’ below). The first reason, the variety of ways agents can connect means to ends, suggests that other factors besides beliefs and desires have an explanatory role. The point is clearest in the case of risk. Suppose two people are given a choice between a sure outcome and a gamble whose expected value is better than the sure outcome but whose worst-case possible outcome is much worse. One takes the gamble and the other does not. The difference may be explained in terms of their different attitude to risk: one is more adventurous and the other is more cautious. It may be explained in these terms if indeed it is reasonable to say that the one person is adventurous and the other cautious. The relation between the intelligibility of the action and its rationality is rather delicate. We certainly have some use for explanations in terms of lunatic risk-tolerance and pathological risk-aversion. For the moment let us simply fasten on the simple idea that often when we explain why a person has chosen an action we make essential use of some concept describing the way that particular person works.

There are many such terms. Many of them involve some dimension along which human problem-solving capacities vary. Some people are wonderful at finding their way in strange environments while others get lost a block from home. Some people can find their way through tricky social situations that completely baffle others. And some people are more intelligent than others, where this means just that they are better at organizing a mass of data and seeing relevant patterns in it. Although there are fools and geniuses and a rough gradation in-between, it seems a pretty clear fact about our species that there is very little correlation between many of these abilities. (The concepts of a fool and of a genius are not simple opposites. A fool bungles everything, while a genius has a flair for some special area. The concept of intelligence, as measured by IQ, is a recent and uncomfortable addition to our folk psychology. If rationality is taken as an all-purpose capacity for finding good answers, rather like intelligence, it may reflect a deep misunderstanding of the ways we think, a poisoned gift of philosophy.) So it should be no surprise when the person who can diagnose immediately the computer problem that has beaten you for months is defeated by a social dilemma whose solution is immediately clear to you.

The list of such reason-describing factors is potentially infinite. Any society’s vocabulary will include only a few of those that might be invented, and there are very few if any that all societies’ vocabularies must have. So the folk psychological resources available at one time and place may be more suited for explaining some kinds of motivation than those available at others. Many of the factors that we can describe are virtues, in that they describe capacities that it is good to have and that one person can have more of than another. One influential definition of a virtue is due to Linda Zagzebski:

> A virtue, then, can be defined as a deep and enduring acquired excellence of a person, involving a characteristic motivation to produce a desired end and reliable success in bringing about that end.
>
> Zagzebski 1996, p. 137

This definition will need some modification for present purposes. A practical virtue such as navigational skill or social facility need not involve a specific motivation. That is, the end in question can be very general, such as getting where one wants to go, or finding desired outcomes to social situations. And the 'acquired' also should be treated with care: any such virtue will come much more easily to some people than to others. So I shall take a virtue to be a relatively basic aspect of a person's psychology, which enables the person to be reliably successful in bringing about some class of ends.[1] On this definition it is easy to see why virtues are relevant to explanation. A virtue specifies that a specific motivational pattern operates in people who have it, and, moreover, is such that its possession is causally relevant to the achievement of the desired ends.

We can also appeal to characteristics that are not excellences when explaining why a person acted as they did. And we can appeal to vices rather than virtues (see this chapter 'Shared problem spaces' and 'Virtues?' below.) For now, let us try to understand better the explanatory role of virtues. First consider the variety of virtues that can be relevant.

A person is looking for his car keys. He has checked that they are not in any of his pockets or lying on any surface in his house. His wife sometimes uses his car keys so they might be in her bag, and they might also be somewhere that he has already looked, but not searched carefully enough. Someone who knows him well is asked what he will do, and predicts, correctly, that he will first check more carefully in all his pockets, and then look in bizarre places where one might absentmindedly put things, like the refrigerator and the box of beer bottles for recycling. When asked how she knew what he would do she says 'He's unusually truthful with himself. He knows he's no fool but he's quite prepared to believe that he has done something dumb, and will check out that possibility before he accuses someone else.' The virtues here are honesty, self-knowledge and modesty. (In a real case the explainer is likely to give more background, to convey the exact forms that these virtues take in this person. Some important aspects of the background are likely to be conveyed more by appeal to the imagination than by explicit description. See the discussion of the inadequacies of the vocabulary of the virtues, taken literally, Zagzebski 1996, p. 135.) Without understanding that this person has these virtues one could not understand why he looked in the fridge before checking his wife's handbag.

Honesty and modesty are moral virtues. To say this is not to invoke a large and definite distinction between moral and other virtues. I do not think there is any very definite distinction. Any feature that it is good that people have –

intelligence, prudence, care, curiosity – can be a source of moral admiration, and can play a crucial role in a person's handling of a moral problem. Or, to put it the other way round, the absence of any virtue can in suitable circumstances make a person behave badly and thus be a moral fault. (And the most obviously moral virtues, for example generosity, can be part of the motivation for actions that are morally wrong. See Trianoski 1987, and Zagzebski 1996, p. 93) It is not important for our purposes how we draw the line between moral and other virtues, as long as we are clear that it is a pretty arbitrary line. (We could take a moral virtue to be any virtue whose general possession we value, or as one whose general possession would have good consequences. These are not the same.)

Obviously any virtue can play an explanatory or predictive role. For some virtues, the actions they are most obviously relevant to involve interactions with others. A person's kindness or fairness will explain why she forgoes an opportunity to gain a small advantage at the invisible expense of another. Such applications of moral virtues are real and essential to everyday explanation. But they are not ideal for furthering my argument at this point. For it is always possible for someone to understand the appeal to fairness in such a context as a shorthand for attributing a complex of desires, such as a preference for equal access to goods and a distaste for procedures that are fixed for the benefit of specific individuals. And while I think that this reaction stretches the concept of desire for ideological ends, it is not a very profitable issue to debate. Better to find examples in which kindness or justice help explain actions which do not focus on others.

Imagine a philosopher. He is considering a claim that he imagines many of his contemporaries might defend. An argument occurs to him, and following it through he sees that a radically unacceptable consequence follows if the claim is conjoined with an intuitively acceptable premise. Or at any rate it seems to follow. He can imagine beginnings of objections to the derivation and he can imagine that some might find the premise not so intuitively acceptable. He does not, however, give these doubts more than a moment's attention, takes the derivation at face value and begins to refer to the claim as 'that ludicrous contemporary dogma'. Some of his fellow philosophers are puzzled by this hasty and inflexible line. Those who know him as a friend or colleague, though, are not. They put it down to a lack of a sense of fairness, a kind of bullying injustice, which prevents him giving equal respect to all the directions in which an intelligent person might take an argumentative situation.

What makes this example work is that much thinking is internal debate, and debate is a social activity so that social virtues apply to it. So even when no actual other people are involved, a person's thinking can be affected by whether she is fair, unfair, tolerant, arrogant, or bullying. The fact is that our thinking is a closely woven fabric. Any deep characteristic of it that shows in one area is likely to show in another. So practical and intellectual virtues are always morally relevant, and non-superficial moral virtues will show up throughout the thinking of those who have them.

Heal's problem

A worry may have occurred to many readers. Although I have not claimed that only virtues can play the explanatory roles I have been describing, it has clearly been important to me that the characteristics I built examples around in the previous section were virtues. But does their virtue-ness come into the story? Suppose that we simply speak of cognitive dispositions, some of which may be virtues, moral or otherwise. Would anything be lost?

In one way nothing would be lost. There certainly are many dispositions that mediate practical reasoning which have no virtue/vice dimension. (And sometimes when you know someone really well you get an ineffable sense of how they operate that corresponds to some such disposition.) But these are not the dispositions that we normally appeal to in everyday explanations. At least one large class of characteristics that we are geared up to think regularly in terms of are ones that we *rehearse.* They are associated with categories of problems, problem-spaces, that we learn a way around, and then use this knowledge to guide our understanding of others' thinking. These are shared problem-spaces; they have to be in order for the rehearsal to give understanding. And so we often focus on characteristics that are associated with shared problems.

That is the argument of the next two sections, in outline. It is quite delicate: I have to demonstrate a significant constraint on our understanding of reasoning, while not overstating it into a universal and exceptionless business. In this section I shall explain about problem-spaces. I shall approach them via the issues that Jane Heal discusses in an important series of papers. (See Heal 1998 and especially Heal 2000. Heal is arguing for a very particular kind of simulation of one person's thinking by another. For other accounts of simulation see Chapter 5 and the references in it.)

Heal's aim is to characterize a family of process which she calls 'cocognition'. One person cocognizes another when she gets a grasp on what that other person may do, or has done, by setting herself the same problem that the other faces, and observing what her own solution to it is. The cases where cocognition is most obviously a powerful and natural device are those where the other person faces a non-trivial but manageable task with a single best answer, such as multiplying 12 by 16 or finding the shortest route from A to B including C and D. If one is asked to predict what the other person will do faced with such a situation – either what she will say if asked or what action linked to the answer she will take – the almost inevitable response is to tackle the problem oneself and to conclude that the person will arrive at the same answer that one did.

Although the range of cases in which Heal's cocognition is obviously right is fairly narrow, there are many other cases in which it is intuitively very plausible that something rather similar is at work. For example: you look out your window one night and you see a woman trying to defend herself against an attacker. You turn on the light outside your door, wait the amount of time it would take to run across the road, and open the door briefly, just long enough for her to slip inside. You expect her to be at the door at that time because running straight to the door is the clear solution to her problem (once the light indicates your presence), which you would arrive at were you in her situation. This particular example makes connections both with the previous chapter (acting on an assumption about how someone else will act) and with Chapter 5 on simulation. For the moment, though, the point is just that we often anticipate someone's action by seeing it as the solution to a problem, where we see that it is the solution by solving it ourselves.

It has become increasingly clear in successive essays that Heal's claims are epistemological rather than psychological. Cocognition is not meant to describe the underlying processes by which we arrive at expectations about others, but the reasons why those expectations are rational. In a recent essay she contrasts two general patterns of justification into which cocognition can fit. (Heal 2000. I am formulating some points slightly differently from her exposition.) On each there is an additional assumption. The first pattern adds the assumption that the person being modeled is like the modeler, so that the justification goes:

> I I arrived at solution S
> P is like me in relevant respects
> Therefore P will arrive at solution S.

The second pattern adds instead the assumption that the person being modeled is capable of solving the problem, so the justification goes:

> II I arrived at solution S
> Therefore S is the correct solution
> P is capable of arriving at the correct solution
> Therefore P will arrive at solution S.

Heal does not deny that the first pattern is often applicable. But she is concerned to defend the relevance and wide applicability of the second pattern. In so doing she runs into a problem, which makes the connection with the issues of this chapter.

The problem is that in many cases – probably most cases – in which something like cocognition is plausible there is not a single best solution, which just about any person who thinks about the problem will arrive at. This is clearest when we consider agents who are generally less or more intelligent than oneself, but as we saw in 'The negative bit' above there are many areas in which one's own abilities are not a good guide to the choices of others. There are three ways in which the argument pattern above can be extended to such cases. First, extending the pattern I

> I* I arrived at solution S
> P can be got from me by transformation T
> If I imaginatively apply T, I find I arrive at T(S)
> Therefore P will arrive at T(S).

This pattern is obviously inapplicable where the person's abilities are greater than one's own. Heal argues that it is intuitively very implausible that when the person's abilities are less than one's own there is much to be gained from applying I*. She considers the example of an expert chess player trying to predict the moves of a beginner. To transform oneself into the beginner one would have to close down one's sophistication, to force oneself not to see combinations and possibilities. But 'shutting down capacities is not something we can do at will or by engaging in some imaginative enterprise, since what it is to possess them is precisely to have certain thoughts strike one or come to one spontaneously on the presentation of certain problems.' (p. 17)

Instead, Heal suggests, we should modify II. We should 'use [our] capacity . . . as presented to generate a rich appreciation of the options available and then to cull from this range those which I judge to be too advanced, leaving only some limited range of obviously attractive moves as ones the inexperienced player might choose.' (p. 17) Putting this in the form of a justification-giving argument, we might try

> II* I arrived at solution S
> Therefore S is the correct solution
> S breaks down into parts S1 + S2 + S3
> P is capable of arriving at part of the correct solution, in fact at S1
> Therefore P will arrive at sub-solution S1.

(In fact this is not quite true to the case Heal discusses, as the choice the novice makes may not even be a partial solution to the problem. I will return to this.)

Heal's aim is to show how an expectation about someone's actions can be justified by relating the person to the structure of the problem and its fuller and partial solutions. Since her concern is with cocognition the grasp of this structure that interests her is that of the person who does the anticipating. And indeed this is often the natural way to approach the situation. How else might one do it? Well, if the person had a *better* grasp of the structure than one's own (or an orthogonally different one) then one might want to assume that there was a structure, but not use one's own grasp of it as definitive. My own concern at the moment is not with cocognition but with the explanatory force of virtues. But there is a close connection: a virtue is a way of seeing and putting into effect solutions to some class of problems. So if Heal is right that when we explain we often relate people to the structure of the situation they face then it would seem to follow that we often explain by relating to the full or partial exercise of a virtue. For example in the chess novice case our grasp of what he will do (or why he did what he did) is in terms of his partial possession of the virtue of chess strategy. And the virtuousness of the virtue is essential to the explanation because it is essential that what he has a partial grasp of is the set of *good* chess strategies.

A distinction will help clarify the connection between the virtuousness of a virtue and its explanatory force. The problems agents face in the situations where Heal's cocognition applies most readily are rather special, as I remarked earlier, in that there is a best solution of them, accessible to human endeavor. Call such problems *contained*: there is a right answer, and enough thought will produce it, and average humans can do the thinking. But very many problems are not like this. Call them *uncontained*. Mere human intelligence, even applied with diligence and care, will not produce a best solution to the problem. A small shift in formulation can move a problem from contained to uncontained. For example, the problem of whether to do A or B or C or D is often contained, and the problem of whether to do A or not to do it is often contained, if we understand it as asking whether one should do A or the most likely alternative. (This point is made by Jackson and Pargetter 1986.) But the problem of what action to do next is almost always *un*contained. For any action that we can think of there is likely to be a better one, and if there is a very best one it is surely beyond human capacity to define it.

It is not essential to uncontainedness that there be no upper limit to the value of the solution. Chess provides an example. Assume that humans will never fully solve chess, in the sense of discovering the winning or stalemate-ensuring sequence of moves that we know must exist. Then although some moves in any given chess situation are better than others, in most situations there is no single unique right move. Sometimes a move is the best we have thought of yet, but this does not mean that a different move may not actually lead to a better result. (What is the best starting move? Pawn to K4 or Pawn to Q4, traditionally. But for all we know Pawn to KR 3 may be the beginning of an unbeatable sequence.)

Most contained problems derive their importance from their position in uncontained problems. The contained problem of knowing whether Q is a truth functional consequence of P is only of interest because it is part of two uncontained problems, that of knowing whether Q is a deductive consequence of P where a wider range of logical structure is involved (notably quantifiers) and that of knowing whether when one learns that Q follows from a belief P the wise reaction is to begin believing Q or to cease believing P. As Gilbert Harman has taught us, facility with the latter problem is a very basic and mysterious epistemic virtue. (See Chapters 1 and 2 of Harman 1999.)[2]

When a problem is contained, there may be only a single not very interesting virtue that is relevant to it: the virtue of getting the right answer. But when the problem is uncontained, many virtues can be brought to bear. There are virtues of general capacity for the topic in question (inference, spatial reasoning, social strategy). There are virtues of coping with the fact that a full solution is not forthcoming (patience, tolerance, modesty). And there are virtues directed at the different kinds of partial solutions there may be. (We don't seem to have easy labels for these virtues: those of assessing the partial-ness of a solution, the cotenability of different partial solutions, the balance between an easy, more limited solution and a more difficult, more complete one.)

Suppose then that we are interested in the actions of a person faced with an uncontained problem. We know what the problem is, though we don't have The Answer to it. (The problem could be that of reconciling love and work, or of dealing with difficult colleagues, or of writing an essay on free will.) And we know roughly what relevant virtues the person possesses. The person may be more or less able than us, with respect to this problem, or have abilities that are incommensurable with ours. (The last is the most common case.) Our first step is cocognitive, in a wide sense. We think about the problem, and get a sense of difficulties, opportunities, traps and imponderables. The second step is to fit the person to the problem. If we know their particular virtues we should know what opportunities they can seize, what traps they may fall into, and so on. Then we can tackle prediction. In some cases, actually quite rare, probably, we can see the partial solution that will present itself to the person. (This is essentially the case that Heal discusses.) In some others we can see the *kind* of solution that the person will come up with. We can do this even when the person's virtues are superior to, or different in kind to, our own. We can for example predict that the person will be able to reconcile two conflicting demands, somehow or other, or that they will be able to overcome a problem that seems to us hard but not insuperable. (We're watching a chess master. We see that his queen is threatened. In that situation we would withdraw it immediately, but he does no such thing, so we presume that he has some way of meeting the threat, or using his queen as bait for a trap.)

Of course no prediction may be possible. But once the person has acted we can often explain even when we could not predict. The process will often be essentially as just described: cocognitive scouting of the terrain followed by location of the particular person on it. To the extent that we grasp the structure of the problem we can see the action in relation to it. We can see that the master moved his rook to the left file because it gave the impression that he was trying to check with it, when actually it was a distraction to cover up the real threat from a knight. We can do this because we understand the problem well enough and understand the person's virtues well enough that given the action we can see the connection with the problem that we could not find in advance. This happens routinely with deductive consequence and conversational relevance. We rarely know what someone is going to say next, but when they speak we can often see that it follows from, or is in some other way clearly relevant to, some assumption that was in the air.

Shared problem spaces

The pieces are now in place for arguing for an instance of what at the beginning of the chapter I called the common focus claim. The conclusion will be that when we are predicting or explaining using the extended cocognitive methods that I have been discussing, our thoughts focus on characteristics of people that are related to the structures of shared problems. Here is the argument.

The problems people face are usually hard, and their solutions to them not easy for others to follow. When we follow another's thinking by reproducing similar thinking in ourselves we usually do it by being familiar with the family of problems of which it is a part, and being able to situate ourselves and the other person with respect to that family. We associate the problem with a problem-space and characterize the person in terms of their ability to negotiate it. We can't do this unless we too have some familiarity with the problem-space. We have to have rehearsed in it, normally both by solving problems of the relevant sort and by cocognizing others' solutions. This takes time and repetition, and the problems rehearsed need to be varied, covering a representative portion of the space. So in any particular case the problem the person is understood to be facing has to be thought of as an instance of a type of problem that the understander is familiar with and has rehearsed. The rehearsal will have happened under two circumstances. The first is when the problem is shared, when it is a problem that occurs in many-person projects and where one person's solution impinges on another's. In this case placing others in the problem-space is vital. The second is when it is a one-person problem that occurs frequently enough in many people's lives, including those of both the understander and the understood. In this second case the skill of placing the other in the space is not likely to have been rehearsed unless there was a need to know what solution others were likely to arrive at. So even in this case the problem will be connected with matters of shared concern. Therefore, in both cases, the problems that we can tackle cocognitively are ones which can be placed in a space of problems concerning matters of shared interest.

The characteristics of people appealed to in this process may not all be traditional virtues. But they are in an unmysterious way normative. That is, their functioning as indicating the links between motive and action relevant to particular agents depends on the existence of a structure of better and worse solutions to a family of problems. (And of course more than that: safe and risky strategies, dead ends and promising leads, and much more. The structure of most hard problems is rich.) The solutions are better and worse in part because we all engage with the problems in the context of particular motivated thinking, in which we have particular aims, which better solutions achieve better. The structure is partly created and partly discovered.[3]

We can see now why, in understanding a philosopher's treatment of rival views, or a chess player's choice of one strategy rather than another, we will usually make some connection with characteristics of the people and of the problems they face that are also found in cooperative or rivalrous activity. We appeal to fairness, courage, impulsiveness and the like. The reason is that by doing so we make ways of bringing our experience to areas where we have rehearsed problem solving, cocognition and the combination of the two, to domains where another person's route through the maze of possible lines of thought would otherwise be beyond our grasp.

A conjecture

I have discussed the appeal to virtues and other characteristics in everyday explanation as if they simply filled the gaps in fixed patterns connecting belief, desire and action. I think the appeal goes much deeper than this. I think that the patterns of practical and indeed theoretical reasoning that we can comfortably apply to one another are shaped by our grasp of the characteristics, particularly social characteristics, that we think people do and ought to have. Making a good argument for this is another matter, however. One argument can be got from Exploration I, where I shall argue that some tendencies to 'erroneous' statistical reasoning have cooperation-inducing effects. If this is right then the theoretically flawed patterns that we typically attribute to one another owe their deviations from correct statistical form to the pressure to anticipate others in socially profitable ways. I shall not follow up this idea, however.

A second way of arguing for the thought is by making a conjecture about reasoning. Begin with what I'll call the Harman–Cherniak problem (Cherniak 1986 and Harman 1999). Suppose that a person is thinking about a problem – what to do or what to believe – and starts with a set of beliefs and desires in mind. Then she infers more beliefs and desires, and possibly intentions and conjectures. There are indefinitely many paths of reasoning she can follow. Some are profitable, many will not be. Some paths of reasoning although perfectly sound in terms of logic are irrational in the sense that they are not good strategies for arriving at a solution to her problem. So there is a deep epistemological problem here, the problem that concerns Harman and Cherniak, of separating the reasonable from the unreasonable among logically correct patterns of reasoning. But there is also a psychological problem. When one person imagines or anticipates the reasoning of another, or attributes reasoning to explain what the other has done, what paths will she imagine or anticipate the other following? Presumably the two questions are closely connected in that there is a large overlap between the two classes of paths (those that it is reasonable to follow, and those that another will expect one to follow). Pure logic – pure statistics – does not determine what these paths are. What does? The obvious suggestion is that one determinant is the person's construal of the kind of situation she is in. That is, one strategy that we use to guide us through the maze of possible consequences of our beliefs and desires runs as follows:

> *stage 1:* characterize the situation as an instance of a familiar problem type, one that one knows one's way around
>
> *stage 2:* employ an easily calculated function which often transforms problems of this type to simpler sub-problems
>
> *stage 3:* check if the result of stage 2 is of a type that one has found to be soluble. If so, apply a solution that has worked in the past. If not, repeat 2 with a less easily calculated function.

Suppose that something of this general sort were the case. Then people in their practical thinking would follow lines of reasoning that they had rehearsed. The rehearsal would be both 'on line' in dealing with similar previously encountered problems, and 'off line' in cocognition of others. And rehearsal would focus, for the reasons given in the previous section, on topics that people shared with those others with whom they habitually interact. So shared activity would shape individual thinking.

Virtues and other characteristics of an individual's thinking would on this model show themselves in the problem-types to which a person assimilates a specific problem and in the strategies for transforming a problem into a possibly solvable one. (Of course this will not do justice to the full range of individual characteristics. It's a very simple model.) Pure logicality, in the sense of a tendency to derive only what is logically or statistically justified given the description of the problem, would be a very rare characteristic, certainly not always a virtue. In articulating an explanation of another person's actions words referring to such characteristics would cue one's audience into applying the appropriate larger problem class (for which neither the agent nor the explainer might have an explicit description) and the right simplifying transformation.

It is a good conjecture that something along these very general lines is right, that everyday practical thinking proceeds more by rehearsal of assimilation and simplification routines that we have rehearsed in shared activity than it does by formally correct inference. But we are not going to discover the real truth here by examining patterns of everyday conversational attribution, or by clever thought experiments. It is enough to have made it clear that we need evidence before concluding that the most basic patterns of reasoning are not shaped by a person's experience of shared activity. This one will have to wait on some psychology.

Virtues?

Given that we make essential use of characteristics of individuals, in explaining their actions, which focus on the structure of shared problems, we can ask 'must they be virtues?' And the simple answer is clearly No, for two reasons. First, there are vices and incompetences. We clearly do qualify our explanations with judgments about what goes wrong with a person's thinking. Faced with a subtle and malicious foe I may have no idea what he will do next, but be sure that whatever it is it will produce maximum consternation to my projects. Discussing a hopeless colleague we may say 'you can't be quite sure how he will mess it up, but somehow he will, so we had better be prepared.' It is interesting that some such vices arouse a kind of despairing admiration, rather like virtues. This is typically when one learns something in cocognizing them, that is, when the effort to understand actions shaped by them is valuable.

Vices are no objection to the common focus claim, since they are defined in terms of the structure of shared problems, and it is those structures that are the object of the claim. Another reason for admitting that many of the relevant characteristics are not virtues is more technical. In moral psychology 'virtue' has come to have a fairly specialized meaning. It applies not just to any good characteristic of a person but to dispositions which need to be learned but are not easily learned by internalizing formulas, and are generally admirable or beneficial when activated in specific circumstances which call for them. Thus intelligence is not a virtue, as it is not learned. Being good at chess is not a virtue, since it could be best not to use the skill, for example when playing against a novice. But tact is a virtue, because in the hard-to-define set of circumstances in which it is called for a tactful person must use it.

The characteristics I have been describing are not all virtues in this sense. They are a large and varied class of dispositions, which we grasp in terms of the structure of problems, many examples of which are virtues. That is all that can be said. There is, though, a reason why virtues in a rather looser sense than that employed in moral psychology do have a special position among these explanation-mediating characteristics. It is an instance of what at the beginning of this chapter I called the good effects claim. We can expect that many of the characteristics that we have names for and which we evoke in everyday explanation are ones that we admire, or want to be widely exemplified. The reason is, again, rehearsal. Fluent explanation requires rehearsal of the concepts involved, and rehearsal brings skill. So we are likely to rehearse more often the skills that we want ourselves and want others to acquire.

Suppose for example that you observe a friend being unusually patient with a silly dithering acquaintance. At first you are irritated with the friend, for prolonging the dithering, and then you come to understand that the patience is patience, rather than just indifference or detachment, and realize that it is appropriate for the particular occasion. Then you have an appreciative attitude to your friend. You see how her motivation works, how she solves the problem and what aims she sets herself in doing it. And in so doing you make this pattern of motivation (and of management of motivation) one that you yourself are prepared to assume when appropriate. In some cases the appreciative attitude is more distanced. If you understand why an act of extreme heroism was appropriate you may well not be capable of wanting, for example, to withstand torture in order not to reveal a secret to the enemy. And understanding that it was appropriate may not allow you to have that motivational structure as part of your own repertoire. But you *wish* that it could be part of your repertoire. To understand that something is a virtue is to appreciate that the virtue represents a capacity that can serve some purpose one approves of.[4]

The link with explanatory force is immediate. If understanding the appropriateness of a virtue makes its motivational structure available, then it is inherently suited for cocognition. When you see that your friend's patience made sense, then you begin to be able to get into the ways of patience, and so can predict, by employing the as-if motivation you have acquired, what other acts may result. And given a specific unpredictable action one's cocognition is more likely to come up with more of the motives and motivation that lay behind it. One reason, then, for trying to understand the actions of others is not to be able to predict what they may do but to help in acquiring their skills and competences. Understanding can bring as well as make use of virtue.

Chapter 3

Belief and coordination

> Well then, we have seen that there are true and false statements, and among mental processes we have found thinking to be a dialogue of the mind with itself and judgment to be the conclusion of thinking. And we have seen that 'appearance' means a blend of perception and judgment. So it follows that judgments and appearances must like statements sometimes be false . . . You see then that we have discovered the nature of false judgment and false statement sooner than we expected.
>
> Plato, Sophist 264 b

> 'Let me go over a bit of it again.' Appleby held up an index-finger. 'He knows we are police.' He held up a second finger. 'But he doesn't know we know he knows.' He held up a third finger. 'He thinks he has hoodwinked us into believing that he believes that we act for Rathbone: he thinks we believe that Rathbone exists: he thinks we believe in the fundamentally scientific character of the whole affair. That is how the position stands now.' . . . 'I suppose you wouldn't be disposed to call it at all complicated?' . . . 'Only when reduced to these compressed verbal terms. The actual situation is fairly simple.'
>
> Michael Innes, *The Daffodil Affair*

Fragile belief

When we ascribe beliefs, by saying things like 'she thinks that he is good-looking', there is a very delicate relation between the actual words we utter and the content of our ascriptions. What is the case with the people the belief is ascribed to, if the ascriptions say something true of them, depends in ways that are not easy to work out on a number of factors that we do not specify explicitly in our words. That much is familiar, though more detail is given below. In this chapter I show that among these factors are coordinations between belief-ascribers and belief-holders, determined by the needs of ascribers and believers to be intelligible to one another. In effect, we fine-tune what we mean by 'believe', from situation to situation in order that the ascriptions we make fit our shared projects. Usually this means that interacting agents should mean the same by 'believe' (and 'thinks' 'judges' and so on). In fact, I shall describe two deep-rooted coordinations. There is a coordination of force, in which we agree to mean congruent things by ' believes'. And there is a coordination of content, in which we agree to take our beliefs as referring to things, real and notional, in ways that will facilitate our shared purposes.

It has seemed to many philosophers and psychologists that the absolute core of the concept of mind lies in the concepts of belief and desire. Or, more precisely, in the idea that states of mind represent situations in the world, either as facts or as aims. Given this, we can attach many optional extras in the form of emotions, styles of reasoning, additional ways of representing, or whatever. But until humanity, or a developing child, acquires the central insight that minds represent facts, it is not on the path that leads to the adult conceptions of mind in developed cultures. The idea is well expressed by Josef Perner:

> children fail to understand belief because they have difficulty understanding that something represents; that is they cannot *represent* that something is a *representation*.
>
> Perner 1991, p. 186

The suggestion is that older children and adults can represent representation, and therefore can have the concept of belief, and are thus initiated into the main body of the concept of mind. The thought that the concept of belief is at the heart of everyday psychological thinking is shared by writers, from Skinner to Churchland, who have a considerably less respectful attitude to belief, and to the propositional attitudes in general. The intensionality of 'believes' is taken as showing how hard it is to reconcile commonsense thinking with a scientific attitude to mind. (See Chapter 1 of Quine 1960, influenced by views of Skinner, whose aim of a behavioristic reformulation of common sense culminated later in Skinner 1971. The story continues with Churchland 1978 – see also Churchland 1998. A more nuanced attitude is defended by Stephen Stich – 1983, Stich and Ravenscroft 1996, Chapters 1, 2. Nothing like belief will figure in science, but some novel propositional attitudes might.)

The assumption common to all these writers is that an ascription of belief makes a definite assertion about a person, describing an imagined fact that might or might not be among the real causes of the person's behavior. This assumption is undermined, I shall argue, by the indirectness and subtlety of belief ascription. There is no simple connection between what is said and any causes or properties that common-sense might be invoking. The simple picture implicit in the quotation from Perner above, in which there are straightforward facts about what a person's mind represents, which one can discover and as a result understand something basic about other people, cannot be right. There are many facts and objects bearing many representation-relations to any person at any time: which ones we take to be objects of that person's beliefs depends on the reasons for which we want to understand the person.

My strategy is to spend the next three sections trying to loosen the grip on the reader's intuitions of the idea that 'a believes that p' makes a clear and definite statement about the person a. Then come two sections ('The coordination of content' and 'Propositions and inferences') which present the core argument of the chapter, which links belief to coordinated intelligibility. Then in the rest of the chapter I make some suggestions why a concept that behaves in the way I have been describing should be a useful device for describing and explaining what goes on in our minds.

Belief as bungled knowledge

A claim that some everyday concept is more fragile than we might have thought is more interesting if it is made clear that other similar concepts do not share its specific fragilities. After all, all commonsense concepts have limits to their easy application; for all there are circumstances in which it is unclear whether or not they apply, which can often be forced in one direction or another by practical considerations. So I shall list several folk-psychological concepts that, wobbly in many ways though they may be, do not seem to me to share the attribution-relativity that I am ascribing to belief.

First, there is the content of a person's stream of consciousness. Images, sounds and words run through people's heads, and, taken just as they are and not as representing anything beyond themselves, it is a reasonably definite fact what they are and when they occur. If the words 'he's lying' suddenly and vividly form in your mind as you listen to a politician then they just do: but it does not follow that you think that anyone is lying.

Second, there is the behavior that a person is disposed to produce at a given moment, or would make if specific circumstances occurred. If you are on the verge of screaming, then that is what you are disposed to do. If you will scream if your colleague talks for more than five seconds longer, then that is what that possible future will produce. But it does not follow that if you scream you will be expressing despair or even commenting on your colleague's eloquence.

And third and most important there is basic knowledge of one's immediate environment. In many cases this is a very clear and definite business, even though there are obviously unclear cases and enormous philosophical difficulties in relating it to other concepts. (But these difficulties are real in part because our reactions to cases are so definite. We are often clear whether we know, and the Yes's and No's form a very puzzling pattern.) People know the colors and shapes of the objects about them. They even know a lot about one another's moods and intentions.

Knowledge is subject to uncertainties that do not affect the content of consciousness or dispositions to behavior. But it is less susceptible than belief to the factors I shall discuss below. I shall largely make my case below with instances of belief-attribution in which knowledge-attribution is not at issue.

I do in fact believe that the situation is very subtle, and that the factors I discuss do in fact affect a refined knowledge concept that we sometimes use. (I think there is a core concept of knowledge that is roughly immune to them, and that in terms of it we can understand a core belief concept, as explained just below. We can then develop delicate attribution-relative further concepts of belief, in terms of which we can understand equally delicate non-core knowledge concepts. But in this chapter I shall ignore these subtleties.)

There are obvious ways in which starting from this basis one can approximate to belief. In Timothy Williamson's phrase (Williamson 1995, 2000 Chapter 1), belief is bungled knowledge; it is what you get when you start with belief and then something goes wrong. Suppose for example that someone knows of each of 13 cats that they eat mice, and knows no other cats. (Just grant this, without specifying what makes these states count as knowledge.) Suppose then that she reasons, inductively, to the conclusion that all cats eat mice. This conclusion is not knowledge, since it is not true. But it is clear why we can attribute it to the person: starting from what she knew she inferred it. It bears the same relation to her knowledge of particular cats that her knowledge that the sun rises every morning does to her knowledge of particular sunrises. It is a bungled attempt at knowing of all cats that they eat mice.

There are other ways of bungling attempts to know. Instead of correct reasoning with an uncooperative world the reasoning itself can be flawed. Yet if it fits into patterns of reasoning whose other nodes are items of knowledge then it too can be seen as an as-if knowing that something is the case. This is quite a limited business, though: what is inferred from non-knowledge is usually not knowledge, even when it happens to be true. When the input links, from knowledge, are stretched too far, output links, to behavior, can take up some of the slack. The most obvious such link is to assertion. A person is shown an object, thinks for a while, and then says 'it's a corkscrew'. Suppose that it is not, and suppose that in fact her reasoning is flawed. Still, her act is the one she would have performed if she had known it was a corkscrew.

Suppose we did not have a real concept of belief, but instead explained actions in terms of knowledge (of a straightforward close-to-the-facts kind) and in terms of almost-knowledge (near-enough-to-the-facts).[1] We would be well equipped for explaining actions directed at particular physical objects. For example, if a predator is chasing a prey, we can explain its catching the prey by saying that it knew where the prey was. Knowing where something is, at a given moment, means representing the location of the object and also being such that had the object been in a (somewhat) different location one would have represented that as its location instead. So if an agent knows where something is, and it moves, the agent will know that it is at the subsequent position. So knowledge as tracking allows real tracking, and catching. Similarly, almost-knowledge can explain actions which are systematically but inaccurately related to an evolving series of facts. For example if a person is spearing a fish but does not take account of the refraction of light at the water's surface then their mistaken thrust can be explained by its being the result of a just-off tracking of the fish's location. (Presumably this is a basic reason for some basic features of our epistemic/psychological vocabulary. See Kornblith 1998.)

One consequence of the suitability of fact-linked states for explaining object-directed action is the frequent failure of psychological explanations to specify the way in which the actions concerned were carried out. (Why did I pick up the glass? Because I saw it on the table, and remembered that it should have been put away. The explanation fails to mention how seeing the glass enabled me to pick it up. It can get away with this omission because seeing involves knowing, and knowing means having reliable information, which can typically guide action.) As a number of philosophers have pointed out such explanations typically lose their force if the terms in them are replaced with terms that do not entail knowledge. (Pioneering work by Peacocke 1981, then Rudder Baker 1982, Carruthers 1987, 1988, McCulloch 1988, Williamson 1995.)

Some of the disadvantages of close links to the facts complement these advantages. When Lois Lane thinks that Clark Kent has taken her notebook she goes to look for him. She doesn't go to look for Superman, though the information she has is information about Superman. A belief – as contrasted with knowledge or almost-knowledge – about an object characterizes the object as related to the believer in particular ways, which determine the ways in which the person will act on the object as so characterized. The object, in fact, can fail to exist, or be essentially different from the agent's conception of it. A child writing a letter to Santa Claus is acting on a complex of beliefs, central ones of which are false. And by articulating them we can explain her action. But we can do this only by preserving the distance of her beliefs from knowledge and taking account of her individual way of thinking.[2] When we want to describe individual ways of thinking we need different resources. We have to incorporate, among other things, allusions to cultural inventions, individual tendencies to turn one way or another at the optional crossroads of inference, and the exact point at which force of evidence leads to assent.[3] The question is, how do we get from relating individuals to a common environment to the creation of these resources?

Eight subversive stories

In the two sections which now follow I shall try to convince you that your grasp of belief, especially when it cannot be glossed as 'almost knows', is not as secure as you may have thought. (In particular I want to make you see that you are more like a three-year-old who struggles with the idea that beliefs can be false – see Chapter 2 'The negative bit' – than you realize.) This present section has no real arguments in it. It is a series of examples together with leading questions, intended to prime your intuitions so that they more readily respond to the analysis of the following section. Some readers may not want to have their intuitions tampered with. (Intuition-pumps without arguments may be like glamorous pictures in advertising or too easily horrifying cases in moral persuasion. But the alternative is being bulldozed by logic, without having any sense that the conclusion proposed is really a reasonable reaction to the argument.) In any case, you have been warned. You have the option of skipping straight to 'What the stories show', and then referring back to this section as needed.

(i) The seventeenth-century bishop Jacques Bossuet, in his *Discours sur l'histoire universelle* gives a chronology of the world since its creation. The events mentioned in the Bible are given dates, and so are episodes in Greek mythology. Thus the battles of Hercules are dated as occurring a short time after Abimelech. Indeed at around the same time, according to Bossuet, occurred the death of 'Sarpedon the son of Jupiter'. But Bossuet also subscribed to the official Christian view that the Greek myths were indeed mythical while the biblical account is literal truth. Hercules never existed, and certainly Jupiter does not.[4]

Question: did Bossuet believe that Hercules and Abimelech were on the earth at about the same time?

Subversive follow-up: Was his attitude to this proposition, and the attitude that he was intending to induce in his readers, something that we cannot recapture using 'belief' in its contemporary sense? Would this attitude have been one that Bossuet and his readers would routinely have attributed to each other, using words that we would translate as 'believes'?

(ii) In B's pious upbringing religious doubt is a shameful thing. As a child he attends church and Sunday school regularly and his parents think he may become a priest. At university he studies philosophy and writes essays arguing that though there are no valid proofs of the existence of God any person can reasonably decide to believe. During this time he stops attending church, saying that his church is in the privacy of his mind. In fact, he only prays in times of crisis, and then he mechanically repeats the words he used as a child. He becomes an investment banker, and is known for requiring the same conditions and collateral for charitable institutions as for profit making enterprises. One of his colleagues tries to change his policy, using Christian arguments, but B simply replies 'that's irrelevant'. He invests a large proportion of his salary for his retirement, saying 'You can't leave anything to chance, and you certainly can't take it with you.' He says grace before every family meal.

Question: Does B believe in God?
Subversive follow-up: Does he simply think he believes? Does he pretend to the world, including to himself, that he believes?

(iii) C was born in Des Moines, Iowa. To impress a woman at a party he once pretended that he was born in Ireland but stolen from a cradle and smuggled to America. This becomes a standard routine at parties, and he finds that it annoys his wife in a way that amuses him. More and more often he tells the story to strangers. Sometimes when filling out a form he writes, under 'place of birth', 'Kilkenny'. (Once or twice he realizes that this will cause trouble, and rewrites the form.) He contracts a fatal disease and as he lies dying he asks for his ashes to be scattered in his birthplace of Kilkenny.

Question: Does C believe he was born in Ireland?
Subversive follow-up: Is he instead deeply wrapped up in a fantasy that involves being born in Ireland?

(iv) D comes across evidence that his neighbor Luigi is a hit-man for the Mafia. He learns that Luigi only pretends to go to his stated work, but actually spends the day in bars downtown, but sometimes in the evening is picked up by scary men in a black limousine wearing leather gloves. On the other hand Luigi's wife confides that she is afraid of her husband losing his job, and in conversation Luigi seems a kind and sensitive man. So an alternative hypothesis is that he has lost his job and is pretending to go to work, perhaps working in the evening at something possibly illicit but not violent. D continues to socialize with Luigi, and even writes a letter recommending him as a volunteer worker in an old people's home, in which he describes him as a caring and gentle individual. But D is very careful what he says in front of Luigi, and when the limousine arrives D takes extreme precautions not to be seen observing Luigi getting into it.

Question: Does D both believe that Luigi is a mafioso and that he is not? Does he believe that Luigi is probably a mafioso, though he might not be, or that he is probably not a mafioso, but just might be?
Subversive follow-up: Might one thread of D's plans and intentions be shaped by the thought that Luigi is a mafioso, and another thread be shaped by the assurance that he is not, in a way that allows no simple description of his motives on the occasions when the two threads combine?

(v) A four-year-old child, E, hears her mother talking about the gas chromatograph she uses at work. (Her mother is a forensic scientist.) She tells her friends 'My mum drives a gas chromatograph.'

Question: Does E believe her mother drives a gas chromatograph?

Subversive follow-up: Does she instead think (or know) that her mother has a belief, or knowledge, which governs her operation of some machine to which the words 'gas chromatograph' apply?

(vi) A sociologist of science, A, is questioning a physicist, B, about her work. B is doing experiments to determine some properties of the Higgs Boson. A has never studied any hard science beyond secondary school; he is now writing a PhD on the ways in which scientists share and communicate conjectures and theories, about whose truth and even intelligibility he remains profoundly agnostic. It all turns on issues of power and fashion, as far as A is concerned, and to maintain this pure attitude he makes a point of making no effort to understand any scientific theory. B's experiments lead her to assert that the Higgs Boson has energy e. This conclusion emerges during an interview between A and B.

Question: Does A believe that B believes that the Higgs Boson has energy e?

Subversive follow-up: Does A have instead the easier belief that B has some belief which she expresses using the to-A-meaningless words 'the Higgs Boson has energy e'?

Second question: Does A believe that B believes that some symbolic relationship constituting a move in the power-relationships between physicists can be successfully defended?

Second follow-up: Which is more appropriate to express what A thinks B thinks: a belief expressed in A's terms which B would not recognize or a belief expressed in B's terms which A could not understand?

(vii) You read stories to a six-year-old. They are scary stories about giants and witches. One day after you have been reading to the child her mother enters and you find yourselves discussing her ex-husband's present wife. Reacting to some scandalous tale you say 'she sounds like a real witch to me.' The small child is visibly terrified. She thinks she has been told that her stepmother is a witch, like some in the stories, but it takes you several minutes to realize that that is what she thinks.

Question: Did you believe that the child believed her stepmother was a witch, or that the child had a witch-fantasy based on stories, triggered by your metaphor?

Follow-up: Beliefs are true or false: could the child's reaction have been based on a thought that was either?

(viii) You have a long drive home from work every day. Last year you shared the drive with a colleague: sometimes she would drive and sometimes you would. Your colleague moves away to another job and so you drive by yourself. There is repair work to a bridge near your house, causing long delays, but you work out an alternative route that turns, just before the affected stretch, to follow a country lane. One day your friend is visiting your workplace and she offers you a lift home. You relax as she drives and only too late do you realize, with surprise, that she has not taken the turning for the country lane. But why should she, since she had not been told about the bridge repairs and the delays?

Question: Did you believe, falsely, that your friend believed that the best way home lay along the lane?

Follow-up: Did you instead believe truly that your friend knew the area, from which it might be natural to infer that she would believe that the best way led along the lane, though if you had explicitly posed the question you would have blocked the inference?

What the stories show

These stories will work in different ways on the intuitions of different people. I would not deny that in each of them 'believes' could be used in an informative way. But in each of them 'believes' could also be misleading, if used without the caution that the particular application requires. And each of them could reasonably provoke the reaction that the normal meaning of 'believes' simply does not allow an answer to the question of whether the attitude of the person in question is one of belief. Yet the stories are quite ordinary (much more realizable than most philosophical examples, though they have features in common with the cases in Kripke 1979, which also show deep tensions between different aspects of the concept of belief). So if our grasp of the concept falters when faced with them, it must be on the edge of faltering in the face of innumerable everyday attributions of belief. We must always be struggling to give 'believes' a meaning that will apply helpfully to the case at hand. Let me begin with a general diagnosis of what the stories show.

Conflicts of anchors

We anchor the concept of belief in several ways. We anchor belief to assertion: typically people believe what they say. (And associated with this we anchor belief to conscious thinking: when someone reasons they entertain possibilities, some of which they are aware of as having a kind of internal assent.) We anchor belief to evidential thought: people often believe what perception, other beliefs or knowledge, or inference gives them good reasons to believe. We anchor belief to practical thought: people's actions are often guided by coherent strands of information. We anchor one person's beliefs to those of others: we often understand a person's state of mind by assuming that their links to others will allow them to relate to a proposition in much the way that others do.

Sometimes the anchors pull in different directions. In stories (i) [Bossuet], (ii) [Piety], (iii) [Kilkenny], (v) [chomatograph] and (vii) [witches] our grasp on the person's state qua belief is based on their own assertions. But this conflicts with other criteria. In stories (i), (ii), (iii), (v) and (vii) [driving] it conflicts with criteria of coherence and evidence. In (ii) it conflicts with criteria of the use of information to guide action. In (iv) 'Mafia' also there is a conflict between what someone says and the information that guides their action. But in that example there is also a conflict between two ways of anchoring the person's behavior in information. The person uses different, incompatible, items of information to guide different acts, in a way that is intuitively intelligible and perhaps not unreasonable, but which does not make either information-base a better candidate for the content of his belief.

In stories (v) 'chomatograph' and (vi) 'sociology' the belief ascription is problematic because of the way the person is related to others. The person uses a word, thus creating an assertion anchor for their 'belief', but does not have the knowledge or the relation to others which would be needed to make that word the expression of the concept which the ascription presupposes. In story (vii) 'witches' the embedding of the person in the community is roughly adequate, but the person does not have the basis of other knowledge to make a normal use of that embedding.

It is worth noting, too, that the anchors tie down different aspects of belief. The link to evidence ties belief to facts that might make it true. The link to assertion ties belief to sentences that might be used in correct ascriptions. The link to practical thought ties belief to action. But a state can be tied down as well as can be in one of these respects and be free floating in another. For example a person reasoning confusedly to the existence of some mythical object can then search for it in a systematic and intelligible way. We can explain the search in terms of the 'belief', but be completely at a loss as to what English sentence to use to ascribe the belief or under what conditions it would be true. Similarly if a person who dogmatically asserts something is to be credited with a belief, it is a belief given by the sentence asserted, but this may be no help in determining what acts besides assertion the 'belief' will lead to.

Problems with higher order belief

In stories (vi), (vii) and (viii) the problems about one person's belief generate problems about other people's beliefs about what that person believes. Story (viii) 'driving' is the simplest case. In it there is a simple mistake: one person relies too much on the default assumption that another shares their knowledge. When this happens there is likely to be a problem about characterizing the first person's belief about the second person. It can be taken as a false belief about the truth value of the other person's belief, or as a failure to have a belief about the other person's belief. Story (vii) is similar in that a person has difficulty attributing a problematic belief. But here what is problematic about the belief runs deeper than falsity in a way that makes it hard to have any unproblematic belief about the belief. The person in story (vii) does not believe that there are witches and that the child thinks that her stepmother is one. Nor is the belief that the child thinks that there are people who cast evil spells and her stepmother is one. For the child's belief is less self-contained than that: the child knows stories about witches and assumes that these link to a more complete account of real witches which adults must have. But this assumption is false and so what the child concludes is not well-formed enough to be a full belief. It isn't about anything; it couldn't turn out to be true. As a result, the belief about the child's not-quite-belief is more slippery than it seems. It's not a belief about the child's relation to witches, nor about the child's theory of evil spell-casters. Perhaps the best way to express it is as a belief that the child is taking the witch stories literally and applying them to her stepmother.

We rarely notice these complications. Adults fail to see that a failure in concept-formation can lead to a non-belief, just as small children fail to see that a failure in inference can lead to a false belief. This way of putting it might obscure an important link, though. In almost all cases in which an ascription of belief is pulled in different directions by different anchors there will be problems also in second order belief. In story (i) 'Bossuet' we are uncertain not just whether Bossuet's state is one of belief but whether someone who thinks the answer is yes can really be described as believing that Bossuet believed that Hercules and Abimelech were contemporaries. For to put it that way attributes to the person a view that Bossuet has misunderstood the true relation between Hercules and Abimelech, which for all the reasons that the witch case brings out is not really right. The relevance of this fact to the problems three-year-olds face in attributing false beliefs is not often appreciated. If Maxi (the puppet figure to whom children must attribute beliefs in the classic experiment) had enormous intelligence – as grown ups do, and perhaps toys in a make-believe do – then he'd figure out that the experimenter would have moved the goodies to another location, and thus would believe that they are where they are. Most attributions of false belief require that we appreciate the ways in which the subject's rational powers are limited. The same over-estimations of a person's connections with the facts – by not appreciating cognitive, perceptual and other limitations – that can make it hard to attribute false beliefs can also make it hard to attribute failed belief. So the child acquiring the concept of belief faces difficult alternatives. Either she assumes that everyone is near to perfect at figuring things out, in which case she cannot mimic their thinking with hers and in any case will usually get a false prediction, or she has to grapple with the different ways in which people reason – the virtues and vices of the previous chapter – giving her a whole complex domain to master rather than one simple concept.

The most telling of the examples, with respect to higher order belief, is story (vi) 'sociology'. The second follow-up question goes to the heart of the matter. A description of what A believes B believes is a report of A's belief, namely that B believes that p. So as a report of A's beliefs it should use concepts familiar to A. But it is also a report of B's beliefs, as attributed by A, and this report has no chance of truth unless it uses concepts familiar to B. But very often we can't have it both ways: we have to use either A's concepts or B's. It is interesting that the link with A's beliefs is more direct than with B's, and yet in the example above, and others like it, it is more natural to use terms appropriate to B rather than to A. He thinks she believes that the Higgs Boson, rather than power-plays among physicists. The reason is most likely that we take the 'believes that believes' idiom to be delivering a report about the world, originating with what B would say and using A's thinking simply as a channel of transmission. In any case, there is a general moral here. *'A believes that B believes that p' is always potentially unstable. If p is constructed from A's stock of concepts then it may not be suitable to express a belief of B's, and if it is constructed from B's stock of concepts then it may not be suitable to express a belief of A's. But it has to do both.*

In practice when one says 'A believes that B believes that p' there are *three* stocks of concepts involved: A's, B's, and those shared by speaker and hearer. And thus a simple 'A believes that p' can lead to the same sorts of trouble if the speaker, the hearer, and the purported believer A do not share concepts. All but the most sophisticated adults ascribe beliefs as if there is a single stock of concepts understood by all, and as if the thoughts of different people differ only in which concepts from this stock any person employs at any time. And the most sophisticated, who can tell stories and devise idioms that avoid this fallacy, may be talking themselves right out of the domain of belief as commonly conceived.

The coordination of force

In many situations agents need to coordinate their actions. Consider situations in which coordination is the primary factor, that is, where the most important thing is that everyone make the same choice.[5] In such situations it helps if all the agents concerned have the same information. Often coordination is impossible otherwise. But it is not always necessary that the information be knowledge. Often it is enough that everyone share crucial beliefs. So if A, B and C are each to choose the same action it is often enough that each believe that p, and that each believe that all three believe that p. (Sometimes more is needed in the direction of fully mutual belief – each believing that each believes that each believes that p, and so on – without the beliefs having to be knowledge.) Almost any standard example of coordinated action will serve here. Consider cases of a kind familiar from Schelling's work, in which two people have to meet in a given town without an agreed meeting place or the possibility of communicating. If there is some location that they both think will occur as a meeting place to both of them, then they will each think the other is likely to go there, and will thus go there themselves. That is, they need a location such that for whatever reason they each think that the other will think that they are likely to go there. One reason might be that one just thinks that the other is likely to go there, which would not be rational unless the other also thought that the first would believe this. Another reason might be that one is inclined to go there herself, and thinks that the other will understand this. A more interesting reason might be that one thinks that the other thinks she is likely to go there. Even if there is no initial ground for the other's thought, it becomes self-fulfilling if the first knows or even thinks he has it. And so on for more remote higher order beliefs. (See Lewis 1968, Schiffer 1972 and Bacharach 1992).

And now the point. Knowledge is not needed for coordination; so what does it require of belief? Suppose that for belief we substitute some variant notion, for example fantasy or hope. And suppose that each agent takes it that the other will act on their fantasies or hopes. That is, suppose that both take it (for good or bad reasons) that fantasy information is treated as factual. Then if A assumes that B fantasizes that A will go to location L, A will conclude that B will act as if A will go to location L and will go there himself. And if B knows that A is making this assumption then he will indeed have reason to go to location L. So their shared belief that each will act on their fantasies is self-fulfilling. *As long as the assumption that a state will fill the coordinative role of belief is shared, the state will fill the coordinative role of belief.*

A variant possibility. We have three people, A, B and C, trying to coordinate. One of them, B, has a defective understanding of belief. On her account of belief if someone assigns a proposition a subjective probability of more than 0.5 then they believe it. A and C have a more sensible conception of belief, according to which on most issues there is a threshold probability for belief, but it is always well over 0.5. Suppose then that A has evidence that C, when considering whether A will go to location L, will conclude that there is a probability of 0.51 that she will. B will conceive of this as belief. And A, knowing this, will conclude that B believes that C believes that A will go to L. And if this is all they have to go on about where they might meet, and they all share the crucial information about each other's information, then they will end up going to location L. So B's flawed conception of belief functions to induce coordination just as more orthodox belief would. (The same considerations would apply with an even more perverse conception of belief.) *What is important often is not what people believe about the objective situation but what they believe they believe, and what someone believes someone else believes depends on the concept of belief the first person is employing.*

These factors are amplified by greater numbers of agents and greater degrees of embedding. The amplification works for any coordination-inducing factor. If I think that one of the twenty of us may have a fairly feeble reason for suspecting that another of us will go for option *O* then I may take it as a possibility that the first will go for *O* herself. So in the absence of any more promising alternative I may provisionally decide to go for *O* myself. And if this is generally known then others may conclude that they should certainly go for *O*, partly because I may well, and partly because its being generally known that I may well makes *O* a salient coordinating point for all. This amplification will be strongest in situations of what one might call monotonic coordination, where there is a pay-off from doing what a number of others are doing, even if some do not so choose, and the pay-off increases as the proportion approaches totality.

In the case of belief-concepts the effect is that if even a few people have an off-center conception of belief, and the off-center concepts are coordination-inducing, then it will be in the interests of the majority to act as if they shared the deviant concepts. If there is a continuum of deviations the majority will typically act as if all concerned shared the exact point of deviation of those who share the coordination-inducing factors. (One could think of a variety of such coordination-inducing factors, besides the two I have discussed: e.g. common views about which kinds of states motivate action in the manner of belief, and about what kinds of evidence justify an attribution of belief.)

I have considered cases in which coordination is needed and the distribution of motives or salience-making factors favors one understanding of belief. When there are two understandings of belief it can be that the discrepancy hampers coordination. Suppose there are three people, who must meet in New York tomorrow and Uno believes* that Dua is likely to go to The Empire State Building, but Dua and Trio do not count belief* as belief. Then Uno's state will not serve as a focus for coordination. Combining this with the other three-person example above we should conclude that discrepancies of criteria for belief-ascription either result in one criterion becoming the norm or in cooperation being blocked. So either we end up with a single concept of belief or we fail to coordinate.

The conclusion is that when there is a need for coordination and it cannot be based on knowledge, we are pushed towards uniformity in belief concepts. We coordinate our criteria for ascription and the psychologies we associate with belief-like states, in order to coordinate our actions. The uniformity can take two forms. It can be case by case or community wide. Case by case we adjust our criteria to fit the situation at hand and the dispositions of the people involved. Thus in a situation in which evidence is scarce we may find ourselves using 'belief' in such a way that a thought held with quite weak confidence counts as belief. On the other hand, in a situation in which there is confusingly much evidence we may filter out considerations that would distract us from working out one another's motives by counting as belief only what is near to certain. Community-wide uniformity prepares one to handle coordination problems in which one has little information about the dispositions of others, by providing defaults criteria for the ascription of beliefs, such that one can be fairly sure that the others are also applying them to oneself.

From this perspective several pieces of the puzzle that we saw earlier can be seen to fit together. We begin with a quasi-paradox. When people are striving to coordinate their actions they often describe their deliberations in terms of complicated patterns of reasoning ascribed to one another. But one source of complication is the degree of embedding involved, as has been evident in the reasoning described in this section. The two aspects ought to conflict, if what was said at the end of the previous section is correct. For to grasp reasoning we have to grasp the particular way in which a person represents a proposition, while to manage iterated attitudes we have to let go of that aspect and grasp the objective proposition which is represented. There is indeed a tension here, and its resolution tells us something about the way that we use the concept of belief. Consider again the Bossuet case, in which Bossuet's beliefs, and the beliefs he intended to induce in his readers, mix together real (or taken to be real) objects and clearly mythical ones. Consider a chain of reasoning much like that Bossuet must have gone through, calculating the intervals of time between events in terms of average life spans, ages at which people have children, and so on. This reasoning is essentially the same whether the events and generations are real or mythical. So a very natural device for describing it is to assume – for the sake of describing the reasoning – that we are dealing with real objects, so that we can describe the reasoning in terms of transitions between as-if-unproblematic beliefs. (We might do this ourselves if we were looking for mistakes in Bossuet's reasoning. 'See, here he assumes that Leda was 24 when she was impregnated by the swan; but I think she could have been as young as 16.') The process here is very similar to that by which off-center notions of belief become taken as normal. The effect is that we take beliefs which are formed in terms of information which does not ground them in the actual world as if they were so grounded, in order to be able to understand reasoning which uses this information. The currency becomes inflated, but up to a point it is in everyone's interest for this to happen. (If some others are treating monopoly money the same as real dollars, you won't be able to do business with them unless you do too.)

Our hesitations in ascribing some kinds of false belief and some kinds of non-belief also take on a new appearance. Suppose that you are coordinating your actions with someone who believes something that you do not, perhaps that Salvador Dali is the greatest artist who has ever lived. Very often you will find yourself acting as if the belief were true. (You can meet at the city museum where there is a Henry Moore exposition, or the state gallery, where there is a Dali show. You want to meet, so of course you go to the state gallery.) And similarly if the other person has some obsessive thought whose marginal intelligibility makes you reluctant to classify it even as a belief. In many circumstances if your aim is to coordinate, cooperate, or do business with that person you will find yourself including that thought in your own calculations almost as if it were a belief you at any rate understood, and sometimes even as if it were one you shared. In this way many of the dimensions of variation of belief – certainty, closeness to knowledge, reference to real things – will get ironed out as people interact cooperatively with one another. But – and this is vital to my argument – there is no uniform way in which they will get ironed out. All we can predict is that on a case to case, and to some extent on a culture to culture basis, as I investigate in Exploration III, people tend to ascribe to others states that are similar to those which the others will ascribe to them.

The coordination of content

Now shift the focus from whether a person's attitude to a proposition is belief, to what proposition it is that the person believes. It is hard to determine a person's exact beliefs from any evidence that is likely to be available. And so, this section will argue, we reduce the indeterminacy with factors derived from the need to coordinate our actions. Among the many ways in which the normally available evidence can fail to pin down what it is that someone believes, there are three deep and systematic sources of underdetermination. The first is due to the holism of the mental. Assume that an attribution of a set of beliefs and desires and other states (emotions, virtues) to a person is reasonable when the possession of those states would explain (or make sense of, or rationalize) the person's actions. Then for any finite set of actions there will be infinitely many combinations of states that it would be reasonable to attribute. For any state can always be replaced by a very different, indeed contradictory, one, as long as complementary changes are made elsewhere. (The contemporary form of the idea comes from Davidson 1970; for some developments of it see Chapters 1 and 2 of Child 1994.) Some of these alternatives will be simpler than others – it is easier to suppose that someone is buying a sandwich because he is hungry than that he wants to die and thinks that it is of a kind particularly likely to stick in his windpipe – but very often there will be equally simple pairs of states which fit what is known about the person's actions equally well.

The alternatives to any belief attribution due to holism are likely to be quite different from it. The next two sources of underdetermination provide more closely linked alternatives. The first, underdetermination, is clearest when states are attributed to non-linguistic animals. The emphasis here is not so much what explains action as what is causally responsible for a state. Suppose that an animal is in a content-bearing state as a result of some situation in its environment. What is often definite is the causal path that leads to the state, but this leaves undetermined what fact at what distance along the causal path the state refers to. There will usually thus be a proximal/distal or a subkind/superkind indeterminacy. Suppose for example that an animal is in a state that is produced by the presence of water in its environment. The animal moves towards the water and drinks. It seems at first clear that the animal is in a state that is about the location of the water. The state might even be described as knowledge or belief that there is water so far in that direction. But it is far from clear, really. The content of the state might as well be that there is drinkable liquid in the direction of some star, or that spring water can be obtained without much effort or that material fit for ingesting will result from moving particular limbs. A considerable and formidable literature has developed trying to add conditions that will pin down intuitively correct determinate contents for animals' representations of their environments. (See Dretske 1988, Millikan 1984, Papineau 1987, Fodor 1990, Sterelney 1995, Pietroski 1992, Godfrey-Smith 1994 and Neander 1995. See Chapter 8 of Dennett 1987 for the potential underdeterminacy of representation.)

The third indeterminacy is in a way the analog of the second for language-users. It creates the rule following problem that Kripke extracted from Wittgenstein. (See Kripke 1982, McDowell 1984, Boghossian 1989, Wright 1989 and Fane 1996. I take the issues of Wilson 1982 to be closely related.) Consider a community of people who apply some word to members of some subtle kind. Butterflies, for example, as distinct from moths and dragonflies. Each person makes 'mistakes' in that they apply the word where the others would not. Sometimes these mistakes are corrected and the correction acknowledged, but most often they are not. Suppose that there are in the world examples to which different members of the community will react differently, some applying the word and others not. As a result there is room for each person to doubt whether each other is thinking about the same kind as they are. For starting with any determination of the limits of the kind there are many variations, which fit the overall pattern of uses of the word just as well. None of these variations fits everything everyone says, and each can be reconciled with what people do say by assuming a certain number of mistakes, oversights and confusions. (But different ones in each case.) So the skeptical thought 'he may not be thinking about butterflies, but about some kind that largely overlaps with them' is very hard to dispel. And the skepticism is not just a matter of the practicalities of evidence, for it is not clear what about an individual links her to one of these variations rather than another. To take an individual's thoughts to center on the variation picked out by their actual and potential responses is to confirm the skeptical thought – everyone has a different understanding of 'butterfly' – and any way of rejecting some of the actual and potential responses as mistakes begs the question. I take it that it is clear that there can be concepts like this, and clear that all concepts are to some degree like this, and very controversial how seriously the resulting undetermination applies to all or most concepts. The crucial issue is what actual facts about a person or their surrounding community can tie that person's use of a word to a determinate objective kind. Some very impressive philosophers have come to impressively different conclusions about this. (My guess is that there is little fact to the matter about whether there is a fact to the matter about whether person p is thinking of kind K. That is, there are legitimate ways of understanding 'belief b refers to kind K' on which this indeterminacy of content is minimal, and legitimate ways on which it is very great. 'Belief' and 'reference' are not determinate enough to settle the question.)

Underdeterminations of all three kinds combine and overlap. The result may be worrying enough for philosophers, but consider the plight of people in everyday life having to face the worries of Davidson, Millikan and Kripke while thinking through simple shared activities. Consider two hunters who have had a vivid meeting with an insect the previous day and who are now setting out hoping to meet one another. Suppose that it occurs to each of them that the other will be going to search for an insect of the same kind, which means going to some particular location. Each will go to the location associated with the kind that he thinks both will take the other to have in mind. But which of the many variant kinds containing the insect in question is this? Each will be happy to focus on the kind that the other takes him to have in mind, even if that is not actually how he classified the insect, or indeed to focus on a kind that neither of them had in mind as long as each can take the other to be taking him to have it in mind. A myth will do perfectly well here, as long as it is a shared myth, for coordinating the contents of their attributions so they can coordinate their actions. But there is no obvious way for them to find truth, half-truth, or myth that will help them.

There is no obvious way, that is, as long as they are starting from scratch, trying to coordinate the contents of their attributions just to facilitate this single activity. What people actually do is to prepare in advance a coordination of their attributions of content, which will make many action-coordinations possible. Suppose that we could pick some definite and intelligible kind and treat that as the canonical version: in the absence of evidence to the contrary this is the variant that we will take one another to have in mind. (It might be the descendants of the primal butterfly as created on the first day of the world. And we might have a belief that *those* insects typically congregate in a specific location, which is where we will both go, even if there is no non-conventional reason to think that the other took the insect in question to be an instance of exactly that kind.) Now of course most often defining such a canonical variant will be as hard as any other instance of the general problem of distinguishing one variant from another. But once we see that what matters is that everyone take themselves to have the same kind in mind, we see that it doesn't really matter whether the canonical version is definite and determinate. We can just declare it to be. And this is what people do. They invent labels – 'butterfly', 'mosquito', 'hummingbird' – and make sure not to invent labels for kinds that vary these. (So there is no simple label for ‘butterfly except for those with orange wings or with wing-flap frequencies above two per second’, and so on.) And then although each person connects with the label, and with the objects it is applied to, via an idiosyncratic bundle of dispositions which potentially define as many extensions as there are people, each takes himself to be having beliefs about *the* kind, and takes others to be, and takes others to take him to be.

We do this for many objects of thought, simultaneously. A crucial element is simply the possession of language, with a discrete set of words for things, kinds and properties. If we make sure that the words of our basic vocabulary do not overlap too much in ways that make hard-to-discern differences between them, we can take them to mark fixed objects of our thinking, even if we have only the vaguest ideas what the shapes of these fixities are. We have to take care not to undermine the plot, though. We have to avoid perverse ascriptions of in-between or variant concepts; we have to correct mis-descriptions instead of taking them as signs of interesting new thoughts. If we play it right then we can act as if we lived in an Aristotelian universe in which there are distinct individuals that fall into a finite number of hierarchically organized unchanging kinds, forms of which can also inhabit our minds. And then if we are setting out to the location of the insect we saw yesterday, there is no problem. It was a butterfly, and each would have thought it was a butterfly and would have thought that the other thought that it was, so we can each set off for where according to our common lore the butterflies are found, safe in the assumption that the other has ascribed the same intention to us.

Propositions and inferences

Not all strategic situations are coordination problems. But most situations will put similar pressures on the belief attributions that their participants make to one another. Very often it will be in the participants’ interest that the attributions be in some way coordinated. (There are somewhat artificial situations in which it is in their interest that they take the contents to be as different as possible.) And the conclusion is in a way unsurprising. With a loose everyday concept like belief we can expect that the exact force of its attribution will vary from occasion to occasion in accordance with the purposes of those attributing it. What the argument of the previous two sections adds to this is that when the interests of all the interacting people become relevant their resolution can become partially constitutive of the concept’s application. (There is another theme in these sections, the special relevance of iterated attitudes – beliefs that beliefs that p – to strategic situations, and the pressure this puts on the coordination of states. I do not know how to extract this factor from the argument and state it in proper generality.)

The marvel is not that belief varies from application to application, but that its variation is part of its power to predict and explain. Part of the reason is found in one of the general themes of this book, that our everyday psychological concepts are most effective when their application is one of the reasons for the phenomena they are used to explain. In the case of belief, this takes the form of the observation that the very fact that people are adjusting the parameters of their understandings of belief to bring them into line with one another is part of the reason that their attempts to coordinate their actions, which they do in part by reasoning in terms of belief, succeed. It follows from this that it would not serve the interests of folk psychology if we were to disambiguate belief into several more precise concepts and then stick rigidly to their literal meanings, as several philosophers have proposed. (See de Sousa 1971, Dennett 1978, van Fraassen 1980, Cohen 1992, Sperber 1996, Stevenson 2002.) We need a concept that can be pulled in different directions for different purposes. We need a multiply anchored concept. We can get a bit clearer about the different pulls on it by attending to some issues about intensionality and logical form.

Take belief to be a relation to a proposition, which is specified by the sentence embedded in the that-clause of 'A believes that . . .'. A proposition can be many things, and 'specified by' can be taken in many ways. The tensions over the interpretation of propositions, in this context, arise from the fact that we use the p in 'A believes that p' for two fundamentally different purposes. Beliefs are true or false, and are also premises and conclusions of inference. It is certainly possible for a thought to figure in reasoning although it is neither true nor false. If you assume that the Jabberwock is gyring in the wabe then you may conclude that something is gyring in the wabe, though it is doubtful that either your assumption or your conclusion has a truth-value. And it is possible for a state to have a truth-value without being capable of figuring in inference. If a salmon knows that this is the stream it must ascend to mate then it has truly represented its native stream, though the representation is not suitable material for inference, at least not inference that exploits the syntactical structure of the sentence with which we specify it.

There are two very general attitudes one can take to this two-sided-ness of that-clauses. One is the Sellarsian view that the coincidence shows the psychological insight of our ancestors. (See Sellars 1963, also Brentano 1995/1874.) On this view, it is a profound fact about mind, perhaps the basic fact, that truth-values and inference take the same vehicles. The model is sentences of spoken languages, which can serve both roles; and, after all, when we attribute beliefs we specify their contents by giving sentences of spoken languages.

The other view, to which I subscribe, is that what we have is as much obstacle as insight, that we are dealing with an economy of language that runs together two essentially different things. One is a worldly thing, the truth-conditions of the belief, the possible fact that if it obtained would make it true, or which if epistemic conditions were met would be what the person knew by holding their belief. And the other is a mental thing, the *way* in which the belief is held, the nature of the representation or other state by virtue of which the person can connect with the possible fact. On this view one could in principle specify a belief by focussing on either the world side of the business or the mind side, then adding some more information to say what was needed about the other. Thus one might focus on the mind side and say ‘she has a belief that there is an object with a certain mysterious property, whose constant existence has been a comfort to her throughout her life’. The Henry James-like that-sentence here gives a partial grasp of what kind of a mental state she is in, while giving almost no hold on whether the belief is true or false. Given a conversational context a hearer might be able to figure out whether it was true or not.

The most influential contemporary version of this attitude prefers the opposite strategy, stating the truth conditions explicitly and letting the mental grasp of them be expressed in a variety of different context-sensitive ways. This view is usually expressed in terms of an insistence that the propositions in question are Russellian rather than Fregean. That is, they are specified by (for example) a list of objective individual things and an objective relation which would hold between the things if the proposition were a fact, and nothing more. On this view the way in which the person to whom the belief is ascribed articulates or represents this proposition is not given by the that-sentence. In fact it is not explicitly given at all; the hearer has to infer it from contextual information.[6] (See Perry 1977, Salmon 1986, Zalta 1988 and Kaplan 1989. It is not at all obvious how to define a Russellian proposition in full generality. I shall not even try. See Chapter 5 of Dennett 1987 for an evocation of quite how different from Russellian propositions our thoughts might be.)

There are two main reasons for believing that this attitude is generally right. The first is that we often do ascribe a belief to someone by using a that-sentence involving demonstratives, indexicals, or anaphoric pronouns, where the only information that can be taken to be part of the meaning of the sentence is the specification of the object referred to. We say ‘Mary believes that you put it there’. It is obviously not part of the content of the ascription how Mary thinks of the you-person, or how she thinks of the location where they are or the thing they are referring to by ‘it’. But, most of the time, speakers provide enough information, by the time they get to make the ascription, that these things can be inferred. (That is not to say anything about *how* they are inferred. My own view is that hearers infer the fuller state of mind of the believer in part by using rich folk psychological capacities, including various capacities to imagine how it is with another person. But that is definitely not to be taken as part of the thesis in question now.) So if we need these inferential capacities when dealing with this large class of belief ascriptions then it would be surprising if we did not use them for other classes of that-sentences as well.

The other reason concerns the defeasibility of the intensional content of belief ascriptions. Put more intelligibly, the reason starts from the observation that it is possible to say e.g. 'Mary believes that Hesperus is visible' in order to convey Mary's belief when she does *not* think of Hesperus as the evening star or Venus-as-seen in the evening. She may think of it as the morning star or just as Venus. For example a bunch of astronomers are discussing the views of someone in the village near the observatory who thinks not only that Hesperus and Phosphorus are distinct but that Hesperus has been made invisible by gods angered by disrespectful scientists. The astronomers are all male, except for Mary, and one of them makes the sexist remark 'well, that's the sort of thing a woman would believe'. Another replies '*Mary* believes that Hesperus is visible, in fact she took a photo of it this morning'. Note that the photo is taken in the morning. As with the first argument, we should conclude that there are delicate context-dependent patterns of reasoning that take hearers from the that-sentence in context to an understanding of the mental content of the person to whom the belief is ascribed.

There are difficult problems for this Russellian understanding of belief. It has to handle ascriptions of beliefs about non-existent or fictional entities, and beliefs about identity. There are problems and hard choices associated with both of these. I am not going to discuss possible solutions here. There is another apparent problem that the position does not have to handle. It does not have to explain how, knowing that a person believes that this object has this property, but not knowing how this person thinks of this object and this property, we can explain why she acted in this way rather than that with respect to it. At any rate, it does not have to explain how this can happen in order to defend itself against the Fregean view. For the Fregean position has the same problem. Whichever view of belief we take, it is not going to be obvious how we get an understanding of the information a person needs to carry out an act. It is just mysterious how we get this from the material at hand: the explicit content of a set of ascriptions of beliefs and desires, what else we know from the conversational content, and our acquaintance with the person. In fact, the difficulty of the problem should push us away from the Fregean position. For the Fregean position insists that belief ascriptions are fully informative in isolation, so the information we need should all be got from the that-sentence alone. But given the complexity of the information needed to guide an action, and the fact that with many large classes of that-sentences we obviously cannot take the information to be given by the meaning of the sentence alone, this seems like an arbitrary restriction of the ways we can describe states of mind.

I shall assume something weaker than the standard Russellian position. I shall assume that when we say 'A believes that s' we intend to communicate both something about A's state of mind and the truth conditions of its content. Very often we do this by using the content sentence to specify the truth conditions and then inferring from contextual and other information how the person is related to these truth conditions. We may also sometimes use the that-sentence to specify essential features of the person's state of mind while leaving the truth conditions to be inferred. (It is plausible that this is what is going on when we say that someone believes that Santa Claus has a white beard.) In either case the hearer obtains a variety of information about the information available to the believer, which can be used as part of an explanation of an action. But very rarely can this attributed information consist of the meaning of a specific English sentence.

If this assumption is right, and it is a position that is entailed by most Russellian accounts and even some Fregean accounts, then when we ascribe a belief to a person we convey two things. We convey that the person has a belief-relation to a (Russellian) proposition, and that the person has a mental representation or other information-bearer with a certain capacity to support inferences and guide action. This suggests a corresponding bifurcation in what I called the anchors of the concept and the ways that they can conflict. For someone to have a belief, there must first be some information that the person holds under some mode of presentation. Call this the information stage. Then this information held in this way must relate the person to a determinate proposition. Call this the proposition stage. There will be uncertainties, problems and conflicts at both stages. And each of the anchors ties belief down at both stages. Speech dispositions give both the verbal form of the information in terms of which the person reasons and the propositions to which the words in asserted sentences refer. Evidence is both intrinsically connected with reasoning and is also evidence about or deriving from facts in the world. Action guidance is both a matter of information management and produces interactions with particular objects. In the problem cases of 'Eight subversive stories' the more troubling cases were the ones where the informational or propositional component of one anchor conflicted with the same component of another. Thus in the 'sociology of science' case the verbal anchor ties the propositional content of the sociologist's belief to the physicist's belief to the Higgs Boson. But the evidential anchor ties the proposition to the power relations between scientists (that's what the evidence is about.) In the 'Kilkenny' case, on the other hand, the propositional content is perfectly clear, but the information is inconsistently anchored, evidentially in a way that disqualifies (rather than simply fails to qualify) it as belief, and behaviorally in a way typical of belief. In other cases things are fine at the information stage, but the information is not such as to generate a proposition ('Chromatograph', 'witch', 'Bossuet'.) Or the information stage can fail in a way that prevents the believer from referring to a single proposition ('Mafia'). (There are other cases in which the information stage is fine but the world does not supply a single proposition. A man marries a woman who is later secretly replaced by her identical twin . . .).

False belief problems and non-belief problems fall into the pattern. Difficulties in ascribing false beliefs can arise from relating a person to a proposition in terms of information that is not in fact true to that person's thinking. The default information, for children and adults, is the information common to attributer and audience. When this is not shared by the person to whom the belief is attributed, that person may be related to a fact via information they do not in fact possess. Older children and adults have overcome such problems by learning to read the implicit information element off a belief attribution in its context. (And by coming to understand how people use the information available to them.) But in so doing they make themselves vulnerable to an opposite problem, of taking the information to establish a belief when it does not determine a definite proposition.

One last link between Russellianism and the wobbles of belief. I pointed out in 'What the stories show' that the conceptual content of iterated belief constructions – he believes that she believes that they believe . . . – becomes more and more problematic as the depth of iteration increases. But the Russellian proposition can remain unchanged. John hopes that Mary thinks her husband knows that her mother suspects him of laughing at her cat. With just a little context-setting this becomes perfectly digestible *as long as* we take it that John, Mary, her husband and her mother all think of her husband and her cat in the same way, indeed in the same way as we are presenting them in the conversational context. (Triple and quadruple embeddings are harder if the same attitude, 'believes' say, is used for all of them. I think this is a matter of linguistic processing rather than of managing the thought to be conveyed.) If we want to use this ascription to explain reasoning by any of the people that requires us to grasp their own particular representation of the proposition, things quickly become unmanageable. Two conclusions emerge. The comprehension of iterated attitudes often requires us to use the default assumption that everyone concerned is using similar information (senses, modes of presentation) about the objects concerned. And in understanding iterated ascriptions of attitudes we focus on the bare Russellian proposition much more than on the individual ways of thinking of it.

Balancing belief-ascriptions

On each occasion when a person truly ascribes a belief to another, or to herself, there is a balance between the various components of belief which characterizes that person's mind. (In the previous section 'Propositions and inferences' and in 'What the stories show', I indicated the kinds of components that seem right. But finding a real and fundamental list is a deep and difficult philosophical and psychological task.) It does not follow that there is a balance between the components that applies whenever someone can be truly said to believe something. Indeed it is consistent with all I have said so far that our occasion-to-occasion balances range randomly and unpredictably over the whole possible range, as the pressures of moment-to-moment strategic situations determine. But it is not very plausible that they do. Individual interactions take place in the context of larger ones, ultimately in that of conventions, long-term implicit social contracts, that generate self-fulfilling expectations about the ways that individuals will interact on particular occasions.

The most pervasive of such conventions is language. In ascribing a belief to a person each ascriber is constrained by the conventional meanings of the words she uses, and by the semantic and pragmatic rules limiting the states that a form of words in a given context can intelligibly be used to attribute. It is conceivable that the latter vary from one language and culture to another. To end this chapter I want to discuss the effects of the former, of the net of reference that we collectively throw over the world and use to relate one another to it. Consider the situation of a person coming into a society in which people already have a rich network of shared beliefs, names for things and properties, authorities for settling questions of reference, and so on. Suppose that the newcomer is capable of entering into the occasion-to-occasion coordinative fine tuning of the concept of belief, sometimes stretching to the allowable limits. The newcomer will still have to operate within the semantic limits of the language. In order to refer to some things he will have to share some beliefs or take some authorities as authoritative. (So some objects will be easy to refer to and others will require complicated explanations or conflicts with received wisdom.) There will be words for some objects and kinds that are generally acknowledged not to exist (Santa Claus, witches, phlogiston), but no words for infinitely many other potential objects and kinds. In this situation the newcomer will find some of his own patterns of thought easy to describe to others and others almost impossible.

Consider for example states of mind that refer to particular non-existing (or at any rate invisible) objects or kinds. If there are no words for these things in the common language, and no resources for non-heroic reference to them, then the newcomer will find that actions that are motivated by thoughts about them are not intelligible to those around him. So such actions will less often get the cooperation of others that they need. So, in the long term, he will perform fewer of such actions. So, eventually, he will come more and more to act in ways that are intelligible in terms of beliefs that can be described using the commonly available resources.

The newcomer will thus come to coordinate his beliefs – not just the beliefs he ascribes but the beliefs he *has* – with those that are generally expressed around him. The coordination focusses on different aspects of belief than the small-scale coordinations described in the previous sections. But it is still a coordination: the newcomer comes to ascribe beliefs in the same terms as those around him, just because those are the terms they use. But there is another, in a way more fundamental, consequence. The newcomer will come to act in ways that are describable in terms of the surrounding belief-supporting network. That is the way to be understood and thus to achieve cooperative results. Thoughts that travel through indescribable territory will become less common, and actions that would result from such thoughts will also become less common. So as the newcomer adapts to the conditions around him the ascription of beliefs and desires by others becomes more accurate as an explanation of his actions and more effective as predictions of them.

So even when ascription of beliefs is part of a successful explanation or prediction of actions, part of the reason is that the agents have been shaped by their membership in a community of belief-ascribers to act on thoughts that can be cooperatively understood by others. (How large a part of the reason for the success of belief-ascription is this factor: the shaping of the believers by the institution of ascription? I don't know, and I discourage you from guessing. It is part of the first not-to-be-begged question of Chapter 1.)

A central purpose of belief – as contrasted with knowledge – is to describe people's individual information, as they conceive it. Often false, always linked to the person's own sources of evidence, patterns of reasoning and armory of concepts. And it is also essential to belief that beliefs be candidates for knowledge, whose purpose is to relate people to a shared objective world. There is a very basic tension here. One fundamental reason for wanting to grasp people's reasoning, and the information it consumes, is in order to manage coordinated action (and other more complex cooperations). And as the previous section has shown this creates a pressure to measure different people's beliefs on the same scales, to make my beliefs comparable to yours. On the other hand we often want to understand the idiosyncratic actions of individual people by teasing out the fine structure of their reasoning, and to do this we want as fine and individual a grasp as we can get of what is going through their minds. So we want our attributions of belief to represent individual states accurately, while not making them more individual than we must.

Is there an ideal level here, a stable compromise between making us similar and making us different? I am sure there is not. When we attribute beliefs we find the balance between degrees of certainty, degrees of conceptual commonality and responsiveness to evidence that will work best for the cooperative or defensive project in hand. We find the best way of making ourselves and others intelligible in the context of this project. And this process is largely invisible to us, for in coordinating our belief-concepts we are tacitly agreeing to take it as factual that our beliefs mesh in the required ways. To think about their differences would be to undo the coordination.

Chapter 4

Explanatory contrast and causal depth

> As with many men who achieve distinction, this feeling was far from self-serving but consisted in a deep love of the general good above personal advantage – in other words, he sincerely venerated the state of affairs that had served him so well, not because it was to his advantage, but because he was in harmony with and coexistent with it, and on general principles . . . even a pedigreed dog searches out his place under the dining table, regardless of kicks, not because of canine abjection but out of loyalty and faith; and even coldly calculating people do not succeed half so well in life as those with properly blended temperaments who are capable of deep feeling for those persons and conditions that happen to serve their own interests.
>
> Robert Musil, *The Man without Qualities*

> It is in pursuit of the good, then, that we walk when we walk, thinking this the better course, and when on the contrary we stand, we stand for the same reason, for the same reason, for the sake of the good.
>
> Plato, *Gorgias* 468b

> [Husserl] a fait la place nette pour un nouveau traité des passions qui s'inspirerait de cette vérité si simple et si profondément méconnue par nos raffinés: si nous aimons une femme, c'est parce qu'elle est aimable. Nous voilà délivrés de Proust.
>
> Jean-Paul Sartre, *Situations I*

We are not always trying to predict what people will do, or coordinate our actions with theirs. Often when we think about people we are simply trying to understand them. We express our attempts at understanding as explanations of their actions. The motive can be future prediction or cooperation, or just intrinsic human curiosity. The purpose of this chapter is to argue that we get better explanations if we focus on ends rather than motives. Or, more carefully put, that explanations that invoke certain factors which induce mutual intelligibility between agents tend to be causally deeper than explanations that invoke individual beliefs and desires. Many of the examples of such factors which will appear are individual and shared goods. This chapter thus makes a link between Chapter 2’s emphasis on the fragility and incompleteness of belief–desire explanation and a very general idea, which surfaces again at the very end of this book (the last section of Exploration IV) but for which the book does not claim to give a complete argument, that our deeper understandings of ourselves will generally give the potentiality for moral insight.

Flawed explanations

Explanations vary in value. Many of the explanations we give in everyday life are fairly feeble. That is, they lack some of a variety of features that explanations have when they give real understanding of why an event occurred or why a fact is the way it is. The sources and varieties of unsatisfactory explanations are very hard to state in a systematic and helpful way. (If it were not so hard philosophers would find their role of guardians against conventional complacency and intellectual fraud easier to fulfill.) But it is clear that explanations are often *causally shallow*. For example we might explain why a car failed to start by saying that it was raining (but why did this car on this occasion find the humidity affecting its ignition?). Or we might explain why a person did not volunteer for a task because it was a hard one (but why was the difficulty more daunting than this other task which on another day she took on readily?). An explanation can also be *unspecific* in its link to the event explained. For example we might explain why bird droppings fell on your head by saying that they are heavy and fall downwards. (But why did they spare my head, when I was beside you and the nest was not directly above either of us?) Or we might explain why someone left the house by saying that they needed to do some shopping. (But why did he leave at this moment when the shopping has been waiting all week, and why go in this direction rather than that?) An explanation can also be *epistemically insecure*: it can appeal to factors whose presence in the given situation can be very hard to determine. For example we might explain why a pebble thrown very hard against a window shattered it by saying that the pebble had enough momentum to overcome the integrity of the glass. (But how would we ever know this was the case, when perhaps instead the pebble could have chanced on a microscopic flaw?) Or we might explain why someone refused help from a stranger by saying that the stranger reminded her of her bossy mother. (But how would we ever know this was the case, when perhaps instead the stranger elicited a desire not to bother her long-suffering over-burdened, though bossy, mother?)

The ubiquity of flawed explanations may be more evident in articulated folk psychology, taking this to be our practice of applying attributions, predictions, and explanations to one another, than anywhere else in our lives. We are complex and mysterious beings, and yet it is essential that we have some hold on one another's actions and dispositions. So we are often satisfied with very imperfect explanations. And sometimes we cheat: we pass off as explanations things that have little or no explanatory force at all. In the extreme case this is what is known as psychobabble, nonsense masquerading as understanding. Psychobabble is hard to diagnose convincingly. One reason is that we sometimes manage to convey a real understanding of why a particular person did a particular thing in the most indirect improvisatory way. We give scraps of information, put together a narrative frame, and invoke imaginative capacities, and the result is that the hearer has acquired a capacity to understand, and perhaps even to predict, a range of the actions and emotions of the person in question. Yet often on other occasions similar rituals involving similar words give only the illusion of understanding.

The aim of this chapter is to engage with these issues in a very limited way. I shall defend a main and two subsidiary claims about the variation in value of everyday psychological explanations. The main claim is that one kind of explanatory force is linked to the contrastivity of explanations: explaining why this *rather than* that occurred. I argue that explanations can gain what I call causal depth by restricting their width of contrast. In the case of everyday psychology one important way of doing this is by means of what I shall call explanation by attractor. Attractors are, crudely put, desirable outcomes to which our motivation can gravitate. So the main claim is that restriction of explanatory contrast is a central aspect of everyday psychological explanation. And the first subsidiary claim is that the restriction often takes the form of a linkage between the agent's actions and results that satisfy some necessary conditions for being objective goods. The second subsidiary claim concerns the need for the restriction to be systematic. The range of outcomes to which people are attracted and the range presupposed by psychological explanation must not be too different, and so, as we shall see, a mutual shaping of the patterns of action and the presuppositions of explanation emerges.

Causal depth and contrastivity

The purpose of this section is to explain a fact about explanation that is presupposed by the rest of the chapter. It is reasonable to call this a fact rather than a position or a claim, since if it is formulated carefully it can be proved. The careful formulation and the proof are not in this section, though, but in the appendix to the chapter. In this section I shall try to give the idea intuitively in a way that makes it clear how the fact connects with our larger themes.

When we explain a fact or event – the explained facts – we do so by referring to other facts or events – the explainers. I shall focus on the particular case in which we explain one fact by referring to a single explainer.

I shall assume that the explainer is real. (One might explain away signs of guilt by giving a convincing line about how the bloodstains were really paint and at the crucial hour one was persuading a stranger not to jump off a bridge. But these will not count as explaining the bloodstains and the gap in one's alibi, if one was lying. They only appear to explain.) I am thus setting aside one category of flawed explanations, those in which the putative explainers do not exist. Then, following a now large literature I shall assume that the explainer is among the causes of the explained. (See Lewis 1986, 2000, McDermott 1995 and Paul 2000.) And, still sticking to a standard line, I shall understand this causal link to require the truth of certain subjunctive conditionals. The basic idea is that when one thing is a reason why another happened the occurrence of the second must be relevant to the occurrence of the first. If it had not happened then that would have been an obstacle to the occurrence of the first. I want a particular kind of conditional connection, though. I want a connection that is strong enough that citing the one event would be a reasonable answer to asking 'why did it happen?' of the other. And I want a connection that is weak enough to survive the presence of actual and possible perturbing factors, wobbles, which can prevent the cause from producing the effect. The reason for wanting to allow inherently wobbly connections is just that the intended application is to the link between motives and actions. The most ideal and overwhelming reason for doing something can easily be blocked by inertia, weakness of will, failure of nerve, or just some episode of mind or brain of which common sense has no grasp. So in allowing that 'because she knew that otherwise she would drown' can explain why someone took the three strokes to the life-ring, we must not allow the fact that she could terrifyingly easily just not have taken them to undermine the explanation.

To see what is needed consider an event c which when it occurs can be part of the bringing about of a situation E. But c is not always accompanied by E. Under some conditions c occurs and E fails to follow. c might be a match being struck, or a person seeing that a freight train is bearing down on her. E might be the match being lit, or the person moving quickly out of the way. So there are conditions P such that when c is accompanied by P, e will follow. c is required for the production of E, it is causally and explanatorily relevant, given P, in that if P holds and c is absent then E is also absent.

It is not easy to formulate this carefully so that it covers the right range of cases. We need conditions that link the cause to the effect which, conditions being favorable, it would produce: the cause must be roughly sufficient. These conditions say that if C occurs and P is present then the way is open for e to occur. And we need conditions that show that the apparent cause is not a by-product or companion of some real cause: the cause must be roughly necessary. These conditions say that if C doesn't occur though P is present then e will not occur. A lot of philosophical effort has recently gone into honing conditions like these. My own versions of them are in the appendix.

(One vital question is how to understand the 'if', especially in the sufficiency condition: given that C and e have both occurred, the claim that if C had occurred e would have, is in danger of vacuity. A reader who knows about these matters may want to quarrel with my treatment of them. Hence the relegation to an appendix since most such quarrels are irrelevant to the main points of the chapter.) At any rate, assume that we have suitable conditional links expressing both a mild necessity, subject to derailment by possible wobble-factors and consistent with some degree of causal indeterminacy or randomness, and a mild sufficiency, which are plausible candidates for saying that the presence of the event C can be cited as a causal explanation of why e occurred.

Sometimes an explanation that meets these conditions will be very weak. Suppose for example that a marble is placed on a sloping surface covered with small bumps and rough patches. It is put on point p and rolls to point q. The fact that it started at p explains why it ended up at q. For if it had not started at p it would have ended up somewhere else – let us suppose – and given that it starts at p, q is where it will end up. The possible wobbles that have to be assumed are the bumps and rough patches: if they had not been as they were then the marble's actual and possible trajectories would have been quite different. Suppose now that we insert several marbles at once. Again one marble starts at p and rolls to q, but now collisions with other marbles are part of the story. Again we can explain why the marble ended up at q by saying that it started at p. But this time the explanation seems intuitively a lot weaker. It seems to appeal to a very fragile connection between events. And indeed it does: if the other marbles had not been placed precisely as they were then the collisions would have been different and the result could have been a different destination. (Indeed the result would be sensitive to more than the insertions of the other marbles. Many-collision processes, physicists tell me, are highly unstable, and tiny vibrations or even the gravitational effects of large objects nearby can change their outcomes.) The wobble factor now for the explanation includes the placement of the other marbles, and that could easily have been slightly different, with very different consequences. (Our intuitions here are affected not just by the delicacy of the background factors but the time at which they are relevant. The fact that the other marbles interact with the marble's trajectory after it has begun inclines us more strongly to think of them as weakening the link between the initial and final point of that trajectory. I think this affects intuitions about cause more than it does intuitions about explanation.)

I shall still count 'because it started at p' as an explanation in the many marbles case. But it is a less good explanation than in the one marble case. It is less good in respect to what I shall call *causal depth*. Given two facts or events C1 and C2, I shall say that C1 provides a causally deeper explanation of an event e, when both C1 and C2 satisfy the conditions of rough necessity and rough sufficiency, but the nearest possible situation in which these conditions fail for the C1 to e connection is more remote than the nearest possible situation in which they fail for the C2 to e connection. Thus the explanation in the one marble case is causally deeper than in the many marbles case because, on the most natural ways of filling in the details, it could more easily happen that the other marbles took different trajectories than that the surface had different bumps and rough spots.

This definition applies plausibly to a range of examples. Thus the presence of oxygen provides a causally deeper explanation of a match's lighting than its not being in outer space does. And a person's seeing a train bearing down on her is a deeper explanation of her getting out of its way than her simply being close in front of it. Still, really deep explanations of particular events expressed in common sense terms are pretty rare. This is at least in part because common sense thinking aims to give quick answers that are accurate in the most commonly found or most significant conditions, ignoring occasional or even frequent exceptions if they do not cause practical problems. This is one reason why common sense lacks analogs of scientific procedures for isolating systems and separating one causal factor from another. But these are just what are needed in order to know whether in the absence of one factor another would still have its normal effect. (Would the ball still have arrived at q if the rough patches had been smooth, or if the other ball that is usually started at p′ had been started at p* instead? You have to do experiments, perhaps inventing some terminology to describe them, rather than generalizing from the more important things that usually happen.)

The causal shallowness of much everyday explanation is undeniable. It applies to everyday psychology too. In fact it applies even more evidently there than in other commonsense domains. It is inevitable given the complexity of human beings and the fact that our underlying causal processes are not open to everyday observation. So the commonsense of mind and action, folk psychology, centers its attention on a few pervasive if unreliable patterns of particular importance in our efforts to live orderly social lives. BUT to say just this ignores something vital. It ignores the distinction between explaining something inadequately and making no claim to explain it. The distinction is between explaining why e *simpliciter* and explaining why e *rather than* some alternative e*. To explain why e occurred rather than a specific alternative or class of alternatives is to renounce any pretension to explaining why e happened rather than some alternative not specified in the explicit contrast. Thus to explain why the marble that began at p ended at q rather than at r is to renounce any claim to explain why it ended at q rather than at some yet other destination s. To explain why someone got angry rather than sad is to disavow any claim to be explaining why they got angry rather than dropping down unconscious or having an epileptic fit.

By presenting explanations as contrastive, as explaining why one thing occurs rather than an alternative, then, we can exclude the relevance of the factors that limit the explanation's force. Framed this way, the central idea of this chapter is not very surprising. *Narrower contrast means greater depth.* That is, if we replace an explanation 'e occurred because c' with an explanation 'e rather than e* occurred because c' then the substituted contrastive explanation will generally have greater causal depth. And more generally, we can replace 'e rather than e* occurred because c' with 'e rather than e^+ occurred because c' where e^+ is a narrower alternative to e than e* is, that is one which is true in a smaller range of possibilities. Then the resulting explanation will have greater causal depth. (Non-contrastive explanation can be seen as contrastive explanation where the contrast is at maximum width: e rather than not e. See Lipton 1991.) There is a simple reason for this fact. The limits to the causal depth of a non-contrastive explanation will show up in the form of possible situations in which the explainer occurs but the explained does not, or vice versa. But if the aim is simply to explain why e rather than e* occurred, a counterexample situation is one in which the explainer occurs but is accompanied by e* instead of e, or vice versa. Situations in which the explainer occurs and some other event, neither e nor e*, are irrelevant. So the causal depth of the contrastive explanation can look past these counterexamples to more remote ones. And comparing two contrastive explanations the same considerations show that the logically stronger contrast event is associated with the greater causal depth. (This argument asks for an explicit definition of 'c causally explains why e rather than e*'. See the appendix, where in fact 'the fact that c rather than c′ explains why e rather than e*' is defined.)

The argument of the previous paragraph may not seem to have much to do with matters of real explanatory force. But the slogan – narrower contrast means greater depth – is borne out in many real examples. Consider the marbles again. Suppose that there is a pattern of bumps and rough patches that guide a marble put in a particular starting position to a point at which it can roll to the one destination but cannot roll to a particular other destination. Compare the explanation 'it arrived at q because it started at p' to the explanation 'it arrived at q rather than r because it entered the zone where it was guided by the bumps and patches (etc)'. Suppose now that other marbles are thrown in at the same time. The connection between starting at p and ending at q is no longer secure – and as a result the first explanation doesn't give much understanding of what happened. It appeals to a sort of fluke, or miracle, that the pattern of collisions was just as it was. On the other hand the second explanation, the contrastive one, is much more secure. Whatever the pattern of collisions, that pattern of bumps and rough spots will guide any marble that enters it so that q is available and r is not.

It is the same with psychological cases. Someone is angry rather than delighted when hearing that her best friend won a scholarship. We explain this by pointing out that she too had applied for the scholarship, and is better qualified, but does not have any relatives on the awarding panel. This is a causally deep explanation in the sense that those facts make anger much easier to arrive at than delight. But the non-contrastive explanation 'she was angry because she was better qualified (etc.)' – if we take it absolutely literally, as a claim to explain why it was anger rather than any other reaction at all – is much less robust. She could as easily have been depressed, numb, confused, or any of a list of other reactions. (The factors that choose between these alternatives may be completely indescribable in common sense terms.) Partly because of the greater depth of contrastive explanations, explanation in everyday life is nearly always implicitly contrastive. And this is particularly so on psychological matters. We may say 'she was angry because . . .', but we nearly always mean 'she was angry rather than delighted/resigned/sad'.

We can expect everyday understanding in most domains to exploit the depth/width tradeoff. We can expect to find that common sense thinking finds ways of shaping the explanations it gives so that they focus on contrasts where causally deep explanations can be given using resources available from pre-scientific observation and concept-formation. (That's what it is reasonable to expect, apriori. There's no guarantee of finding it.) So we should look for systematic sources of causally deep explanatory contrasts. There is one well-known such source. It is the idea of an attractor, as it is used in the physics of non-linear systems. (See Schuster 1984, Crutchfield *et al.* 1986, Skarda and Freeman 1987, Gleick 1988, Morton 1989 and Rueger and Sharp 1996.) To see how it works consider another marbles set-up. I will describe it at some length because I think it is analogous to folk psychology in several respects.

This time the bumpy rough surface onto which the marbles are put has several depressions on it, with holes at the bottoms of the depressions. Marbles are rolled onto the surface. They are rolled with varying speeds and directions, with and without spin. Most end up going down a hole; a few descend some way downwards but have enough momentum that given the trajectory they take they arrive at the far edge of the depression and escape. An observer with a rough knowledge of a marble's initial conditions – how fast, what direction, whether spinning – will be able to make a rough prediction of its fate. The prediction will be very rough and will often be very inaccurate about the exact path that is followed. And when the prediction does happen to be accurate it will often embody a mistaken explanation. It will attribute the marble's path to its speed and direction when in fact the spin was essential. Such an observer will with somewhat more accuracy be able to predict whether the marble will end up in the hole or escape over the edge. The accompanying explanations will still be suspect, though. But there are accurate predictions to be made. The observer can very reliably predict that if the marble has entered a depression it will end up in the corresponding hole rather than anywhere else in the depression. And with somewhat less reliability the observer can predict that the marble will end up in the hole of the depression it is in rather than that of another depression.

Another image: similar at an abstract physical level. A hanging pendulum is given a push and left to swing. An observer with roughly accurate knowledge will have limited success predicting the periods and extents of the first few swings. But after those first few the pendulum will settle down into its natural frequency, which an observer can easily calculate. And after some time at that frequency the pendulum will become hard to predict again, before ending up, after another hard to determine interval, at the absolutely predictable destination of immobility.

In these and many similar cases we can predict that a physical system will eventually gravitate towards one of a set of states, its attractors. The route it takes towards them may be very hard, even impossible, to predict, but the same factors that lead to these varied and delicate routes also determine that their destinations are limited. Usually an attractor is associated with a range of earlier conditions which predispose the system to gravitate towards it, such as the depressions in the marble-rolling surface. Sometimes one has a pattern of transitions between attractors, such as the pendulum's going first to its natural frequency and then to rest.

These considerations have gone from contrast to depth. There are also connections in the opposite direction. Suppose that we can explain why a physical system ended up in a state having one property rather than another, and suppose moreover that this explanation is causally deep, in that the connection between initial and final properties would have held under a considerable variation in the surrounding circumstances. Then there will be a range of states 'near' the initial state, those that the system would occupy in such variations, from all of which it will evolve to a state with the one property rather than the other. So the set of states having this property will constitute an attractor. It is a destination that the system robustly tends to under a variety of circumstances.

Some rough generalizations emerge. Systems vary, as do the requirements on a good explanation and the kinds of interfering factors that create differences of causal depth. So neither of the principles below is a Metaphysical Law. But very often we can expect the following.

- An explanation that depends on the fact that a system will tend to one of a set of attractors, is usually causally deeper than an explanation that depends on the fact that from some initial state there is a route, given circumstances as they are, to some final state.
- When a contrastive explanation is causally deep it usually explains why a system tended to one attractor rather than another.

Human beings faced with the task of explaining and predicting the actions of other humans, without the aid of science, face problems that are very similar to those of the marble observer. The final positions of the marbles correspond to actions, the initial states of the marbles (positions, momenta, spins) correspond to the environmental influences, and the locations and shapes of bumps, rough spots, depressions and holes correspond to the states of the nervous system. At any given time several different action-directed processes are spinning their way through the system, colliding with one another. No ordinary observers would have a hope of predicting or understanding the exact trajectories of the marbles, so what chance would they have in predicting or understanding the output of the vastly more complex nervous system?

Marble observers can exploit the attractors of their system. These are identifiable inductively, straightforwardly describable, and lead to easily stated causally deep contrastive explanations. (Not all attractors of all systems will be easily identified or described, though as I argue in the appendix there is a sense in which all lead to causally deep explanations.) So what might the attractors of human action be? In the rest of this chapter I describe several.

Desire-attractors

We often understand what someone does not by deducing a prediction or an explanation from what we already know they want, but by knowing a destination towards which their desires and the pattern of their actions may gravitate. Consider an example. A person is setting up the living room in her new house. She needs a comfortable chair to sit in listening to music, a settee where she can sit with friends and watch videos, and a desk for her computer with a good chair. She sets these things up one way for a few days and then moves them around. Something is wrong about the chair and she tries putting a different cushion on it. Something looks wrong about the arrangement of the room and she varies the spacing of the furniture and replaces the curtains. Eventually, after a series of such re-arrangements the urge to change fades; she is now content.

The final disposition of the furniture satisfies our person. Why? She will be able to give part of the answer. The chair is now comfortable, the colors of the settee no longer clash with those of the curtains, light from the window does not now make reflections on the computer screen. Why is this particular balance between comfort, attractiveness and practicality the right one? She probably cannot say. Yet it probably is pretty nearly optimal for her purposes; if it is not she will find herself wanting to change some features. And there inevitably are other factors of which she is not aware. The low rumble from a distant road may disturb her work more if the desk is in one place than if it is in another. The beginnings of sciatic pain of which she is not yet aware may influence the postures in which she will comfortably sit. The settee may be a more friendly place if it is neither too near nor too far from the radiator. These factors too have to enter the equation. So if pushed beyond a point about why she finds the room satisfactory she is likely to confabulate, to make up answers that pass by the real reasons why she has arranged it well.

This is an example of a middle-sized process. The person adjusted her moment to moment preferences over a few days or weeks in order to satisfy deeper needs. The same happens in the course of a few minutes as a person settles into a comfortable sleeping posture or a promising conversational strategy. And it occurs in the course of years as a person finds the right partner in life or the right sources of emotional contentment. Folk psychology can understand none of these things in terms of a fixed list of desires attributed to a person, for they concern the ways desires *change*. One reason why they change in the course of very ordinary practical actions over a short space of time is that as we act we discover ways in which the goods we are aiming at can be better achieved, and discover other goods, which we had not anticipated. As this happens desires for situations and actions that get us nearer to these goods are generated. People can never state explicitly all the goods to which their actions aim, and even for those goods which they can to some degree describe they can very rarely articulate the subtle means by which they balance and trade off one good against another. This homing in on states in which desire and situation are in equilibrium occurs in every little practical activity as much as in large life-shaping behavior. (The position described is thus a naturalistic moral realism. For a survey of moral realisms see Little 1994. Though moral realism has few terrors for us now, a lot of credit must go to writers such as Platts 1991, who first overcame the inhibition about talking of moral facts.)

A rough generalization about our folk psychology thus emerges. We expect that people's actions will very often result in situations which are stable and satisfactory in ways that those people can rarely describe clearly. People find their way to such situations because of systematic ways in which their beliefs and desires change with respect to them. Moreover, among the changes in our desires there is a tendency for them to gravitate towards satisfaction of deep human needs. We both wobble between different needs, and follow trajectories homing in on them. This is no accident, as Peter Railton has pointed out:

> Humans are creatures motivated primarily by wants rather than by instincts. If such creatures were unable through experience to conform their wants at all closely to their essential interests . . . we could not expect long or fruitful futures for them . . . Since creatures as sophisticated and complex as humans have evolved through encounters with a variety of environments, and indeed have made it their habit to modify their environments, we should expect considerable flexibility in our capacity through experience to adapt our wants to our interests.
>
> Railton 1986, p. 181

It is a basic fact about us that our desires change, and do so in systematic ways that meet our needs. And it is almost inevitable that our folk psychologies have ways of taking account of this. I believe that we have here an aspect of everyday explanation that is as pervasive as explanation by reasoning from motives. Call it *explanation by attractor*. I shall use this term to cover a variety of explanatory patterns. All that is required is that they cite some characteristic of an agent and some class of outcomes which can be described independently of their relations to agents, and then explain the occurrence of the outcome by the tendency of agents with that characteristic to produce outcomes in that class. There are trivial and not very informative explanations satisfying this description, as when we say that someone broke his leg because he is accident-prone on holiday, always returning with some injury. My main target, though, is explanations where the outcome came about because of motivated actions by the agent. The explanation, though, rather than describing specific ways in which existing desires lead to particular actions appeals to the kinds of desires agents like the one in question form and the kinds of results their desires tend to produce. The explanation of the person arranging her room is a typical example. There are also many less elaborate examples. For example some appeal to, rather than being puzzled by, weakness of will. We explain why someone took a chocolate off a table and ate it, though he was trying to lose weight and though he knew the chocolate was meant for someone else. We say 'there it was, and it looked so enticing, and he had been so good for weeks, so it isn't surprising that he found himself wanting it, and finding a way of getting it into his mouth without being noticed.' Note that the explanation not only implicitly appeals to a desire-production mechanism, but also appeals to ways in which the desire produced fits into and takes a priority among the whole complex of the agent's existing desires. It does not describe how this happens; it just asserts that in such circumstances then somehow or other it does happen.

The focus of explanation by attractor is sometimes, as in this example, the particular intentional actions that are the objects of explanation by reasoning (and of standard philosophical interest.) But it can also be smaller or larger. Smaller when it concerns the sub-intentional component actions, such as a musician's hitting a particular note in an improvisation or a person using a particular word in a persuasive speech. Larger when it concerns the general tendency of a sequence of actions, such as the overall aesthetic effect of the improvisation or the persuasive effect of the speech. The attractive outcomes here are most often not even conceptualized by the person, let alone deliberately aimed at. (For more on the contrast between larger and smaller see 'Attractors and goods' below.)

Three aspects of explanation by attractor are particularly important. I shall briefly mention them here, and both will get more attention later in the chapter. The first is the use such explanation makes of the fact that different people are acute in different ways. This was illustrated in Chapter 2 with stories about geniuses and practically handicapped colleagues. In those stories explanations were implicit, which recreated the connections between the individual's skills or failings and the kinds of outcome to which those skills or failings naturally tend. We use explanations like this wherever people's capacities vary. Differences between people's ability to think through social, spatial, arithmetic, or moral problems can explain why they get different results when they face such problems. We explain actions by relating their authors to characteristic outcomes.

The second important aspect of explanation by attractor is that the attractors, the goods to which people's actions are related, can often not be described in neutral terms, which would be available to someone who had no concept of them as the particular goods they are. A just person's actions will tend towards just outcomes. But to know what these are you have to have some idea what justice is. A person inclined to physical comfort will tend towards comfortable situations. But to know what these are you cannot just operate with a list of the person's present sources of comfort and discomfort. Their dispositions will change, and a balance between the resulting goods and bads will be struck, a balance which you will have trouble grasping unless you can replicate the attraction to that particular stable combination of comforts. In general, explanation by attractor depends on the way the desires (and related states) of the person explained *change*. And while there are some commonsense resources for explaining changes of desire in terms of general principles, the default method is to share in fact or imagination an attraction to the object of attraction. If you were indifferent to comfort no list of the desires of the room-arranging person above would make the drift of her actions intelligible to you. If you are indifferent to justice, no list of the principles of justice that a person subscribes to will allow you to anticipate how she may change those principles as she reacts to details and arguments about a particular hard case.

And the third aspect is fundamental contrastivity. The explanation says why the agent gravitated towards this outcome rather than that; it does not say why the agent adopted this means rather than that to attain the outcome. If the explanation of the person arranging her room is correct it says why she made the room comfortable rather than elegant, uncomfortable, or some other general way it could have been and some other person might have made it. But it does not say why she made it comfortable in this way rather than that. It does not say why she chose between different possible ways of making it comfortable. Perhaps a chair could have gone where the sofa has gone, and vice versa, just as well, and perhaps there is a good explanation of it, either via a finer-grained explanation by attractor or in terms of means to a desired end. But this explanation does not give it. So the claim that the explanation explains *what she does* has to be taken in just the right way. The explanation says why one generally described result rather than another came about, as a result of what she did. But she could equally well have done many things.

Strategic attractors

Consider a paradigmatic situation from the introduction and Chapter 1. You are about to enter a zone of lawless traffic. You cannot make an absolute prediction but you can make one conditional on your own behavior. If you drive in an orderly, predictable way then others will be able to avoid you and in order to do so they will themselves be (somewhat) predictable. Now suppose that several people are entering the maelstrom, adopting this tactic, and that the situation is progressively calming. Choose an arbitrary driver, who begins to drive in an orderly way. Why is he doing so? It might be because he has reflected on the chaos around him and adopted predictability as a tactic. It might be because others around him are adopting the tactic and this is forcing order on him. It might be some delicate mixture of the two. The details of his behavior may reveal which reason is operative, but they will be very hard to discern. It may in fact be unclear what it would take to make one reason rather than the other be the better explanation of his actions. And even if his action is the result of one rather than the other it would have taken only a very slight change in the situation for the other to have been operative. It will be very hard to predict when he will begin to act as if the majority of those around him were no longer road maniacs. What is clear is that at some point he is very likely to adopt the new pattern. It is an attractor.

Or, consider the coordination situations that have often featured earlier. Several people must, without communicating, do the same thing. Suppose there is a solution, in the form of an act which if chosen by all will give good results to all, and which is distinguished by some interesting feature from other such acts. We can expect the people to choose it. But we can't tell how long they will deliberate, or what their individual motives will be. Some will choose it because they will expect others to, others because they will expect others to expect them to, others because they just think it is what a wise advisor would suggest. This contrast becomes even greater when the solution involves a tradeoff. Suppose that it is better for almost all, but at a serious price for a few, but there is no alternative which does not either impose a comparable price on more, or have no marks to distinguish it from similar solutions. Then again all will eventually come to choose it, even those who must pay the price since the price is less than that of not choosing as others do. But any guess about why any particular person made their choice is fraught with uncertainty. And any person's reasons could easily have been those of another.

The most important point is not about certainty but about explanation. Since we can very easily be wrong about the causal relevance of factors which could rationalize one individual's doing one act rather than another, it is easy to give explanations that do not actually say why someone did something. On the other hand we do know why it is that all the would-be coordinators converge on the chosen outcome. It is because it has features that attract motivation in a variety of ways, including no doubt some that do not feature in easy rationalizations. The fact that the coordinators are in a situation that has a potential outcome with these features explains, causally, why the co-ordinators, individually and collectively, act in such a way as to bring it about. (I have chosen the wording here so as to avoid appearance of backwards causation or illicit teleology. This should be no more teleological than an explanation of a physical change in terms of the fact that the entropy was lower in the final state.)

I will return to issues about causation and causal depth in the next section. First I must try to describe the general pattern behind the two examples. Call an outcome O of a particular concrete strategic situation involving a number of particular individual human agents *stable* when O results from similar action-choosing processes in the agents, such that they will all tend to expect O to emerge, and tend to expect one another to expect it to emerge, and which most of them will accept, given that it results from the expected choices of the others. ('will accept' means 'those same action-choosing processes, when given the additional information that O has occurred or is about to occur, will not revise their choices in a way that will undo O'. When the number of agents is small, for 'most' read 'all'. Compare the definition to that of a solution in Chapter 1. That definition was put in terms of congruent choice processes, while this one is in terms of their results.) This is a rough version of the game theoretical idea of an equilibrium. (It takes the fine smooth surfaces and deliberately files and batters them. More specifically, it appeals to whatever decision processes happen to be in use, rather than 'rationality'.) But it diverges from the official equilibrium concept in some cases, notably those in which although some participant would have a motive for deviating, all involved know that that person will not act on that motive. Examples are those prisoner's dilemmas where as a matter of human fact almost everyone cooperates. These examples show the importance of the fact that stability is defined for particular token situations involving particular real people while equilibrium is an attribute of outcomes of abstract types of situation whose agents are defined only by the specified preferences.

A stable outcome has some of the features of moral acceptability. In many situations if agents were to negotiate among themselves they would decide on a stable outcome. If agents are equally well endowed and are prevented from making threats, then the outcomes of negotiation are very likely to be stable. But of course agents usually vary in their bargaining powers and usually can make threats. But, still, when there is a common good it will be a stable outcome. And when there is a reasonable and sustainable compromise between competing interests that too will be a stable outcome. Or, rather, the outcomes in these cases will generally tend to have properties rather like what I have called stability. For the point is not that stable outcomes are always better ones. Rather, *one* of the components of a good outcome – the reconciliation of competing interests – tends to result in stable outcomes. For participants in a stable outcome accept it as an outcome; given that it has occurred they are not motivated to change it. Some stable outcomes certainly have very little to do with the good, notably those shaped by implicit threats and some (not all) prisoner's dilemma-like situations in which all concerned can foresee that they will be forced to choose in a way that all will find unsatisfactory. Even in these cases 'stability' labels attributes that explain the occurrence of the outcome. The outcome happens because it is the kind of outcome that the agents, whether they think their options out carefully or trust to their social instincts, will tend to find their way to.

Contrastivity is essential here too. We know why each driver's actions tended towards the eventual orderliness, but we do not know why any particular driver began driving more carefully than others around him (say). We can explain why the result was order rather than chaos, in fact we can explain why it is some kinds of order rather than some other kinds. But we cannot explain why the transition to order happened in exactly the way it did. (I would add the intuition that this may well be just inexplicable, in anything like common-sense terms.) This generates another difference between the everyday approach to strategic situations and game theory. In game theory an outcome is defined as a combination of specific actions of specific individuals, and the explanation of the outcome thus explains why each agent did exactly what he did. Not so in everyday thinking. What we explain is why the outcome that is the causal consequence of the choices of those involved made, has a certain general character, which could usually have come about in many different ways from many different combinations of fine-grained actions.

Causal depth

Assume that there are desire attractors for one-person choice and strategic attractors for interacting agents. What advantages do explanations that appeal to them have?

I think it is not hard to see that an explanation that appeals to the fact that a person is in the range of a desire attractor will very often be causally deep. (Deeper than one that simply appeals to the fact that the agent wanted an outcome and understood a way to attain it.) The reason is built into the definition of an attractor. Put crudely: states of mind that are similar to the person's actual state will also tend towards outcomes in the attractor. So had the situation been somewhat different the person would still have ended up in the attractor. And that is causal depth. More carefully put, had the person's state been slightly different in a way that had left it in the range of that attractor rather than another, the result would have been an outcome in that attractor rather than some other. (For an even more careful formulation, see the appendix to this chapter.)

An example may bring out the kinds of explanatory robustness in question, and also the non-mysteriousness of the concept of desirability involved. A person is deciding how much she is prepared to contribute to her child's college expenses. The decision is how to divide the expenses between herself, her ex-husband, and the child. She decides that she will pay for half of the child's board and tuition, and nothing more. The reason she gives is 'At that age one should be becoming independent of one's parents, and I have other expenses, for example a much-needed vacation'. The person also wants her ex to pay now to compensate for lack of support for an older child, and feels less close to this child than to the other child, but does not cite these reasons and does not think of them as relevant. She may be right; it may be for these motives that she acted. It is hard for her to know, though: it is always even harder for people to know what the causal connections are between their motives than it is to know what it is they want and what they think. (I take this rather than the more general fallibility of self-knowledge to be the conclusion that emerges from the classic work in the 1970s by such people as Nisbett and Ross, summed up in Nisbett and Ross 1980.) And the reason it is hard to know these causal connections is that they could easily have been different. Had her mood been just slightly different, or had she been in a slightly different physiological state, she might have done the same action, not because of the vacation and the child's independence but because of her resentment of her ex and the limits of her maternal affection. (There are lots of motives hanging around in any person's mind; at any moment many of them might volunteer to be the sponsors of a possible act.)

So, though the explanation 'because she needed money for a vacation, and she thought the child should be partially self-supporting' may be correct, it is unlikely to be deep. It rests on factors that could easily have been otherwise. On the other hand, the explanation 'because that level of expenditure strikes a stable balance between her concern for her child and her self-interest', or some idiomatic variation on it, is more robust. If her choice had been caused by some other combination of the available motives it would still have been explained by the way it balanced the competing motives. It would still have been true that had that balance point not been such that on considering it she was happy with it she would not have chosen it. And this brings out the way in which the outcome is desirable. It satisfies not the combination and balance between desires with which her deliberation started but the combination and balance that deliberation brought her to, and which would as a result of that deliberation remain as they are under some degree of perturbation.

Similar points can be made for other attractors. Suppose the choice she made was such that she could think of it as a fair reconciliation of her interests and those of her ex and her child. And suppose that fairness was in fact a factor. Then even if the conditions which allowed the operation of the actually effective combination of particular beliefs and desires had been different, she still could have been led by her propensity to choose the fair solution to a solution that exhibited some degree of fairness. The choice was not inevitable; a fair-minded person sometimes wobbles and makes a transparently unfair choice. But if the result was fair and the fact that it satisfies some criteria of fairness was evident to an agent with the virtue of fairness, then the facts explain the result in a way that would hold under a considerable variation of background detail. Again the causal stability of the outcome is linked to its being good for the agent. She faced a problem reconciling different commitments, a problem that is eventually solved. Had the outcome not been stable, it would not have been a solution.

Very similar observations apply to outcomes of strategic choice. The main point was made above when I introduced the idea of a stable outcome, and has been argued in various ways since Chapter 1. It is of the nature of the outcomes of strategic situations that it is very hard to capture the process that leads a particular person to play their part in a satisfactory outcome in a way that is at once psychologically plausible and rationally compelling. People find their way to coordinative, cooperative, and otherwise stable outcomes in ways that vary greatly in their detailed motivational structure. But all the ways have in common that they are recognitions of the fact that the outcome is one that once achieved no one would want to un-achieve, given that the choices of the others were as they were. But this is just to say that whatever the motives and other states of mind that produce one participant's action are, they could have been different in various ways without preventing that person from finding their way to their part in the outcome. And that is to say that the explanation linking the person to the stability of the outcome is causally deep.

Attractors and goods

An explanation gains explanatory depth when it can appeal to attractors. And attractors are desirable, for the most part. (I mean, they are capable of being wanted.) And desirable things are good for those who get them, for the most part. It might even be taken as definitive of what is good for a person that it is something to which their desires tend: something which is stable under reflection about what one wants and which is not undermined by the experience of getting it. And as remarked above, stable outcomes of strategic situations are closely related to potential results of fair negotiation. So the argument connects in a vague general way with another idea, that often we explain by linking an agent with something good. The purpose of this section is to explore that connection, partly because I am sure that there is something to it, and partly to separate out that claim from the other ideas in this chapter and make clear its independence from them.

It could not possibly be true that *any* characteristic that makes an outcome desirable contributes to the explanatory depth of explanations that appeal to it. What is more defensible is the suggestion that there are some characteristics of some kinds of desirable outcomes such that the more an outcome has them, the causally deeper are explanations that appeal to them. (I might add, though, that I do expect that the phenomenon is a lot more general. I expect that a variety of explanatory virtues are associated with connections between agents and the value of the things they are acting towards.)

An objection would be natural at this point. The attractiveness of the attractor and the causal depth of the explanation have nothing to do with the outcome being desirable, in the sense of *good*. Heroin addiction is an attractor, and stable: someone in the range of the attractor has a good chance of becoming subject to it, and once there the result can be very stable. Certainly more stable than the mild glow of a good deed well done. The coordinations of criminals are stable outcomes of their criminally strategic situations, and it is in their mutual interest to adhere to them. What people want is often bad; that people don't want to undo something once it is done is no sign that they should have done it in the first place.

The objection misunderstands. It confuses the possession of important features of a good, which attractors generally have, with being a good, all things considered, which a fair number of them are not. Suppose that a situation is in someone's interest, and they understand its nature. Is it not reasonable to suppose that they will find some way of achieving it? Once achieved, will there not be a strong pressure to keep it? Is this not the difference between needs and whims or obsessions? Is not a large part of the reason that peace, giving and receiving affection, accomplishment, and the respect of others are basic goods, that when we see them clearly we want them, and when we have them we do not want to give them up?

Or to put it differently: suppose we were to find that an outcome is such that people will under no conditions individually or collectively strive to achieve it, or that once they do achieve it they wish they had not. We would find it impossible to regard it as a good. The assumption here is a mild internalism: if something is valuable then there are situations in which it will be wanted, and closer acquaintance with it, most importantly the acquaintance that comes with actually achieving it, increases the motivation towards it. As internalisms go, this is fairly mild. It requires only that there be situations in which the good is wanted, and requires only that for each particular good this be true as a matter of psychological fact. (But if it were not a psychological fact that item would not be a good for human beings.) It is consistent with some positions labeled externalist. For it allows that for any good there are likely to be situations in which some person does not have any inclination to achieve the good, and allows the possibility that neglect or failure to maintain a grasp of some good things sometimes can be perfectly natural. (For the flexibility available here see the appendix to Scanlon 1998 and Svavarsdottir 1999.)

This defense against the objection runs the danger of mis-stating the claim. For it makes it seem as if attractor-hood is just one of many necessary conditions for an outcome being desirable. But it is surely a particularly important necessary condition. For it captures several central characteristics of what is good. I'll list five.

1 *independence of present desire:* what is good for a person or what is a good outcome for several interacting people is determined more by the desires that they would develop and retain while deliberating and reflecting than it is by their actual present wants.
2 *stability:* what is good should keep its desirability once attained.
3 *sharearability:* what is good for one person may not be good for another. But it should be intelligible to one person why the actions of another tend towards a good.
4 *discoverability:* many goods may be described with familiar labels – tranquility, love, cooperation. But many are subtle hard-to-describe attributes which we half-blindly find our ways towards. Some very fundamental goods in human life may have no useful names.
5 *resolution:* people's desires are rarely consistent, and even when consistent they include incomparable desires whose priority cannot be settled in any mechanical procedure. A resolution of the conflict of desires or a compromise between incomparable desires that gives a balance with which the person can operate, is an important complex good.

Attractors have all of these central characteristics. The fact that something is an attractor, that it strikes an equilibrium between the desire-potentialities of a single person or the preferences of a number of people, goes a long way to establishing that it is a good for that person or those people. It goes far enough that we can sensibly ask what prevents all attractors from being goods. The answer, it seems to me, is that any attractor would on its own be a good, were it not for the presence of other perhaps more important ones. Thus comfort is a *prima facie* good, but there certainly are times when it is over-ruled by the importance of self-preservation or that of alertness. (Heroin is a good, but getting it interferes with greater goods. And as Peter Railton points out to me it is lacking in stability, in that the addiction requires more and more of the drug as time goes on.) Indeed it is possible that there is a fundamental good of human life, of which we are at best dimly aware, consideration of which would undermine the status of almost all the qualities we take to be valuable. Similarly with strategic attractors. If there is an outcome that affords mutually profitable cooperation for a number of interacting people then it is a *prima facie* candidate for being a good for them. But of course it may not be part of a larger satisfactory outcome for a larger group of people. It may block such an outcome. (The harmony of the bandits is a disaster for everyone else. A more interesting case is a cartel preventing a market from operating in a fully competitive way.) And there are conflicts between desire attractors and strategic attractors. A resolution of the present preferences of a set of interacting people may put some of them in a situation in which their desires change so as to make the result unstable. A situation that develops and resolves the desires of a single agent satisfactorily may block developments that are in the interest of others.

The obvious way of summing this up is to say that individual and strategic attractors are *prima facie* goods. They would be goods if not prevented by their position in a structure of other *prima facie* goods. There is also a less obvious conclusion to draw. It is possible that when we adopt a really critical attitude we may conclude that there are no ultimate goods. Perhaps for every *prima facie* individual good there is another that over-rules it. Perhaps there are no satisfactory solutions for the competing preferences of large numbers of people (however we arrange it, someone is sleeping in the street.) Perhaps there are irresolvable conflicts between what solves the problems of individual people and what solves those of groups (so there is always an uncooperative option that would really be better for any one person.) Or perhaps, though conflicts are partially resolvable for every partial resolution there is a better one, so that the attractors of individual and communal life form a structure of local and more global goods, with no ultimate global goods. Perhaps, and perhaps not: the answers here cannot be given apriori (see Morton 1996b.) Usually when we treat something as a good all we need to be sure of is that it is a *prima facie* candidate that is not over-ruled or inconsistent with any other candidate that seems to be relevant in the context at hand. It doesn't often matter too much whether there are ultimately stable and fundamental goods. (Though it does matter for everyday psychological explanation, whether one *prima facie* good is ruled out by the presence of another.) So, though nothing I say will depend on it, I would be quite happy to say that all attractors are good, in the sense in which all elephants are big, even though some are bigger.

The size of actions

The central purpose of this chapter is to work out one aspect of everyday psychological explanation, the trade-off between causal depth and explanatory width, in the special case of explanation by attractor. A subsidiary question, though, is how much of our everyday explaining we could do – or could best do – by appealing to attractors. This section gets a small grip on that question, along the way tidying up a couple of other outstanding points. And then in the following, and last, section I return to the main line of argument.

It would be foolish to attempt to explain all actions by appealing only to underlying attractors. Any such scheme of explanation would be pretty feeble. There are kinds of actions, though, which seem to call for explanation-by-attractor. These can be thought of as 'larger' and 'smaller' than actions to which 'Humean' explanation, by appeal to the best ways of satisfying an agent's current desires, given their beliefs, naturally applies.

The larger actions are long-term results or accomplishments. Consider how we might explain why a couple have managed to stay together through a variety of troubles, or why a person's life work amounts to what it does. We often cannot adduce a desire to accomplish that particular end, for example the specific balance between conflicting inclinations and temperamental differences or the renunciation of some ambitions and the concentration on others. People do not usually have such desires, or beliefs about means to such ends. They have very vague and general aims in life and then much more specific desires governing their day to day activities. And when they do have goals of this breadth and specificity they rarely accomplish them.

And yet there is often a lot to say about why a couple achieved the equilibrium they did, or why a person's life had the overall results that it did. The explanation links together the many small component choices that, together with unanticipated facts and accidents of fortune, added up to the result. It does so by describing the people, what they are like and how they developed, in a particular way. We might say of the couple that they were tolerant, hopeful and un-hasty. Each of these attributes links to some aspect of the balance they struck and maintained, or to the route towards it. Each attribute can be used in two ways. First, simply to connect the people and the attractors. We say of the couple that they are together after all these years and all these troubles because they could always keep a larger perspective than the issue of the moment. Second, to explain the beliefs and desires that lie behind some of the actions that led to the eventual result. One member of the couple might have left a well-paying job to allow the other to further her career. We could explain this by saying that he wanted her not to miss a rare opportunity and felt sure that something equivalent for him would open up soon. These desires, and the role they played in his deliberation, are then explained by appealing to patience and optimism.

Other large actions are smaller than these. Consider a group of jazz musicians improvising. Their playing has a result of a particular character. They didn't intend to get that particular result and thus didn't want it. But it wasn't an accident either. Far from it, easily nameable qualities of the musicians are responsible for very general characteristics of the result (without determining it), and much more subtle qualities of individual musicianship and social temperament are responsible for much more subtle qualities of the music. Some of these more subtle personal characteristics cannot be described except by reference to their typical results: this is an effect he often produces; when these people get together something like this can often result.

Long term outcomes connect with the outcome of debate about the meanings of words and the shapes of concepts. Imagine a child who wants to give her little brother a nice present for his birthday and is shopping with her mother to find one. We can predict that she will buy a present that is nice not in terms of her own first reactions to candidates but in terms of something nearer to the adult conception of niceness. In the longer term we can predict that her presents over the next few years will gravitate towards what would be characterized as nice among more mature members of the community. In the yet longer term we can predict that she and other members of the community may choose presents and other such things that tend towards some real complexes of properties which could sometimes be summed up as ‘nice’: tastefulness, attractiveness and sensitivity to the intended recipient.

There is an important general pattern here. There is a connection between the long-term destinations of thoughts and the dependence of their contents on factors external to the thinker. As Timothy Williamson puts the point ‘The need to think about connections between earlier mental states and later actions is largely a need to think about connections between mental states and actions separated by seconds, days, or years . . . How and whether one puts [these states] into effect depends on one’s interaction with the environment in the intervening period . . . Extended actions involve complex interaction with the environment.’ (Williamson 2000, pp. 75–6, but see also Chapter 2 ‘Heal’s problem’ and Chapter 3 ‘What the stories show’ of this book.). Moreover (and as Williamson stresses), the contents of many of our motives are thoughts that are *externally based*, in the sense that what an individual person’s belief is about depends on that person’s actual environment. (The terminology comes from the tradition stemming from Burge 1979, for the current state of the tradition see Bach 1998.) By consulting with experts, responding to scientific discoveries, evaluating evidence, and so on (varying according to the kind of concept, the variety of externalism) the person’s acts will focus on some actual objects of the concepts rather than any objects which they fit. (In the simplest case, as discussed by many authors, an action is guided to the *de re* object of a belief or desire by its external relation to it rather than by the internal content of any beliefs.)

In an article arguing that a generally externalist attitude is appropriate to a very wide range of concepts, Tyler Burge argues that the content of many beliefs depends on the actual defining characteristics of things to which competent speakers apply a word. A particular speaker may be ignorant of these characteristics, although her beliefs are nevertheless about those things. As a result, there can be a debate between users of the concept about the meaning of the relevant word and, simultaneously, about the true nature of the objects to which it applies. As Burge (1986) writes:

> These disputes usually concern two matters at once. One is how correctly to characterize the relevant entities: whether all chairs have legs or must have legs. The other is how to state the meaning of the terms as such: whether according to (by) definition chairs have legs . . . what is important is that . . . the second is typically pursued by trying to answer the first. Questions of meaning are pursued by attempting to arrive at factually correct characterizations of empirically accessible entities, the examples . . .This point rests on two

> features of the dialectic . . . One is the central role that *examples* play in arriving at meaning-giving characterizations for ordinary terms . . . The dialectic aims at capturing archetypical applications in a way that sharpens explications that competent speakers naturally give . . . The second notable feature of the dialectic is that, as the participants work toward an expression of communal meaning, they typically do not discuss the matter as outsiders. Usually, all participants begin the discussion without being able to give a precisely correct normative characterization . . . But this does not entail that any lack object-level thoughts expressible, by the rest of us, with the term whose meaning is in question.
>
> Burge 1986, p. 705

Burge goes on to argue (p. 714) that as a consequence

> Moreover, understanding the meaning-giving characterization for a term does not necessitate acknowledging it as true. It is sufficient to be able to apply the term correctly and to give approximations to standard characterizations that may be adjusted under reflection, criticism, or new information . . . The consensus of the most competent speakers can be challenged. Usually such challenges stem from incomplete understanding. But . . . they need not. One may always ask whether the most competent speakers' characterizations of examples . . . are *correct*, and whether all examples usually counted archetypical *should* be so counted.

So two things go together here. There is the focus of an agent's behavior on the real features of the world that actually satisfy her states, and there is the way an agent's grasp of her belief leads her to interaction with other people and with the objective features themselves. Reference to things and entanglement with others go together. Even for scientific and natural kind concepts this can make obstacles to simple explanation simply in terms of the content of a person's belief and desires as it is for that person at a time. In order to explain what a person who collects chairs will do, you have to either have perfect knowledge of what chairs are (as no one does) or understand the processes by which people consult with one another and with evidence to determine true-er extensions of 'chair'. But we rarely have explicit knowledge of these processes: we just know how to take part in them, which also provides us with our only way of knowing what cats are. So in order to predict the behavior – all the potential behavior – of chair collectors you have to know what will happen when they consider evidence and consult with others. The obvious route to knowing this is being able to do it yourself.

The equilibriation on the outcome determined by the facts and by the externalist content of an agent's thoughts can take a long time. Among the longest times are those associated with a moral content. If a person wants to do what is fair, in a certain situation, she is likely to do what seems fair to her given the time she has to reflect. If she and another person have an ongoing disagreement about the fair resolution of a certain issue then there is a pressure – not more – towards an outcome that accommodates both their deeper intuitions to what is fair, and some of those of other people whose influence may come to bear. If a whole society is faced with a complex of issues about fairness, there is some tendency – obviously not more – towards judgments and actions that are somewhere further on the way towards some ideal balance between intuition and general belief, that is towards what is actually fair. (If such an ideal balance exists. The allusion is obviously to Rawls 1971. See also Daniels 1996.)

There is thus a rough and general connection between the size of an action, measured by the amount of time that it takes for an initial motive to determine a result, and the degree to which the content of the motives is given by externalist factors. A long term prediction of an action governed by motives with a strong externalist component will have to take into account the influence of this component. And often the best way to do this is to figure out the right answer to the question the agent's action poses.

At the other extreme from the large actions are the small ones. Consider the jazz musicians again. A bass player hits a particular B-flat at a particular moment with a particular finger. He didn't plan it, in fact he may not later know he did it, but this action too was not an accident. Not a slip or a fluke or something that just happened. Hitting that note with that finger at that time allowed a sequence of chords to happen, and that sequence at that point was essential to the group's improvisation having that electric-relaxed quality that it did. When we explain why he hit that note with that finger then we just say: it got him from this chord to that one and he was aiming at that dark glowing sound to complement the trumpet's shining purple. There are innumerable examples like this, of things people do that are not properly described as intentional but which are far from accidental, which we explain by saying what effects they produce in the presence of an intention to produce a larger and less specific result. Often, though not always, they are constituent parts of outcomes that are intentionally produced. Very, very often – my main point – they are explainable in terms of their relations to characteristic attractors.

We use different language when we explain large and small actions. When explaining a long term tendency we might say 'what she was after was peace (or companionship, or distraction from her thoughts, or . . .).' Or 'these things give him a way of influencing younger colleagues (or a way of pretending he is important, or . . .).' When explaining the result of a sub-intentional action we tend more to embed the attractor in the accomplishment. We might say 'by hitting that note he managed to create that special effect' or 'by phrasing the request in just that way she made it hard for him to point out her hypocrisy.' In both cases we tend not to describe the attractor very explicitly. We describe something less specific (companionship or hypocrisy, when a very particular kind of companionship or hypocrisy is what matters) or we refer to a particular instance of something more general. We do this because we usually are not capable of describing the exact quality or result that that particular person's actions were aiming for. Sometimes we are able to fill in something more specific from what we know of the situation or of the person. Sometimes the premise is implicitly existential: the target is some quality or result which is suitably related to this more general feature or this special case.

Shared contrasts

We need explanation by attractor in everyday life. We use it in explaining change of desire, the long-term results of actions, the effects of little improvised acts, and the choice of actions in strategic situations. Not that it is our only resource for any of these. But its suitability for all of them should be clear by now. And the case should have been made that it can provide explanations of all these kinds that are causally deep, or at any rate causally deeper than the alternatives available by considering the beliefs and desires that a person happens to have and which happen to have played a role in producing her action.

Since we think in terms of attractors, we think in terms of contrastivity. (And vice versa; the two go hand in hand, in ways explored in 'Causal depth and contrastivity' above.) When we say why someone accomplished something we are almost always explaining why they did one thing rather than another. We rarely state these contrasts very explicitly. We don't have to, since we know implicitly what they are. They are in fact generally the contrasts in terms of which we act, choosing one option *rather than* any of a list of alternatives (and not rather than other alternatives either not considered or not discriminated from those considered.) If the contrasts that shape our actions were not closely related to those that we find in our understanding of the actions of others, we would find our attempts at shared activity enormously encumbered. Indeed even in antagonistic interaction it is important to understand the space of alternatives within which the other is choosing, and then to adapt one's own pattern of choice to it.

It is worth at this point asking the direct question of how the ubiquity of contrastive explanation and its link to attractors of individual and joint action bears out the theme of understanding that succeeds because it is shared. Do we here have examples of understanding because one is understood? The answer is Yes in the following respect. The basic fact is the incomplete nature of any set of contrasts and attractors that enter into any form of human life. We never engage with more than a selection of the goods that we could organize our lives around, and those that we do engage with are always relatively superficial compared to others that we could conceive of. The labels we give the axes of choice and explanation – pleasure, comfort, esteem, peace, justice and safety – point to deep and universal goods while being employed in practice to designate parochial limited versions of them. (As any Buddhist can tell you, there are forms of calm which, obvious when noticed, require remarkably special conditions in order to be noticed; the idea of harmonious relations between people in which it is not settled who gives orders to whom is absent from most human calculations.) So in acting and explaining we draw on a limited range, nearly always essentially the same limited range that those we are explaining use in their actions and their explanations. And the explanations would not work otherwise.

This acknowledgement of a limited range of partial goods can be achieved intellectually, or imaginatively, but most often it is achieved simply by acting, choosing from within menus whose structure makes salient the contrast-space in terms of which the particular people you are understanding, members of a common culture, act. They choose from options structured in these ways in part because they are attracted to the corresponding goods, in part because they are engaged in strategic interaction with others whose thinking is shaped by these contrasts, and in part because they use these contrasts in explaining the actions of others. These factors are inseparable and reinforce one another: cooperation with others will not be sustainable unless it bears some relation to basic human goods; when people who will interact learn their way around the same contrast space their familiarity with it pays off both in coordinated action and in understanding. The actions one tries to understand have in these ways been shaped by the fact that their agents are also engaged in a shared process of understanding.

Much of this has been put very concisely by Robert Graves.

> Sigh then, or frown, but leave (as in despair)
> Motive and end and moral in the air;
> Nice contradiction between fact and fact
> Will make the whole read human and exact.

APPENDIX: ATTRACTORS AND CONTRASTS

The purpose of this appendix is to give a rigorous presentation of some of the connections between the ideas of contrastive explanation, causal depth and attractors used in Chapter 4. The appendix states and proves some results relating these ideas. To do that, of course, it needs some definitions. They are not the only conceivable definitions. But they are intuitively on the right track, and I add some informal explanation to help make this visible. I expect that many variant definitions will lead to similar results. One reason the definitions I am using may be of interest is that they are entirely non-quantitative.

They include a qualitative formulation of the idea of an attractor in a dynamical system. (When writing this appendix I have seen closely related unpublished work by Laurie Paul, Christopher Hitchcock and Cei Maslen. Particularly close to my themes are Hitchcock 1996 and Paul 2000.)

Part 1: causal links

I assume events c, c', c'', e, e', e'' etc. some of which are caused by others. I shall treat events like propositions and consider conjunctions and conditionals of them. Thus 'if c then e' means 'if c occurs then e occurs'. The causal connections I am concerned with are unreliable and delicate, as is appropriate when the subject matter is psychological explanation. The image is one of unknown and hard to describe additional factors, in the presence of which the causes can produce their effects. (So c may be my desire to see a film and e may be my going out into a blizzard to see it. Had things been slightly different in my brain, then though I had the same desire it might not have produced the same effect.) I shall thus usually talk of causal connection rather than causation, to stay away from the image of effective mechanical effect-makers. The first aim is to define an appropriate notion of causal explanation.

It may help to keep a slightly paradoxical thought in mind. When we look for causal connections between events we want the connections to be at the same time quite delicate and quite robust. If the connection between one event and another is too robust, then this can detract from the explanatory value of the causal connection between them. All people die, but if you ask why a particular person died the answer 'because she was human' is not very satisfying. Tighter causal connections link the fact that a cause happened in some very specific way delicately to the fact that an effect happened in an equally specific way: if the cause details hadn't been just as they were the effect details wouldn't have been just as they were. But now in explaining the need for delicacy the counterfactual link has entered, and it seems to require a kind of robustness. If things had been different in various ways the effect would still have followed.

Of course the delicacy in question and the robustness are not incompatible. The factor we want as much of as possible is causal depth, and the factor we will accept only if doing so yields causal depth is causal narrowness. Here is an image. We can get towards various destinations by taking various signposted ways, which branch before us. There are other ways too, leading to other destinations. If we turn onto a signposted way then most of the paths that branch from it will get to the posted destination. If not only the main and easily taken ones but also the byways and remote detours arrive safely, then the connection from that way to that particular destination is robust – there is causal depth in your progress. If by choosing a way you can choose exactly which destination you will get to the connection is delicate. Delicacy and robustness are compatible, but usually you have a trade-off. Usually there are other ways with no signposts or posted to quite different directions. Then the ways that connect reliably, robustly, with destinations of interest may be few among all the ways. But, still, if you choose the right destination then there will be a reliable way to it. In that case the connection is deep but at the price of being narrow: for nearby destinations there might be no reliable way. So by careful choice of signposts and destinations we can ensure that if we take this way rather than that one most of the paths we are led to will lead to this destination rather than that one. But the choice may have to be pretty careful.

For events c, c', e, e', where c and e are actual and c* and e* are possible incompatible alternatives to them, say that *c rather than c* is causally linked to e rather than e**, or (c/c*) C→ (e/e*) when:

> there is a true proposition P such that
> for all P-worlds w* such that c* occurs in w* there is a P-world w such that w and w* are near to each other and to actuality, such that:
> e* occurs in w*, and
> c occurs in w, and e occurs in w, and
> there is no P-world w′ nearer than w to w* such that c* occurs in w′ and e* does not occur in w′, and
> there is no P-world w″ nearer than w to w* or equally near such that c occurs in w and e* occurs in w

(This is the definition that seems to me to capture best what is intuitively required. But I formulate a simpler version below.)

Essentially, this requires that if c occurs then e be possible and e* be excluded, and that if c* rather than c occurs, then e be ruled out by the occurrence of the incompatible e*. (And the effect is *almost* to say: take the nearest situation in which both c and c* are possible; at that point c would leave e open and close off e*, but c* would exclude e in favor of e*.) The details of the definition strengthen this gist in one way and weaken it in another. They strengthen it by requiring that these connections hold in a range of worlds near actuality.

The respect in which the definition is weaker than the gloss is in its concessions to indeterminism. The effect of the proposition P is to introduce a wobble factor that can make any causal link fail. Even ignoring this, the occurrence of c does not completely exclude that of e*. All that is required is that any world in which c occurs and e* occurs be more remote than the nearest (to w*) in which c occurs and e occurs. Note that there is an asymmetry in the treatment of c, e and c*, e*: the link between c* and e* is tighter than that between c and e. But this tightness is typically compensated for by the greater diffuseness of the foil e* compared to the target e. If c* occurs then some event falling under the heading e* will follow, but if c occurs then the specific event e will follow. If (c,c*)C→(e,e*) for all alternatives c*, e*, then e will occur in all nearby worlds in which c occurs: c will determine e. So this fairly weak indeterministic contrastive causal link generates a deterministic link when the contrastivity becomes redundant.

Mention of determination invites comparison with David Lewis's accounts of causation and of contrastive explanation, which do not presuppose determinism (Lewis 1986, 2000.) Lewis's account of causation is based on a requirement that, in effect, given the cause the effect can occur and without it it cannot. It is easy to see that this does not by itself generate an account of contrastive explanation. For consider situations like the following. A ball is dropped onto a sharp blade; if it swerves left it ends up in the left pan, and if it swerves right it ends up in the right pan, and if it is not dropped it is thrown away. It is dropped, swerves left and ends up in the left pan. The ball's being dropped causes it to arrive in the left pan, on Lewis's account. And plausibly so: once it had been dropped the way was open to the destination, which would otherwise have been closed. But it is not an explanation of why it arrived in the left pan rather than the right pan. It could have gone either way. The ball's swerving left does explain why it ends up in the left pan. So there is some further condition that is satisfied by the left swerve that is not satisfied just by the ball's being dropped. It cannot be that once the ball has swerved left it must end up in the left pan. For there could be a further blade, a right swerve which could destine it for a middle pan. The initial left swerve would still explain why the ball ended up on the left rather than on the right (but not why it ended up on the left rather than in the middle.) This and similar cases suggest that the added condition has to be that swerving left excludes the rightmost destination, or at any rate makes it less accessible than the leftmost one. And thus the definition above, which can be seen as starting with Lewis's core requirement and adding the conditions needed to make it into an account of contrastive causal explanation.

Two more points from Lewis:

(a) He defines cause as the ancestral of the counterfactual, if not c then not e. This step is not needed for our purposes since we are concerned with the one-link causation between an initial state and a whole subsequent trajectory. But such considerations do raise the issue of explanations which first show that $(c,c^*)\ C\!\rightarrow (e,e^*)$ and then explain some event e' which is a part of e by the fact that it was such a part.

(b) He has his own account of contrastive explanation, which uses his account of cause, unmodified, and adds to it in a different way. The effects are very similar:

> The causal link $(c/c^*)\ C\!\rightarrow (e/e^*)$ is *deeper than* the causal link $(d/d^*)\ C\!\rightarrow (e/e')$ when both links are true and the nearest world w* in which either (a) c* occurs and e* does not occur or (b) there is a world w at which c occurs and e does not occur which is nearer to w* than any

> world at which c occurs and e occurs, is more remote than the nearest world w^+ in which the corresponding failure occurs for the d to e link (i.e. either d* occurs and e′ does not occur or there is a world w at which c occurs and e does not occur which is nearer to w* than any world at which c occurs and e occurs.)

Note that this definition does not exclude worlds in which wobble factors P fail. So if a causal link holds only because of the occurrence of a fact which could very easily not have been the case, it will be a shallow link.

> The causal link (c/c*) C→ (e/e*) is *wider than* the causal link (c/c*) C→ (e/e′) when e′ occurs in all possible worlds in which e* occurs, but not *vice versa*.

Note that depth is a condition on the cause side of the link, and narrowness is a condition on the effect side. I have defined width rather than narrowness because depth and narrowness are the two desirable features. The widest causal link holds when e* is not-e. But note that there is another virtue of explanatory links, easily confused with width. That might be called sharpness/dullness. A causal link is sharp when it explains why a very precise event occurs, rather than explaining some more generally described event. (Explaining why e_1 occurs is sharper than explaining why e_1 or e_2 occurs.) Sharpness would be measured by the range of worlds in which e occurs. Not distinguishing narrowness and sharpness can make for confusion, because very often when we have a blunt connection – that is we explain why some rather diffuse event occurred – the causal link contrasts that diffuse with an alternative diffuse event. So a kind of width is achieved, but at the price of bluntness.

To illustrate: one marble, M, rolls along a surface that can have very variable texture, rough and muddy or smooth. It hits another, N. This sets off a chain effect involving six other marbles, and a marble T drops into a hole. The features of the surface are the background properties P. Take the P in question to define a smooth surface. Then, since the laws of mechanics apply unproblematically, we know that:

(a) M's striking N rather than missing it is causally linked to N's moving rather than not moving.
(b) M's striking N at angle θ rather than at angle θ* is causally linked to N's moving in direction δ rather than direction δ*.
(c) M's striking N at an angle rather than straight on is causally linked to N's moving in a different direction to M rather than in the same direction.
(d) M's striking N rather than not striking it is causally linked to T's falling into the hole.

(a) is deep, since the laws of mechanics hold in worlds to some remoteness. And it is wide, since the effect event is contrasted with a large alternative.

(b) is deep and less wide, since the effect event is contrasted with a small alternative. (It cannot be made wider since mechanics does not say why N does not disintegrate, or absorb M's kinetic energy as heat.)

(c) is deep and dull, since the effect event is large, but also fairly wide since the effect event is contrasted with another large event.

(d) is shallow, since such chain reactions are very sensitive to outside influences, so if any nearby massive objects had been in different positions T would not have fallen into the cup.

There are many ways in which one can specify the comparative nearness of possible worlds to make (a)–(d) come out in this intuitively right way. A simple way is to make the laws of mechanics hold in all of a set of worlds that vary only in the initial location of the marbles. Call one world more distant than another from actuality when the sum of the distances of the eight marbles from their actual locations is greater. Of course, (a)–(d) will follow on more complex models too.

It should be intuitively clear at this stage why there is generally a trade-off between causal depth and contrastive force. The causal depth of a link (c/c*) C→ (e/e*) depends on how near the nearest counterexample world is, that is, the more remote the first world in which either (c* and not e*) or (c and e*). If there are many worlds in which e* then a counterexample world in which c and e* may not be a world in which c and e′ for some e′ occurring in fewer worlds. The effect of the first conjunct is just the opposite: a counterexample world in which c* and not e* is more remote the 'larger' e* is. And in general, causal depth is increased by the following factors, using 'small' and 'large' to designate events that can occur in few and many ways:

> small c; large c*; large e; for a given small c and large c*, small e*; for a given large c and small c*, large e*.

The combination that includes the most depth-increasing factors is small c, large c*, and small e*. And that is the combination that is appealed to in many everyday explanations. We say that the fact that a ball bounced in this direction rather than in some other explains why it went through that window rather than that other window. We say that the fact that a person reacted with outrage rather than some other emotion explains why she defended herself against the charge rather than paying the fine. It is only when science allows deep and precise causal laws that we find frequent explanations which abandon these sources of causal depth in order to gain sharpness. The fact that the ball struck the wall at angle θ rather than θ′ explains why it bounced off it at (π-θ) rather than (π-θ′).

In the rest of this appendix I shall suppress the contrastivity in the cause argument of the causal linkage relation. I shall work with a relation c C→ (e/e*), and I shall simplify its definition as follows:

> c C→ (e/e*) if and only if there is a property P such that
>
> (a) given c, e is possible (there are nearby P-worlds in which c and e, none of them more remote than P-worlds in which c and not e)
> (b) if c then not e* (in all nearby P-worlds in which c, not e*)
> (c) if not c then not e (in all nearby P-worlds in which not c, not e).

And, to accommodate this simplification, I shall work with a simplified notion of the depth of a causal linkage. I shall say that a causal linkage c C→ (e/e*) is deeper than another d C→ (e/e*) when the nearest world in which c and e* is more remote than the nearest in which d and e'. These definitions will be much easier to work with, though I believe the full definition is a more accurate representation of contrastive explanation. (Though I would really want the definition of contrastive explanation to be *more* complicated than either one. I would want the link between the cause event and the possibility of the effect event to hold at least as wide a range of possible situations, as that between the cause event and the non-occurrence of the contrast event.) And the full causal linkage relation allows a better exposition of the ways in which explanations can be more and less powerful than one another. It is in considering these comparative relations between explanations that I expect insight into the nature of explanation to be found.

(The conditionals I am using here are somewhat different from Lewis's. They are the kind of conditionals that are coming to be used frequently in conditional accounts of knowledge. They are occasionally referred to as Nozick conditionals. Instead of saying that 'if p then q' is true if q holds in all *nearest* worlds in which p we say that the conditional is true if q holds in all *nearby* worlds in which p. This is a stronger condition in that it fails if q holds in the nearest but not another just slightly near world in which p. In epistemology the benefit is that it makes more sense of conditionals with true antecedents. That advantage is useful here too, but is also appropriate when comparing the remoteness of the possible worlds in which counterexamples to conditions occur to state those conditions in terms which apply through a range of worlds. The price for these advantages is an intrinsic vagueness.)

Part 2: attractors

Assume that we have a domain D of objects, a set T of times, and a partition Π of D, i.e. a set of properties holding of members of D at times in T such that one and only one property holds of every x at every t. Assume that every x ∈ D is such that for every t there is some 'initial' property I and some 'final' property F, (I, F∈ Π) and some t′ ∈ T (t′ ≥ t) such that Ix,t and Fx,t′ and

$$(Ix,t)\ C\!\rightarrow (Fx,t'/Gx,t')$$

holds for all $G \in \Pi$. Abbreviate this as $Ix\ C\!\rightarrow F$. Call these *dynamic links*.

An *attractor* is a union Q of members of Π such that there is a union R of members of Π such that $Q \subset R$ and:

> for all $F \subseteq Q$ there is an $I \subseteq R$ $(F, I \in \Pi)$ such that for all $x \in I$ and all t
> $Ix\ C\!\rightarrow F$.

Call R the *range* of Q. Call Q a *strict attractor* with respect to R if Q is an attractor with range R and for all $I \subseteq R$ and all x, t $Ix\ C\!\rightarrow F$. Assume also that R has modal significance, in that if $x \in I$, where $I \subset R$, then in the nearest possible worlds in which $x \notin I$, $x \in I'$ for some $I' \subset R$. (If x had been wobbled out of I then it would still have been in some other part of the range of the attractor.)

The idea is that an attractor consists of destinations of states in R. When the attractor is strict it consists of all the destinations of states in R. Since the connections between a member of D being in a subset of R and being in Q are causal links rather than full causation, each one will lead from its initial state to its final state only if some property P holds. So each transition is subject to 'wobbles' that may cause it not to occur. The degree to which these transitions are affected by wobbles is a major factor in determining their causal depth.

Note first that the set of causal links at the level of initial and final properties generates another link at the level of ranges and attractors. For all objects x and times t there is a time t′ such that:

$$(Rx,t)\ C\!\rightarrow (Qx,t'/Q^*x,t')$$

for any Q* such that Q* is a union of members of Π such that $Ix\ C\!\rightarrow F$ for no $I \subseteq R$ and $F \subseteq Q^*$. For suppose that x is in R. It is therefore in some $I \subseteq R$. There is therefore an F and a t′ such that $(Ix,t)\ C\!\rightarrow (Fx,t'/{\sim}Fx,t')$. But $F \subset Q$, so in any world in which Ix,t it is the case that Fx,t′, thus Qxt, and for no x in R is it possible for x to be in some $I \subseteq Q^*$ at any t′, by the modal significance of R. Call this the attractor link for R and Q.

(It is worth relating the inherent contrastivity of attractors to the motivating image of contrastive explanation used above. A range R corresponds to the situation of the dropped ball just after it swerves left after encountering the first blade, with the attractor Q being the left pan. If there is no second blade then the attractor is a strict one. But if there is a second blade then the fact that the ball is in R explains why it is later in the left rather than the right pan – there are sub-situations of R which are initial states of final states in the Q. It does not explain why it is later in the left rather than the middle, since R contains states leading both to the left and the middle.)

The attractor link has a contrastivity built into it, generated by the underlying causal connections. These also generate an inherent contrastivity to the dynamic links. But the contrastivity here is trivial. If Ix C→ F holds then (Ix,t) C→ (Fx,t′/F*x,t′) for every F* distinct from F. This has consequences for the causal depth of the dynamic links. As in Part 1 above, measure the causal depth in terms of the nearness of the nearest possible world in which the link fails. Then a counterexample for a dynamic link will consist of a world in which the initial state leads to a different final state. But then the initial state cannot, in worlds near that world, lead to the actual final state. So the world is a counterexample for the dynamic link whatever contrast is employed. The conclusion is important: *the causal depth of the dynamic links is insensitive to contrastive breadth.*

This conclusion does not apply to the attractor link. A counterexample world to the attractor link for an x and a R will be more remote the larger the Q. That is, if there are Q, Q′ such that both are attractors with range R and Q $\subset$ Q′, then the nearest counterexample to the attractor link for R and Q is no more remote than that for the attractor link for R and Q′. Moreover, the causal depth of an attractor link is always at least as deep as that for any of its associated dynamic links. For consider a dynamic link Ix C→ F and an attractor link Rx C→ Q, where I $\subseteq$ R and F $\subseteq$ Q. A counterexample to the dynamic link is a world in which Ix and F′x where F′ is not F. (I am suppressing the time references, for simplicity.) If F′ $\subseteq$ Q then this is not a counterexample to the attractor link. On the other hand a counterexample to the attractor link is a world in which Rx and Q*x where Q* is disjoint from Q. In such a world x $\in$ F for some F not $\subset$ Q. So this is a counterexample to the dynamic link. So any counterexample to the attractor link is a counterexample to the dynamic link, but not vice versa.

Again an example of falling balls and blades may help. Suppose the balls fall through an array of little tubes. Beneath one set to the right there is a blade with a wobbly tip, and beneath another set to the left there is another. Beneath the blades there is another array of tubes, some to the right of both blades, some between them, and some to the left of both. A ball passes through a particular upper tube on the left almost directly above the left blade, and ends up in a particular lower tube on the left. The fact that its *initial* location was that particular upper tube explains why its *final* location was as it was rather than any other. The fact that it was in the left set of tubes explains why it ended up in one of the tubes to the left rather than one of the tubes to the right, but does not explain why it ended up in one of the tubes to the left rather than one of the middle tubes. The explanation in terms of sets of tubes is causally deeper than that in terms of individual tubes, in that if the tip of the blade had wobbled just a little the ball would not have ended up in the very tube that it did, but would still have ended up in a tube on the left.

To sum this up: attractor links are at least as causally deep as dynamic links, and unlike dynamic links their causal depth increases as the contrast class becomes wider.

Chapter 5

Learning to simulate

I wish that for just one time
You could stand inside my shoes
And just for that one moment
I could be you

Yes, I wish that for just one time
You could stand inside my shoes
You'd know what a drag it is
To see you
Bob Dylan, *Positively 4th Street*

You know my methods in the usual cases, Watson. I put myself in the man's place, and having first gauged his intelligence, I try to imagine how I should myself have proceeded under such circumstances.
Arthur Conan Doyle, *The Musgrave Ritual*

We rarely anticipate or understand one another purely out of curiosity. We are more often after some information that we need for some purpose. This book focusses on the purposes for which it is important to be understood as well as to understand. In this chapter the topic is the congruence between understanding and being understood. There are basic reasons why we need to understand by methods which are congruent with those with which we are being understood, as sections 'Enter simulation', 'Kinds of simulation' and 'Learning to simulate' below show. One result is a scope for learning. Each person has to acquire ways of anticipating and interpreting others that mesh with the ways the others have already acquired. And each person has to acquire the capacity to fine tune these ways, on particular occasions, to mesh with the fine tuned procedures that others are employing. Moreover, I shall argue, a very attractive form for such congruently tunable procedures puts them in the general area that recent philosophers have labeled 'simulation'. In fact, in the kinds of situation that I discuss most in this chapter, in which it is in people's interest *not* to know in advance exactly what one another will do, I think that the only procedures that have much chance of working have the general appearance of simulation.

So, in a nutshell, this chapter argues that each person must have a variety of generally simulatory procedures matching the variety used by others with whom they interact.

The loonie game

The need for congruence is illustrated very clearly by a game which distills crucial features of many everyday situations. It also illustrates an important and little-remarked fact, that sometimes when people deal with one another it is in the interests of all of them not to have too detailed predictions of what one another will do.

Two people X and Y are playing a game. There is a pile of 100 Canadian dollar coins – known as loonies from the picture of a loon on the reverse – on the table before them. X and Y take it in turns to take either one or two coins from the pile, and they keep the coins they take. However, as soon as either of them grabs – that is, takes two coins at once – the game stops, and the rest of the coins are cleared away. So long as they each take only one coin when their turn comes, the game continues till the pile is exhausted. Suppose the number of coins is quite large, and even, and X has the first turn. What should she do?

Most people faced with this problem begin at the beginning. They think they should clearly take a coin to begin, because if they take two coins then that is all they will get. And they see that if both people put off taking two coins till near the end then they will both do better than if either had grabbed two early on. However, there is an argument for grabbing at the first move. The argument begins at the end, as follows. If ever there was a moment when there were only two coins on the table then the person whose move it was would take both, ending the game. Suppose there comes a moment when there are only three coins left on the table. Then the person whose turn it is will surely take two of them, ending the game. For if they only take one then the other will take the remaining two, and they will have only got one of the two, rather than the two they could have had. But three coins can only be there if at the previous move the other person will have been faced with four on the table. At that stage they will know that if they only take one then their opponent will take two next move, when there will be three on the table, as just argued, ending the game and leaving them with no more gains after that one. So they will take two. So at the previous move to that, by similar reasoning, the person who has that move will take both, ending the game. So rolling back in this manner from the end to the beginning we seem to have an argument that whoever starts first should take both coins. They should be satisfied with just two, knowing that if they had only taken one the other player would have grabbed and ended the game at the second move. (This game is often known as the 'centipede' game. It is related to the much-studied chain-store ‘paradox’, and to roll-back arguments for defecting from the beginning of a series of prisoner’s dilemmas. See Kreps *et al.* 1982, Pettit and Sugden 1989, Chapter 4 of Kreps 1990, Chapter 4 of Myerson 1991, Bacharach 1992, Morton 1994, Bermúdez 1999 and Broome and Rabinowicz 1999. One reason games like this are studied is that they seem to suggest problems with standard assumptions about equilibria as solutions to strategic situations, and to motivate refined versions of equilibrium.)

To get to this paradoxical recommendation each person has to have quite precise thoughts about the other person’s thoughts. Or, to put it more carefully, each person has to think that the other one will form an exact and rationally calculated plan of action on the assumption that she herself has an exact and rationally calculated plan. As a number of writers have pointed out, the reasoning involved breaks down if this is not the case. But real people do not have such precise thoughts about one another. (And in this case they benefit by not having them.) They do have some expectations. If either thinks that the other is going to grab very early, then she will too, if possible even earlier. Each also needs to know that the other’s expectations are not too different from her own plans. If either thinks the other thinks she is going to grab very early then she will change her plans and grab at least as early as the other expects. But more striking than the need to have expectations is the need *not* to have or to induce expectations that are too precise. For if it is known to both of them that either of them expects the other to grab at any precise turn, then the inevitable cascade of inferences is released, leading to grabbing at the first move. And this outcome is in the interest of neither participant. (It is not surprising that sometimes it is important not to be too predictable. Submarine captains in World War II used to throw dice to decide on their route, to avoid accidentally being predictable. The present twist on the unsurprising fact is that it is sometimes in the interest of the potential predict*or* that the potential predict*ee* be somewhat random.)

Real people playing the loonie game do not grab very early. Nor do they grab at a fixed late stage. They grab late enough to get a reasonable return from the game, and at a varying enough point to keep one another uncertain of their exact choices. They form *imprecise* expectations about one another’s choices. As indeed they must if they are to do well. Moreover, the imprecision of the two players is congruent. Each will form an expectation that is precise enough to allow her to hold off grabbing until there is a significant danger of the other doing so, and vague enough that it does not entail that she should grab early. (There do not seem to be proper empirical studies of how people behave in loonie-type games. My amateur experiments with undergraduates playing for pennies is that grabbing rarely occurs before the middle of the sequence, and it will become steadily earlier as the game is played repeatedly, and then become later, and then tend to fluctuate around a fairly late focus.)

We thus have an argument for congruence. In situations like this agents need to have expectations that match one another's. That is, each interacting participant needs to form her expectations of the others' actions by a method that will give good results given that the others are using the methods that they do. Moreover, these expectations must not be perturbed by knowledge of the expectations formed by others. (So I not only expect you to do A and you not only expect me to do B, but I expect you to expect me to do B, and you expect me to expect you to do A. And so on. And reasoning from these higher order expectations must not undermine the simple expectations, or make each cease to be good given the expectations of the other.) Consider the special case in which congruence requires that the expectations of two interacting people be identical. The expectations might in theory be based on cognitive patterns that are structurally different though they had the same output. The two people might get to the same answers by different methods. But in practice this would be unlikely, if the methods were ones that gave the same answers over a range of variations on the situation, and generated the same higher order expectations.

Enter simulation

In these and many other situations people have to think about one another's decision making in ways that are congruent with the ways the others are thinking about them. If there was a single obviously correct theory of rational action, accessible to all people, then congruence might be achieved by striving to act rationally and by assuming that others will so act. But there is no such theory. Theories of rationality that handle even slightly hard cases correctly – where risk, causation, or strategicality play any non-trivial role – must be expressed in technical terms that are remote from the language of common sense. These theories moreover describe actions that are very different from the choices people actually make. And, most important at this point in the argument, they don't give good results when shared. If everyone subscribed to any manageable version of an orthodox theory of rationality and assumed that everyone else did too, everyone would defect in prisoner's dilemmas and grab in loonie games. (In fact, the thrust of most recent work on roll-back arguments – see the works cited above, especially Pettit and Sugden 1989, Bacharach 1992 and Broome and Rabinowicz 1999 – is that it does not pay to expect others to have explicit rational plans of their future actions or to expect them to expect each other to. If explicit rational expectations about others means 'what would follow from a complete theory of rationality' the standard conclusion is thus 'it does not pay to understand others in these situations by using a complete theory of rationality.')

These are problems of shared explicit theories encoded in public language. The same problems would arise with a shared implicit theory. We might have some representation of a complex, sophisticated and accurate theory of choice, which we were not able to express in words or consciously articulate, but which gives us reliable expectations about one another's decisions. The fundamental problem here is simply prediction. As the loonie game shows us, in many situations we do not want to share precise predictions of one another. And yet in others that is just what we do want. In fact, in many situations people can do well only if they share precise common expectations of their choices. (The simplest examples are situations of pure coordination.) A procedure that gives satisfactory congruent expectations must sometimes produce the most precise predictions that it can and sometimes produce explicitly imprecise general expectations, of the particular species of imprecision that the situation demands. How can this be?

There are fairly tight constraints on the thinking involved. It has to be sensitive to a variety of information about the other. In particular it has to be sensitive to intuitions about the other person's arational dispositions. (If one just gets the feeling that the other is a nervous type who will panic and grab early for no good reason, one will be pressured to grab early oneself.) It also has to be *plural*, by which I mean that it must take a description of a situation facing a number of interacting people, and produce a prescription for their interlocking actions. This prescription can be the basis of expectations about other people's actions and intentions about one's own. And, thirdly, it must be congruent, in that it works best if both people employ it.

There may be several ways of meeting these constraints, but one seems particularly natural and appealing. It is a version of what in Chapter 1 I called 'solution-based thinking'. One first thinks of an outcome which one can imagine the other person or persons both would want to achieve and would believe that one would try to achieve. One then thinks out a sequence of actions by all concerned that will lead to it. Lastly, one performs the actions that fall to one's account from this sequence (until something happens to suspend the plan) and expects the other(s) to do their corresponding actions.

It is not hard to see in outline how solution thinking can meet these constraints. At the first stage information, rational and intuitive, about the other fixes a reasonable two-person ambition, for example how far into the loonie sequence to get before cooperation collapses. The second stage is essentially plural in that it finds a route to the goal (a sequence of single coin-takings) and then deduces what each person must do. And the whole procedure is congruent, in that if both people aim for, say, getting to very near the end before either grabs, then this aim will be achieved. (But if only one is thinking in these terms, success is unlikely.)

And it is simulation. First of all it results in an understanding of others: you have an idea what the other is going to do, and you have the materials for putting together an account of why they may do it. And the understanding is got by reproducing rather than representing their thinking. For if all the interacting people are choosing their actions by this procedure then each will gain an expectation of what the others may do by performing the same thinking routine that they are performing. I find that when discussing this topic with others this is a point I have to underline. Simulation is understanding others by going through the same thinking as they do. As it is described by Gordon, Goldman, or Heal we tend to think of a person reproducing the thinking of another as a detached observer, insulating the output of the simulated thinking from their own actions ('off line'). But exactly the same procedure can be followed if the simulating and simulated person are acting together and if the simulated procedures result in action. The important thing is that one's understanding of what the other person may do comes from the fact that one is running through the thinking that results in their actions: if it results in one's own actions as well nothing is essentially changed. In fact the procedure is simpler, since one does not need to manage a delicate balance between keeping the thinking live enough to reproduce the other person's actual thoughts and yet restrained enough that it does not produce one's own actions. And there is then no reason why each person may not be simulating the thoughts of the other, simply by approaching the same problem in the same way and by framing their approach to it with the intention of using the result as a guide to the actions of the other. (Mutuality, each person having a grasp of the other's thinking, and indeed of the other's grasp of their grasp of her thinking, is thus achieved without complex embedded thoughts.) A simple, child-like way to mediate shared activity thus has essential characteristics of sophisticated and reflective understanding of another.[1]

Moreover, there really is no alternative to thinking in some such way. Suppose that you are one of 13 people needing to coordinate their actions. You might try to represent to yourself the thoughts of each of the other 12, including each (first) one's thoughts about each (second) one's thoughts about your (13th) thoughts. But that clearly will be beyond your capacities. The only feasible general technique is to think out what would be a satisfactory joint reaction to the situation. (This is not to deny that there are many feasible non-general techniques. You might for example know that this particular group of people always acts in some particular way.) And then, time and resources permitting, you may consider of each person whether she would see and act in accordance with this joint course of action.

The application to situations such as the loonie game is straightforward. Each person has to see that both will do well if they set out not to grab too early, and this thought has to be shared. And then they have to set off on that course, assuming that the other is also setting off on it. Simulation – shared solution thinking – can deliver this. And it is hard to see how any non-simulatory procedure can. For any procedure that separates the reasoning of the two people will require that each have an explicit prediction of what the other will do. And given that prediction, each can then modify their intention from the very beginning, which will either undermine the prediction or result in mutually disadvantageous choices. In the rest of this chapter I elaborate this point in several ways, but this is the core idea. Only simulatory procedures will give interacting agents access to parallel paths through situations in which prediction undermines cooperation.

Kinds of simulation

It is almost inevitable that in many strategic situations people reason in ways that reproduce rather than represent the reasoning of others. But there are many forms this can take. Some of them contrast sharply with thinking in terms of explicit thought-out predictions of action, and some will overlap with 'theory' based methods, using propositional information whose manipulation is largely a matter of inference from learned facts. (Few now believe that there is a single mind-understanding procedure that we can label 'simulation', or that there is a sharp contrast between simulation and theory.[2] The consensus now is that there is a large variety of cognitive routines which combine and overlap with each other and with routines that are fairly labeled 'theory'. See Goldman 1992a, Currie 1995b, Gordon 1995a, 1995b, Davies and Stone 1995a, 1995b, Perner 1996, Gopnik and Meltzoff 1997, Perner *et al.* 1999, Goldman 2001. The two philosophers who have thought hardest in recent years about simulation, Goldman and Heal, have refined their versions in such different directions – what I below call cocognition and modeling – that the default assumption at this point is that there is little psychological unity here.) Ideally we would like to know precisely which kinds of thinking can mediate which kinds of strategic situation. That is an extremely hard question. The rest of this chapter is structured around a series of gestures towards and partial answers to it. To begin, here are four basic contrasts between the wide range of processes that may fairly be called simulation.

Modeling versus cocognition

In what I shall call modeling the person attempting to understand puts herself through a cognitive process which is in some way related to the situation of the person to be understood, and then uses the history or output of that process as a guide to features of the thinking of that other person. In cocognition, on the other hand, the person attempting to understand puts herself through a process which is based on the problem the person to be understood is confronting, and then uses the result of that process as a guide to the solution that the other may have found. Modeling asks 'what would I go through in that situation?' while cocognition asks 'what is the answer to that question?' (Or more generally modeling asks 'what do I simulatedly do if I prepare myself in this way?' while cocognition asks, 'what is an acceptable answer that fits what I know about that person?') The two are easy to confuse because if the other person is thinking correctly then in cognizing them one is usually also modeling some aspect of their problem-solving. The difference emerges clearly when understander and understandee are not on a par with respect to the problem. Suppose, for example, that you are trying to anticipate the answer a person will give when asked to find 13 cubed by mental arithmetic. As it happens they are extremely good at that sort of thing and you are not. After cocognizing you emerge with the prediction that they will give an answer of 2,197. If you are using your efforts to model them you may also come up with the inappropriate prediction that they will be flustered and on the edge of a headache. It would be more appropriate to model this person's state of mind after confronting this problem by getting yourself to cube 5. (The clearest contrast is between Jane Heal and Robert Gordon. See Gordon 1995a, Gordon 1995b, for modeling, and Heal 1998 and Heal 2000 for cocognition. I take Goldman 1992, 2001 also to be presenting an account of modeling. See also 'Heal's problem' in Chapter 2 of this book.)

Fundamental versus folk-psychological

Any understanding by simulation will center on a transition between states of the understander. These states are then the basis of attribution of states to the person to be understood. What kinds of states are we talking about? Some theorists talk of using one's own decision processes as a model for those of another person, and conceive of decision as a procedure that goes from beliefs and desires to intentions or actions. But beliefs and desires are states postulated by commonsense. Whatever it is that goes on in a person when she pauses for a few moments and then takes one course of action rather than another, its fundamental nature is not going to be fully expressed in terms of beliefs and desires. Some of the relevant processes may be describable in terms of sub-personal content-bearing states of a kind that cognitive psychology could understand, and others may only be fully intelligible in more fundamental neural terms. Writers on simulation often motivate an account of simulated decision with examples expressed in terms of beliefs and desires, and then apply the account to cases in which something more fundamental must be involved. For example Gregory Currie gives a fairly standard description of off-line simulation in which a person feeds into their own decision processes another person's beliefs and desires, and then goes on to talk of simulation in which one takes on another's perceptual state. It is clear from Currie's exposition that what he has in mind in the latter case is imagining oneself in the physical situation of another, and then taking account of whatever one's thus-adapted perceptual system produces. (Currie 1995, p. 145. It is hard to know exactly how far apart folk psychological and fundamental simulation are, just because it is hard to know how significant a part in shaping our actions the states we have folk psychological names for are. That is one of the questions that in Chapter 1 I urged us not to beg.)

The difference becomes important when we consider what has to be done to get a simulation going. When it is a matter of simulation by folk psychological states one needs to have to hand an attribution of such states. One needs to know for example what the person to be understood believes or wants. On the other hand when the simulation is in terms of more fundamental states one needs a capacity just to get into those states, whether or not one can describe them. Consider an example. The task is to predict the choices of someone much less capable at an activity than one is oneself. (The activity could be managing a delicate social situation, or playing chess, or getting home at night across an unfamiliar city.) One way, folk-psychology-oriented, would be to identify the person's somewhat naive beliefs about the topic, and then to carry out one's own problem-solving routines guided by these flawed beliefs. The result might not be completely useless, but it would be of pretty limited validity. (But then the task, of getting a grip on the decisions of people who understand something significantly less well than one does, is very hard. That is one reason why good teachers are rare and valuable.) A very different method might be used by someone who over a period of time had acquired an intuitive sense of the way the person in question handles this kind of activity. (Perhaps the mother of a mentally handicapped child, who finds that she can tell where the child will have got lost, or how the child will react to teasing.) Such a sense need not take the form of simulation. But it can be clear simulation: the understander may find that she can report half-way-there incomplete solutions, and then when she arrives at a prediction it may take the form of a suppressed impulse to do the act in question. Then the adaptation of the person's own thinking to the predictive task is much more subtle than in the folk-psychology-oriented case. It is the thinking itself rather than its inputs that is tuned to match the target. The tuning is very unlikely to be describable in any commonsense terms. The explainer may have no useful description of what she is doing and how someone else might do it. It is extremely plausible that we do a lot of this basic-states-oriented simulation in our everyday social life, that the capacities it draws on vary greatly from person to person, and that we can rarely tell in advance when we can rely on them. (Some evidence that fundamental simulation can operate from a very different basis than the off-line assumption of conceptualized states required for folk psychological simulation is found in the work on mirror neurons cited in Goldman 2001.)

There is a connection between this contrast between kinds of simulation and a contrast between two kinds of cognitive illusions. On the one hand we have illusions such as the framing effects investigated in the psychology of decision making, notably by Tversky and Kahneman. In these cases the illusion is simulable, in that if you ask one person how another would respond to a problem as framed in a particular way then the person will be able to frame their simulation of the other person's thinking with the same frame, and will come up with an accurate prediction. These must correspond to a first approximation with cases where simulation in terms of folk-psychological states is reliable, for it is in terms of such states that the framing is described. On the other hand we have illusions such as those in which people radically underestimate their understanding of most topics, or a variety of cognitive dissonance cases. In these cases a person will not simulate the mistake made by another. And, presumably, the reason is that an instruction to oneself to get into the state of mind of the person being simulated, expressed in folk psychological terms, will not manipulate the actual variables concerned. (Which is not to say that it is beyond our powers, but that the capacity to do it will be hard to describe and hard to learn, with skills of simulating different illusions in different sparse distributions among the population. Salespeople acquire one kind, teachers of rhetoric another.) And, touching again on a central theme, we must form our patterns of everyday interaction in such a way that we avoid having to predict the kinds of illusions that we find hard to simulate. (The simulation-illusion link was suggested to me in conversation by Josef Perner. See Tversky and Kahneman 1992, Stich and Ravenscroft 1996, Johnson and Keil 2000.)

Centered versus non-centered imagining

This distinction is clearest when it is actions that are being imagined. Suppose you are imagining that someone on the far side of a crowded room might come over to talk to you. You may imagine a path that she could take and imagine her walking along it. You may think of the path – the obstacles to a direct route, and the ways around them – in terms of locations and directions as seen from *your* location. That is non-centered imagining, because it is not centered on the location or point of view of the person whose actions are being imagined. Alternatively you may think of the path in terms of locations and directions as seen from the other person's location. That is centered imagination, because it is centered on the location or point of view of the imagined person. (The distinction comes from Wollheim (1984). See also Goldie 1999, and Chapter 7 of Goldie 2000. Wollheim and Goldie say 'central' and 'acentral'. They also speak of peripheral imagining, when there is a center, but it is not that of the imaginer. I have been writing as if non-centered imagining is also centered, but not on the imagined person. But non-centered imagining may have no center at all. You might think of the route from there to here in terms of a map of the room on which your location and that of the other person were just two points marked with Xs.)

It may make a big difference to how one thinks of an action whether one imagines it in a centered or non-centered way. Suppose that there is a direct route from where the other person is standing towards your location, but which is blocked by an obstacle near to you, which would be hard to see from the other person's location. (A low table, say, on the far side of which there are some people talking.) Thinking non-centeredly, you might imagine the person avoiding that fruitless path, and going around the edge of the room. So if you wanted to meet her you might set off to intersect that peripheral route. But thinking centeredly, you would see that from the other person's point of view it made more sense to begin by coming across the middle of the room. So if you wanted to meet her you might go towards the obstacle, planning then to intersect the work-around she'd improvise when she found her route blocked.

This is not to say that the predictions that flow from centered imagination cannot be got by non-centered thinking. The information can always be translated into the other format. But doing this adds to the cognitive burden of the simulation. Some imaginative tasks are most easily done in centered terms and some in non-centered terms. Very often we oscillate between them, as when you walk towards your friend, who is finding her way around tables and guests, and you plan your route in terms of how things seem to you, but also in terms of where you expect her to go next, thinking this out in terms of how things will seem to her. Quite hard work.

Singular versus plural

One can imagine the concerted actions of several people, and one can imagine the reasoning that lies behind them. Sometimes the simulation that predicts or explains one person's behavior is inseparable from that which predicts or explains another's. There are many examples in strategic thinking: situations of pure coordination will again illustrate the point. When two people who need to meet but have not agreed on a location both head for the camera obscura rather than the docks or the bus station we can think of the reasoning that goes through each of their heads. We can represent each as figuring out what to do given what they know about the other. The reasoning is usually quite complicated, even when the situation is intuitively simple. Or, alternatively, we can think about the problem which each of them is grappling with, and use our reaction to it as a guide to theirs, to what they will do as a pair. In this latter case the simulation is plural: we are using our thinking to give us a grasp of the thinking of more than one other.

Plural simulation is particularly common, and particularly invisible, when the plurality consists of oneself and one or more others. (I argued for this in Chapter 1.) Then the thinking that tells you what you should do also yields a prediction of the other people's actions. Neither can precede the other. So your understanding of what they will do is got not by reflecting but by doing, not by thinking about their thinking but by thinking yourself about what to do. Of course they can be taken to be doing the same, so that various embeddings collapse nicely: in simulating them you are simulating their simulating you.

All of these contrasts are independent. There can be processes that combine one feature from any one of the four pairs. I would go so far as to say that there are processes of all possible combinations, which are on occasion used to predict, explain or otherwise grasp actions and motives. (Not that it is an apriori truth that *any* form of simulation is central and indispensable in everyday understanding. There is a strong case to be made, but amazing evidence could always emerge.) There is a tendency to cluster, though. Modeling, fundamental states, centrality and singularity tend to be associated, as are cocognition, folk psychological states, and non-centeredness. (And plural simulation tends to be cocognitive and non-centered.) For – to consider the first cluster – if we are modeling, instantiating in ourselves something like what is actually going on in the other, the accuracy of the result will depend on psychological details that are unlikely to be captured at the level of folk psychology. It may depend on differences between ways of accepting propositions that are lumped together by 'belief', or differences between ways of getting from one idea to another that are lumped together by 'reasoning'. And simulating with fundamental states is cognitively demanding – it takes a lot of time and thinking to get accurate results – so that one is less likely to be able to do it for more than one target person. The other cluster occurs in part because thinking cocognitively is usually thinking about a practical or intellectual problem which one poses in folk-psychological terms (what to believe given this evidence, what to decide given this aim, these means, and these facts). And it often results in a conclusion about a right answer whose rightness is independent of its point of view. But, as I said, neither of these clusterings is exceptionless. In particular, it is important that one can have the capacity to know what a group of people will do, without thinking in terms of the individual motives of the group's members. When this can be described as simulation it is plural and fundamental. Sometimes, also, when anticipating someone's solution to a problem one follows an intuition about which of several possible lines of thought is appropriate that cannot be expressed in everyday language. That is cocognitive but not completely folk psychological.

Three levels of cooperative thinking

Now to correlate these kinds of simulation with the thinking needed to get through various strategic situations. I focus on situations in which there is an outcome that is in the interest of all the interacting agents, and in which the problem for each of them is, intuitively, to find a way to such a cooperative solution. I shall discuss three broad classes of such situations, which present progressively greater obstacles to cooperation: simple coordination, gentle approach concavity, and abrupt concavity. (The loonie game is an example of gentle approach concavity, and the prisoner's dilemma is an example of abrupt concavity. 'Concavity' is a standard term, but the gentle and abrupt qualifications are mine.) In going from the first to the third of these classes we move from a plural non-centered cocognition (in effect what in 'Enter simulation' above I called a 'simple, child-like, way to mediate shared activity') to a greater reliance on centered and fundamental modeling.

Begin with the simplest, plural cocognition. Suppose that you are one agent among others facing a strategic situation. Assume a list of agents, including yourself, and of the actions available to them, and of the desirability for each agent of what will result from each combination of the agents' actions. Suppose that these desirabilities are expressed first of all as preference rankings, though you can access finer-grained information, if you need it. So you have a matrix of outcomes, with one cell for each combination of an act by each individual, and in that cell the position of that outcome in each agent's preferences. Now add an *acceptability criterion* for outcomes. This criterion will select an outcome that is generally acceptable to all concerned. This must meet a number of conditions. There is a practicality condition. It must be easy to apply, in particular it must be clear after a very finite procedure whether an outcome meets the criterion or not. There is an optimality condition. An outcome meeting the criterion must be as good an outcome for all concerned as can be found. And there is a fall-back condition: if after a finite search no good enough outcome is found there must be a systematic way of retreating to another. There are obviously many ways of meeting these conditions, and most of what I say will be true of any procedure that meets them. To make things definite, consider the following procedure, which is based on a very simple acceptability criterion defined by four clauses:

(i) An outcome is acceptable in the first instance if it is ranked first by all agents.
(ii) An outcome is acceptable in the second instance if it is ranked first by a majority of agents and no lower than second by a majority of those for whom it is not ranked first.
(iii) An outcome is acceptable in the third instance if it is ranked no lower than second by a majority of agents and no lower than third by a majority of those for whom it is not ranked second.
(iv) An outcome is not acceptable if it is not acceptable in the third instance.

This criterion can be applied or rejected in a finite number of steps. It will never count an outcome acceptable when there is an alternative that is better for a majority. And it will automatically fall back from first to second to third instance, then to a verdict of unacceptability, after a search of limited length.

The natural decision procedure to apply this criterion to situations is as follows.

1 Search through the whole matrix for outcomes that are acceptable in the first instance. If they exist, save them and go to 4.
2 If there is no outcome that is acceptable in the first instance, search the whole matrix for outcomes that are acceptable in the second instance. If they exist, save them and go to 4.
3 If there is no outcome that is acceptable in the second instance, search

the whole matrix for outcomes that are acceptable in the third instance. If they exist, save them and go to 4.

4 (a) If there is exactly one saved outcome, output it as an ordered sequence of agents and acts

(b) If there is more than one saved outcome compare all pairs of saved outcomes and reject an outcome in favor of another when the lowest preference it satisfies is lower than that of the other, or when there are a greater number of preferences satisfied at this lowest level. If this results in exactly one saved outcome, output it as an ordered sequence of agents and acts.

(c) If there is more than one saved outcome after (b) choose one saved outcome arbitrarily, and output it as an ordered sequence of agents and acts

(d) If there are no saved outcomes, output failure.

Call this procedure RC, since it searches for a rough unstated consensus between the agents about what is in their common interest. I have described it in probably unnecessary detail in order to make it clear that it can be carried out mechanically and that in most cases it will produce an output in a small number of steps. (In the appendix to this chapter I take the issue of the comparative expense of decision procedures a little further. 4(b) is obviously a non-essential, element, indeed one that is expensive in comparison with the rest of the procedure. It could be replaced with a drastically simpler routine.) There are situations in which it will not come up with any acceptable action, obviously, and there are situations in which it will come up with an intuitively wrong answer, as we shall see. It is particularly important that the acceptability criterion in question does not explicitly look for equilibria, though in many cases the outcomes it finds will be equilibria. It is more accurate to see it as an approximate search for Pareto-optima, that is, for outcomes for which any alternative makes some agent worse off. (The match is not perfect, but the procedure will come up with Pareto-optima more often than it will come up with equilibria.) Procedures that look for Pareto-optima are generally less expensive cognitively than those that look for equilibria, as is argued in the appendix to this chapter.

RC is a particular instance of cognitive plural simulation. It is plural because it arrives at a decision for the thinker and predictions about other agents as a result of the same process. And it is simulation because it allows one person to draw conclusions about what another person might do by using her own decision-making capacities, rather than beliefs about those of the other. Reproducing rather than representing. It is well suited for finding good solutions to coordination problems. That is, if there is an outcome such that each interacting agent will understand that if all or most choose it then all or most will do well, that most will do badly if their choices diverge, and that most also understand these facts, then RC will find it. For the outcomes in which all agents do the same thing will be picked out in the first three stages, and will be the only outcomes picked out. As a result if all of the participating agents employ RC then nearly all will do reasonably well.

RC is not so ideally suited for guiding agents through other situations. Consider again the loonie game. (And since it is defined for only two agents, read ‘all’ for ‘majority’ when applying RC to it.) First consider what would happen if the two people could decide in advance how many dollars they would take at each stage. (If they get to it; the other person’s choices may determine whether they get that far.) The first preference outcome for each agent involves the other agent taking just one dollar at each stage until the agent in question can either end the game by grabbing the last two or take the very last dollar. But that is the second preference outcome for the other agent. There are thus two acceptable outcomes, and following RC (as is intuitively natural) each will make an arbitrary choice between them. Suppose that each player chooses the outcome that is their own first choice. Then the player who plays first will grab at the fiftieth round, making it impossible for the other to carry out her conditional plan and take the last dollar.

That is a reasonable outcome for both players, even though one has two dollars more than the other. But it is *not* what we can expect to happen, and not what actual players would expect to happen. So something must be wrong with it as a model of actual strategic thinking. RC is not right in this case. To see why imagine that the game is actually being played. It has got to the 49th round and the player who will move second can see that if she takes two dollars then the other will surely take two in the next round. This leaves her with $48, while if she takes two now, at round 49, she will get $50. So she grabs now, abandoning RC for simpler non-strategic reasoning available in the simpler case that has now developed. Real players will not follow this reasoning all the way back to its conclusion that they should grab at the first round. (I will return below to the question of why they will not.) But at some stage they will lurch from something like RC to some form of the reasoning just described. They will suddenly abandon plural thinking for singular.

What we have here is a special case of a very general phenomenon. There is nearly always a possibility of inconsistency between the results that a reasonable decision-making procedure gives when applied to a situation, and the results that it gives when applied to a sub-situation. (And in fact two reasons for this coincide in this as in many cases. The outcome of a procedure that takes account of our cognitive limitations when applied to a whole situation is likely to be incompatible with the outcome when applied to a sub-situation. And the outcome of a cooperation-producing procedure applied to a whole situation will also often be incompatible with its outcome for a sub-situation.) That is, if you reduce the number of actions open to the agent then what was the most acceptable choice for an agent may cease to be so.[3] As a result, if the situation develops in such a way that the options open to agents reduce with time, it is always likely that one or more agents will switch from the action specified by the procedure. And these switchings can be very much against their common interest. So they are faced with what is in effect a meta-coordination problem. Inasmuch as they must adopt mixtures of starting-off and continuing strategies they must adopt strategies which fit one another. Without much loss of generality we can take this as the problem of finding a strategy to start off with, so that when deviations from it occur they will be minimally damaging. And, the point that matters for this chapter: this problem cannot be solved by any cocognitive routine, which will give too definite expectations too far in advance. Some deeper plural thinking is called for.

This is what I shall call gentle approach concavity. 'Concavity' refers to the fact, shared with prisoner's dilemmas and similar situations, that there is an outcome which cannot be improved for any agent without spoiling the situation of another, but which at least one agent can improve on by unilaterally varying their choice. (So the Pareto-optimum is not an equilibrium.) And 'gentle approach' refers to the fact that the agents can arrange their choices in a sequence so that the hard concavity-facing choices are postponed. Then, as remarked above, it is in their interest to coordinate the postponement. But knowing when one another will break from the initial strategy undermines that strategy.[4] A procedure that gets the sequence going without giving too much definite information about the length of sequence might run as follows:

1 Choose an initial sequence of actions which satisfies RC for the first several choices.

1(a) (optional) Consider whether the other(s) is likely to have got the same result from (1). If not, exit.

2 Perform the first member of the sequence.

. . . .

(n) If the result of performing the (n–1)th member of the sequence is acceptable, perform it. If not, exit and choose by some other method.

Call this procedure VC because it forms a vague consensus about how to begin, which may then dissolve as the plot thickens. The range of initial sequences that might be chosen at stage (1) is often large. In the loonie game it might include: take one coin for the first five moves, take one coin for at least thirteen moves, or take one coin at move n with probability f(n). The first of these would presumably be ruled out by the optional step 1(a). This step is itself an anticipation of step (n), which asks the agent to consider what would happen if the agent continued with the sequence, in effect what the other agent would do at the following stage. There is a more and a less demanding way of understanding this. The less demanding way is to ask what would be the output of RC or some other plural cocognitive routine applied to the situation. The more demanding way is to apply a thicker simulation – a centered fundamental singular simulation – to ask what this particular person would do in the situation you are considering putting them in. (This is what must lie behind a hunch that the other's non-grabbing patience is about to run out. Or which might abort the whole program at stage 1a if one senses that the other is a game-theoretic Martian who will find it natural to grab at the first opportunity.) I return to the difference between these 'indicative' and 'subjunctive' readings in 'Learning to simulate' below.

VC gives good results for the loonie game. If a pair of agents are both employing VC then they will both choose some option such as 'wait until near the end till grabbing', since such options define later games in which the payoffs are higher. And then when playing the later game they will begin, at any rate, in accordance with such actions. (And then of course at some unpredictable move one will flip to grabbing mode.) And the result will be that each will be assured that grabbing will be delayed until well into the game, without having the nasty surprise of finding that the intended plan is not adhered to.

The relation between RC and VC is a special case of another very general phenomenon. Very often an intractable strategic situation can be tamed by embedding it in a larger context, in which the crucial decisions are transformed into a coordination problem. (Not necessarily an easy coordination problem, but by the nature of coordination problems not intrinsically intractable or mysterious.) The best known example of this is the embedding of the prisoner's dilemma in a series of iterated games, where agents have to choose a strategy for playing the whole series. (The classic work here is by Axelrod 1984 and Taylor 1987. See also Binmore 1998 for doubts about Axelrod's approach, and for a general attitude to the transformation of cooperation into coordination, and Morton 2001a for appreciation and doubts about Binmore's approach.) But VC does not provide the right embedding for prisoner's dilemmas and related situations. In these, abrupt onset concavity consists in the fact that there is no step by step approach to the crucial decisions. So there is no avoiding the fact that, in the prisoner's dilemma for example, each agent, if committed to cooperation, provides an easy target for defection by the other, and knows it. So the loonie game phenomenon of agents breaking from their chosen plans to act in their immediate interest can be expected to happen from the very beginning in such situations.

People often do choose the cooperative option in situations of abrupt concavity, naturally as part of their ordinary routines. (And very often too they do not.)[5] When they do it is often because the situation has been embedded in some larger one which transforms it in some crucial respect. There are many such possible embeddings, and not all of them have any special consequences for the procedures by which people choose and anticipate actions. One such embedding is, however, an easy progression from VC. To get a sense of it imagine that you know that you will soon find yourself in an unexpected prisoner's dilemma situation with another person, with whom you cannot communicate, for high stakes. If you both choose non-cooperatively (both defect) then the result is pretty bad for both, though not quite as bad as the outcome for the cooperator if the other defects. But if you both cooperate then both will come out acceptably. To make a clear separation of different components of the situation suppose that you are now at a stage – call it the deliberation stage – at which you know that the two of you will face a specific option later, at the choice stage, which will have the form of a prisoner's dilemma with these stakes, though you do not yet know what the options will be. You desperately want to find a route to cooperation. You know that the other will also want to find a route. Any route that gives either of you the confidence that the other will cooperate in a way that is separable from your own intention to cooperate will fail, as it will be available to the other as a ground for defection. What is needed is some way of mentally taking each other's hand and leaping into the deep end. That is, you need a way of choosing an act such that your knowledge of what is going through each other's minds first commits you to the cooperative act and only then informs you of the other's likely choice. That is far from impossible; all you need to do is to form an intention to choose in a particular way at stage two, which you will stick to if and only if you are confident the other will too. And there are simulation-and-choice procedures which deliver this. Here is one.

1 Consider procedures for choosing at the choice stage, and select one which will produce an acceptable outcome if employed by both parties.
2 Reproduce stage 1 from the point of view of the other.
3 If the result of stages 1 and 2 is an intention by both parties to use the selected procedure on the assumption that the other will, proceed to stage 4. If not, leave the procedure and choose in some other way.
4 At the choice stage choose from the choices that become available an act in accordance with the selected procedure.

Call this SN for 'simulated negotiation', for stage 2 amounts to an imagined conferring with the other resulting in an imagined promise. An example of a procedure that might be selected at stage 1 is one that chooses the act with the best outcome given that the other person has made the symmetrical choice, or one that chooses the act with the best average outcome for both people, or which discards a subset of options in order to transform the situation into a coordination problem. There are others, depending on the details of the situation. It is important to see that the confidence that the other will choose cooperatively cannot be based on testimony, induction from that person's past behavior, or a theory of their psychology. At least it cannot be based on any of these alone. For if that were its source, it would induce non-cooperation on one's own part. So two sided cooperation will not result from that kind of confidence.[6] The crucial fact about SN is that the expectation about the other is based on a reflection on one's own reasoning: if that reasoning then develops to give an intention to defect, the expectation of the other's behavior will also change. One either cooperates expecting the other to cooperate or defects expecting the other to defect: there is never any hope of being on the winning side of one-sided defection.

The simulation at stages 2 and 3 in SN must be a fundamental centered modeling. You must actually put yourself in the shoes of the other and consider what effect the deliberation process will have had. For no formal solution procedure that does not take into account the state of the particular person at the particular time will come up with a reliable answer.[7] The material for this simulation may be very varied. One can imagine it along the lines presented in the previous section: the two people have undergone similar experiences and internalized each other's reactions, so that they have come to be able to model each other's thinking.

To call a process fundamental centered modeling is not to say how it occurs. The label describes what happens, not how it does, and there must be many cognitive processes that produce the right kind of results. Many of them are best not given fake descriptions by philosophers. So it is comforting to see that we can turn back towards the ordinary here. In some cases the simulation needed to get agents through abrupt concavity is something extremely down to earth. One very down to earth factor is pair-bonding. Two individuals undergo a period, long or short, in which they come to identify with one another's projects, and then when the crunch comes, simply act in accordance with their mutual interest. Bonding is clearly a fundamental process in primate and other mammalian social life. So it frequently happens that something like SN occurs and the deliberation stage consists of any of a variety of shared experiences, which in members of your species will induce bonding. Then the transition to the choice stage will only occur if you know that bonding has occurred in both of you. The simplest way for this to occur is if bonding is intrinsically two-sided: if you are bonded to the other the other is bonded to you. In that case your knowledge that the other is bonded just consists in your intention to proceed to the second stage. In a more complex context bonding may often enough be one-sided that you need to perform a re-centered re-creation of the bonding from the other's point of view. If the outcome of that is positive, you proceed to the choice stage.

So, filling in some details, one way of instantiating SN which is clearly an instance of off-line fundamental simulation, proceeds as follows:

1 Consider – vividly – selecting a pattern of coordinated action which will be chosen by individuals acting as a coalition. Consider your grounds for taking the other as a reliable actor in a coalition with you.

2 Activate your bonding module taking as input your experiences in stage 1 taken from the other person's perspective.

3 If the result of stage 2 is that the other is bonded to you, and the result of stage 1 is that you are bonded to the other, proceed to stage 4. If not, leave the procedure and choose in some other way.

4 At the choice stage choose from among the options that become available the one that most represents acting in coalition with the other.

Call this SN via bonding. I should say why this is simulation. The reasons are very similar to those for plural cocognition. An individual undergoes a process that produces a plan of action for two people. As a result of forming this plan the individual has information about the likely actions of the other, which are reliable as long as the other is also forming a plan by a process whose workings are regularly correlated with it. (The individual has no knowledge of how the process works. The justification of beliefs about the other is, in the jargon of contemporary epistemology, firmly externalist.) So, uncerebral as it may seem, taking a deep breath, pausing, and then going ahead trusting your instinct that the other will not betray you because she is trusting her instinct that you will not betray her, can be a kind of simulation. It is that as long as what happens in the pause is the operation of a bonding mechanism which has as inputs facts about the environment, and whose giving the go-ahead signal is a reliable sign of the operation of a parallel mechanism in the other. You know what the other will do; you know why she will do it (because she trusts you); and you know these things because her trust has the same basis as yours. (The link between bonding and strategic thinking is discussed in Morton 2000, where I also discuss the interaction between innate and learned cooperation-inducing procedures.)

Both cocognition and fundamental modeling come in basic and sophisticated versions. With cocognition the basic version is the plural thinking described in Chapter 1 and the sophisticated version is the fine-tunings of this described in Chapter 2. With modeling one basic version is cooperation-via-bonding as just discussed. And sophisticated versions include the idiosyncratic grasp individuals have of other individuals they know well. To say this is obviously not to give any sort of psychological account of how these simulations work. It is, however, to allude to constraints that basic and sophisticated forms must satisfy. They must allow agents to choose actions that mesh with those of other agents, without generating the kinds of too-specific predictions that theories of rational action would. And in abruptly concave cases they must do so by generating intentions and expectations that are paradoxically inseparable – neither can exist without the other but if either is formed separately from the other it will undermine it. So these basic types of strategic choice force interacting agents to use routines for simultaneously choosing and anticipating which themselves mesh – if they differ the actions will not combine. And so agents are constrained to anticipate others' actions by running through routines that the others are also running through, and indeed to use similar routines to those the others are using. We have to coordinate our thinking in order to cooperate in our actions.

Learning to simulate

The procedures described in the previous section were not meant as literal candidates for processes whereby human individuals come to know what they and others will do. (They bear roughly the relation to real attribution and decision procedures that the Prolog programs described in Danielson 1992 do to moral deliberation.) No doubt we tackle strategic situations with an enormous variety of procedures. Some of them are even simpler than these three ('if it's a friend, cooperate; if an enemy, resist'), and others much more complicated. But any such procedure that is used for a range of situations must have important features in common with the three I described. In particular it must respect the limits of our thinking powers; it must not demand too much time or memory. And it must give good results when shared; if several people approach the same transaction using the same procedure they must do at least as well as they would had they been using different procedures. No limitation-respecting procedure will handle all strategic situations. So individuals living lives of any significant complexity will have to master a range of procedures. And they will have to acquire the right range, given the procedures that others are using. Here too, and perhaps more importantly than anywhere else, the need to be understood – just enough, not too accurately sometimes – creates a pressure to understand others by methods congruent with those they are using to understand you. (The methods will no doubt be very varied, and sometimes shockingly simple, like the inference and decision procedures studied in Gigerenzer *et al.* 2000.)

It follows that we learn to understand others, shaped by a core of robust innate routines and fine-tuned by the particular routines used by others around us. And if it is true that cognitively affordable procedures that handle strategic situations will tend to involve simulation, then it follows that we learn to simulate. That is, we learn to deploy a rich range of different kinds of simulation. In this section I shall describe an idealized path this learning could take. I describe a possible route by which some more deeply ingrained capacities can scaffold the development of some more delicate ones.

Begin with *conditional thinking*, which in its simplest form consists in forming action plans contingent on possible events. You are walking across the road and you decide that you will run *if* the bus leaves the stop before you reach the other side. Or you are wondering whether to cross the road and you realize that if you do and the bus leaves the stop you will have to run. Simple conditional thinking is required for all but the most primitive plans of action. (It is probably a fundamental frontal lobe function.) In more complicated cases one must explore the branches of a dense tree of possibilities, which ramify more thickly the further into the future one looks. From a logician's point of view ideal conditional thinking would consist in adding a supposition to your stock of beliefs, then removing and adding beliefs from this stock as required in order for the supposition to make sense, and then drawing consequences. The second stage is very problematic. No one knows the rules for doing it. (The question of what the rules are is often called 'Goodman's problem'. See Goodman 1973 and Lewis 1973. For a survey of recent work on conditionals see Edgington 1995.)

Take an imaginary person who has mastered this kind of plan-oriented conditional thinking to a high degree. She can make complex plans that include complex conditional sub-plans. But she is a human person, living with others, and as a result many of the conditions upon which her decisions depend are actions of others. All the difficulties of strategic choice arise. But there is a way of managing them with skills she already possesses. She can make complex plans, perhaps with complex conditional sub-plans, on behalf of those with whom she is interacting. When she can evaluate what would be a good outcome for a given group of people, whether it be in terms of what would be preferred by all or agreed to by all or what would simply be good, she can think through ways of achieving such an outcome. These plans will define actions to be performed by her and by the others. (So she will have in hand routines in the same category as RC. But see below for VC-type, subjunctive, aspects to her situation.)

She can use these plans to predict the actions of others, and to explain them when others act in accordance with the plans she would have prescribed. The predictions will only be accurate under conditions that may be hard for her to articulate, but let us suppose that they are right often enough to guide her through practical life. A basic tool for anticipating and understanding others will thus have been acquired. Conditional thinking plus a concept of the good for others yields plural cocognition.

Our conditional thinker's plan is supposed to tell her what to do, as well as to predict the actions of others. But the form of the prescription will often take some translation before it is usable as a guide to her action. For example the group plan may require that she walk down the alley on the north side of the main street. She is walking down the main street, and needing to know what to do. She has to be able to translate that prescription into 'turn left three steps ahead'. The individual action-descriptions need to be translated from objective to self-centered coordinates. In order to do this, she has to keep track of the changing relations between her own egocentric coordinate system and the objective system of reference to places and directions in terms of which the plan is formulated. But once she has this skill, it is a small step to applying it to keeping track of the relations between *other* people's egocentric coordinate systems and the objective references. The algorithms are the same; she just has to keep track of the data. As a result, she has acquired the capacity for a rudimentary form of centered simulation, in which she can recreate the actions of another person from that person's point of view.

(The ability to translate from objective to self-centered coordinates is needed in much purely individual conditional thinking, as well. A long-term plan is usually thought out in terms of objective descriptions of objects and possible developments. For the available information about the relevant states of the world beyond the immediate present is usually expressed in objective terms. But then when the plan is being acted on, the agent has to convert act descriptions in these terms into usable egocentric ones: she had always meant to run out the side entrance if the guards came in when she was defacing the manuscript, so when they do come in she has to find the way from the 'here' she is at to some 'there' which is the path to the side entrance. The conclusion here is similar to ideas in Bermúdez 1998. Bermúdez argues that information a subject gains by perception about the relation of the surrounding world to itself can usually be transformed into information about the relation of the relation of the self to the world. Knowledge of the environment generates one kind of knowledge of self.)

At this point the capacities that have developed from basic conditional thinking plus an ability to define a satisfactory outcome for a group of people can link with a more advanced conditional thinking. This is subjunctive or counterfactual thinking. From a first person future-directed point of view it is the difference between thinking 'if e occurs, I will do a', as part of a plan, and thinking 'if e occurs then I shall in fact do a' as a factual prediction of one's future behavior. The distinction is one that every decision maker has to heed. (You are going to a party where it may be hard to avoid a congratulatory glass of champagne. If this happens then the sensible thing to do is to take a taxi home. But you know that you would be more likely to drive. If the champagne is offered I shall not take it; but unfortunately if it is offered I will in fact take it. So it might be a good idea to take a taxi to the party, even though it is far from certain that the champagne will be unavoidable.)

Suppose now that our conditional thinker is making a plan, which at some point in the future will depend on the action another person will choose in some possible situation. (This time it is another person who may or may not resist the champagne if it is offered, and may or may not order a taxi.) She could predict the other's action by deriving it from the best group plan in the eventuality in question. But this is no more likely to be accurate than it is in her own case. (That is, it often is accurate, but there are circumstances, which need to be learned, in which it is not.) She needs to be able to say what someone would or might do if various possible events occur. She may do this by observation and inductive learning ('this is how people behave after a glass of champagne'). Or by application of some theory learned from others ('car salesmen often strike bad deals in order to have made one sale in a day'). Or by imagination: non-centered, fundamental, modeling. Each of these ways will give her a way of performing the transformation function needed for the CT routine of the previous section.

(If this general line is right, it links issues about cooperation to issues about dynamic choice. Agents' opportunities of cooperation are greatly increased if they can make plans based on reasoning about what they and others would do at later stages of those plans. The most influential work on this is Hammond 1976. Hammond is not an easy read. See also McLennen 1990 and Rabinowicz 1995.)

(Even in completely non-psychological cases it seems intuitively that when thinking 'what would happen if?' we often model the situation in an analog way, following the trajectories of mental representations of objects from imagined initial to projected final positions. Consider how one plans a speedy route through a crowd of moving people. This example suggests close links with the debate on mental images. There, too, what seems to some as intuitively analog and non-propositional thought can be described in terms of the manipulation of suitably chosen symbols. See Shepard 1982 and Tye 1995. It is very attractive to suppose that this thinking is also often plural. You learn what would happen if you and another try to talk without quarreling, or what would happen if two drunk people who are attracted to one another share a taxi, *without* first thinking out what the individual people would do in such circumstances. In Exploration II I argue in an impressionistic way for ubiquitous and fundamental 'interspection': that people are often more sure about what is true between them than about what is true of each of them.)

A crucial stage has now been passed. The transition is from simply being able to translate descriptions of acts and situations into person-centered form, to being able to simulate a counterfactual decision. This is like the transition from being able to take account of the visual and informational perspective of others to being able to ascribe false beliefs to oneself and others. (In fact, there are close connections between the capacity to think what one or another would have done and the capacity to ascribe false beliefs. See Gopnik 1993 and Morton 1993.) After the transition a much richer set of possible considerations opens up. Our conditional thinker may be said now to be capable of centered modeling. As a result, she will now be able to undertake the delicate thinking required to assess whether it makes sense to proceed into situations where cooperation is a delicate matter, such as those with the surface structure of the prisoner's dilemma. These are best avoided unless you know that you and the other person would behave in them.

In fact, the capacity for plural counterfactual thinking (whether or not it is irreducibly plural) can be applied in many other contexts too. In the previous section I noted them by inserting a 'subjunctive aspect' flag at several points. The topics in question were salience, acceptability and simulated negotiation. Being able to think in subjunctive or counterfactual terms what a person or group would do increases one's capacity to assess the salience of outcomes in a coordination problem. For one wants to know not just whether the outcome in question has features that are in another person's interest, but whether it will seem to that person to stand out from the alternatives, in a way that may be thoroughly arational and idiosyncratic. (Or, more subtly, whether it will seem to one person that an outcome will stand out for another: double idiosyncrasy.) Subjunctive thinking also helps assess the acceptability of possible outcomes in any situation in which compromise is required. For what one wants to know for some purposes is which outcomes would in fact be accepted by people when they occurred. (For other purposes it is more important to know what outcomes will seem more acceptable to people in advance. What they want before they get it. Indicative thinking is more effective for this task.) And the relevance of subjunctive thinking to simulated negotiation should be clear by now: you need to know what situation you would actually be in if after negotiation you proceeded to the transaction. None of these tasks has to be done by subjunctive thinking. The functions of simulated negotiation, in particular, can be performed by an enormous range of devices, from explicit theories of fairness to elemental bonding between individuals. But it certainly helps if you also have a quick and intuitive grasp of answers to 'what would follow if this is what we did?'

(It will be clear that I have not presented any model of modeling, as opposed to cocognition. Perhaps there is some standard form that centered modeling of basic states typically takes. If so, I doubt that the resources that I can deploy will uncover it. But it is also possible – likely, I would suggest – that there is no standard form. Modeling would be defined by the role it plays in filling the gaps in other forms of thought, notably those based on cocognition. Each person would learn their own ways of doing it, to the extent that they can, in ways that are not determined by innate routines or cued by socially transmitted lore. The effect of normal social life, itself shaped by innate mind-ascribing skills and by social lore, would be to put each individual in a situation in which it is very much in their interest to acquire as much modeling skill as they can.)

The development from non-centered cocognition to centered modeling is not a trivial one. It cannot be accomplished just by having the capacity for conditional thinking and then applying it mechanically. Essentially new elements are added at three stages: learning what counts as a good outcome for a group of people, learning to translate objective descriptions into person-centered descriptions, and learning to apply subjunctive conditionals (what she would do if) to oneself and others. And the end point of the development is still a long way short of the rich delicate centered modeling that we are on occasion capable of. *And* it is far from obvious that the development of richly psychological thinking in any actual human takes the course I have described, though some of the stages that result do correspond to stages in the development of a child's understanding of mind. What the possible development I have described does show is that the capacity for cocognitive non-centered simulation – which is potentially grounded in the capacity for conditional thinking – can be a framework on which more problematic modeling and centered simulations can grow.[8]

One last observation. The main thrust of this chapter has been to argue that we press one another into congruent ways of deciding and understanding. The argument has focussed on pressure to simulate in congruent ways.

Someone might argue that some or all of the simulatory routines I have described could be replaced with procedures that gave the same results but could not be described as simulation. I cannot imagine how this would work: I cannot see how, for example, deductions from innate or learned general principles could have the same effect. But suppose that I am wrong, and that there is some other family of cognitions that will do the jobs. The main argument is unchanged. There will be a contrast between the simpler and the more complex cognitions as described in this chapter; there will still be contrasts between cognitions along the cocognition/modeling, folk psychological/fundamental, and centered/non-centered dimensions, and there will still be a delicate fit between cognitive process and strategic situations. So whatever these cognitive processes are, if people are to use them in strategic situations then interacting people will have to use the same, or congruent, ones. And if they are to be usable by finite creatures such as ourselves then which members of the family are used depends on which strategic situations are encountered. We equip ourselves with the cooperative and defensive devices for understanding others that fit those that they are using to understand us.

APPENDIX: MAXIMIZATION AND COMPLEXITY

> The fascination with what is optimal in thought and behavior does reflect a certain sense of beauty and morality. Leibnitz's dream of a universal calculus exhibits the aesthetics and the moral virtue of this ideal, as does Laplace's omniscient superintelligence. Cognitive scientists, economists, and biologists have often chased after the same beautiful dreams by building elaborate models endowing organisms with unlimited abilities to know, memorize, and compute. These heavenly dreams, however, tend to evaporate when they encounter the physical and psychological realities of the waking world, mere mortal humans cannot hope to live up to these dreams, and instead appear irrational and dysfunctional when measured against their fantastic standards. On earth, heavenly dreams become nightmares.
>
> Gerd Gigerenzer 2000

Some arithmetic about solution concepts

It is inevitable that we use short cuts and heuristics for thinking through real-life choices in real time. Therefore we must anticipate the actions of others in a good enough way. How acute is the need for approximation: how complex could non-heuristic reasoning about real social situations be? Very.

It is easy to see that the examples in Chapter 1 of situations where agents have to consider each other's beliefs about each other's beliefs and desires are only the beginning of the possible complications. It is easy to produce two person situations in which the agents are choosing between n actions each and must consider n-deep embeddings of states in order to arrive at the unique solution. (For example, with such a 2-person 3-action case an agent has to consider what she thinks the other thinks she thinks.) Must, that is, if they are considering the problem exactly and completely, and if solving it is taken to mean finding one or more equilibria, that is, situations in which each person cannot do better for herself given the choices of the others. Normal subjects find four-deep embeddings difficult, but we routinely enter into situations in which each agent has a large number of available options. This strongly suggests that in everyday life we are in the midst of situations that if thought out in explicitly strategic terms would stretch our cognition beyond its limits. (See Russell 1997 and Kinderman, Dunbar and Bentall 1998, on the limits of our capacity to comprehend embedded attributions. Kinderman *et al.* find that there is a sharp threshold between four and five embeddings, so that very few five-embedded sentences are understood.)

And those were in a way particularly simple cases: only two agents and only one equilibrium, only one outcome which it is in each agent's interest to go for given that the other is going for it. If we have many agents and many equilibria things are much more complicated. Game theorists have studied the complexity of the n-person case, partly because they want to create software that will compute the equilibrium outcomes for arbitrary games. There is a very technical literature here, which has not yet crystallized into a few central results, but the drift is clear. As the number of agents and the number of options open to each of them increases, the number of equilibria grows rapidly, in the worst case exponentially. If one chooses situations at 'random' then (given suitable definitions) the number of equilibria of an average game increases exponentially with the number of options. Moreover, the computation of equilibria must be complex: the computation time taken to find the equilibria of a game, with the algorithms presently available, also increases exponentially with the number of players and options. (See McKelvey and McLennan 1996, 1997 and McLennan 1997. These results are for Nash equilibria of mixed strategies. It doesn't look as if the general character of the results will change for different equilibrium concepts, except that the calculation of a subgame perfect equilibrium must be a very demanding business. See also Chapter 10 of Rubinstein 1998.)

One fact underlying these results is very simple: the number of cases that have to be searched before one can know the equilibria of a situation can be very large. If there are n agents and each has m possible actions then there are m^n outcomes. (Not mn: with n agents the matrix of the game is a n-dimensional cube with m strata on each face.) For each outcome one will have to check whether some agent is better off if they alone had chosen differently. This means checking nm possibilities, so a total of $m^n(nm) = nm^{n+1}$ outcomes may have to be checked. This is not a practical way to make a decision.

So in a nutshell the situation is this: if you make strategic decisions with official methods you have two choices. You can reason in terms of the agents' motives, which will involve enormously complicated embeddings of beliefs and desires. Or you can search through the matrix of outcomes for equilibria, which will involve considering an enormous number of cases. In either case you are likely very often to surpass the limitations of human memory and reasoning power.

These complexity calculations have been based on an assumption, though. The solution to a strategic situation has been assumed to be an equilibrium. An equilibrium is, in effect, what a completely self-interested agent who completely understands the situation will go for as the best he can get, given that all the other agents are completely self-interested and completely understand the situation. These are not realistic assumptions. See what happens when we simultaneously let go of the idea that the agents are completely self-interested and that they aspire to complete understanding of the situation. Suppose that, given the complexity of the situation and the need to come to terms with it in a limited time, they are instead looking for 'good enough' outcomes. And suppose their conception of a good enough outcome is one in which they do not do badly, and which is likely to be arrived at by others because they too do not do badly.

As a first stab at such a conception suppose that we read this as taking Pareto-optimal outcomes to be acceptable solutions to strategic problems. (Remember that an outcome is Pareto-optimal when there is no other outcome which is more desirable for all agents than it.) These are not so difficult to find. The reason that when looking for equilibria we must check through nm^{n+1} outcomes is that for each candidate outcome we must check for each agent whether some other action of that agent might have led to a better outcome, holding the actions of the other agents constant. (So for *each* outcome O we must check for *each* agent whether the outcome O' consequent on *each* other action is better than O: triple *each*.) Now suppose that we are instead looking for Pareto-optimal outcomes. At first sight it might seem that the situation is parallel: for each outcome we have to check whether there is another that is better for some agent than it is. Again triple *each*: for each outcome O we must check for each agent and each action, whether the resulting O' is better than O for some agent. But this is a grossly inefficient way of running through the possibilities, and results in making the same comparisons many times. Here is a much more efficient way:

> Go through the set of outcomes in any linear sequence, collecting candidate Pareto equilibria, that is, outcomes which are at least as good for some agent as any outcomes considered yet. When considering each next outcome discard all candidates to which it is Pareto-superior. When you have gone through the entire set of outcomes your set of candidates will be the Pareto optimal outcomes of the situation.

The number of outcomes that will have to be compared in this procedure will be at most the number of comparisons of each outcome with all previously considered ones, that is $\sum i$ ($i = 1, 2, 3, .., m^n$) = $m^n(m^n + 1)/2$, which is considerably less than nm^{n+1}. In fact the upper bound is considerably lower than this, since one does not have to check an outcome against all previous ones to know that it can be eliminated as a candidate.

It is not easy to see an analogous shortcut for finding Nash equilibria. The reason is that to determine whether an outcome is a candidate Pareto optimum one needs only compare with other candidates, while to determine whether an outcome is an equilibrium one needs to compare it with all the other relevant outcomes. A proof of the nonexistence of a more efficient algorithm would be very nice: but proofs that algorithms are optimal are notoriously hard to get. So I shall rest content with what is clear, that there is a procedure for finding Pareto-optima which is less expensive than the obvious procedure for finding Nash equilibria. And on it I rest the natural conjecture, that this is a difference between the best procedures in each case.

Pareto-equilibria were chosen simply as an example of a good enough outcome. Others are even simpler to calculate. For example, suppose that a satisfactory outcome is one where the benefit for all concerned is above a given threshold. Then one only has to consider the m^n outcomes, comparing each to the threshold, and one will typically have to stop searching long before all have been considered. All these solution concepts are symmetric between agents, in the sense that if interacting agents use them on the basis of the same information they will arrive at the same solution. As a result each agent's decision is not undermined by considerations about what other agents may decide by thinking along the same lines.

Though the argument is based on a pretty crude analysis of conceptual effort it raises an important possibility. The examples of good enough thinking we have considered have tended towards collective rationality. In many situations if all interacting agents use them then they will be better off than if they think in terms of equilibria. Combine this fact with the fact that they are cognitively less demanding than equilibrium thinking and a thought is almost inevitable: people may be pushed towards collective rationality by the difficulty of thinking in more individualistic ways. Or we can emphasize another side of the thought. Suppose a person decides to think through her decisions in a way that would be generally beneficial if those she is interacting with use similar ways of thinking but which would have bad results if some others reasoned in a more individual way. Then some assurance that they will not do so is afforded by the greater effort required to do so. (For satisficing, that is, thinking in terms of good enough, see Simon 1982, Slote 1989, Chapters 1 and 11 of Rubinstein 1998.)

This must be a vague and tentative conclusion. There will be many exceptions. It is based on many simplifying assumptions. Some of them are standard in game theory, but no less unrealistic for that. The conclusion would be undermined if there were reasons to believe that describing choice in a more realistic way would diminish the difference between equilibrium and good enough thinking. But there is no reason to believe this. In fact, in real life they may be even further apart. For one thing, the strategic complexity of real life is potentially greater than game theory allows for. For example, we care not only about outcomes but also about the reasons for which they come about, so that our valuations of them are likely to change as we deliberate and come to conclusions about the motives that will move others. This is a complication that is completely ignored in the standard theory.

Good enough thinking can be taken to have advantages, then. But good enough thinking and plural simulation are closely related. Good enough thinking is the special case of plural simulation in which the outcomes are evaluated in terms of thresholds. So some kinds of plural simulation are candidates for efficient ways for thinking creatures to anticipate one another's actions.

Recursion and maximization

This section makes a very speculative point. Even its relevance to the main issues is speculative. It is definitely skippable, but it is also interesting.

Many computationally hard problems ask for a maximization or a minimization. What is the quickest, shortest, least, greatest? Many of these questions resist the search for efficient algorithms. Consider an example, which brings out a frequent feature. We have a number of points on a plane and we are trying to join them with straight lines in such a way that the total length of the lines is as small as possible. We must join each point directly or indirectly to each other point, but we are allowed to add additional points. This is a problem with clear practical applications, for example in setting up a telephone network. Compare the solutions for four points and for five.

If we have four points arranged as vertices of a square the shortest network is got by adding a fifth additional point in the center, like a five at cards, and then joining each point to each other in the obvious way. Now ask the same questions of the same four points and a fifth non-additional point. Depending on where this new point is, the additional point added to the original four may not be needed. In general, the pattern when a (n + 1)th point is added may be very different from the pattern for some given n points. So an algorithm that solves the problem for n may not work for n+1. This makes the general case very hard to solve. The inductive approach, reducing the n+1 case to the n case, does not apply. But maximization and minimization problems are very often like this: the solution for n and n+1 are different. So there is a general danger: extremal problems often resist recursive thinking. (See Garey and Johnson 1979 and Chapters 30, 38 and 50 of Dewdney 1989. Very often the difficulties for recursive calculation I have been describing can be avoided by an analog computation. Dewdney conjectures that this may not be so for the most difficult problems.)

How does this connect with the issues that concern us? Official game theory situates an agent in a situation with a definite number of other agents, each of whom has a definite number of options. It then expects each of them to find the very best solution to the problem of maximizing the satisfaction of their preferences given that each other agent will also find their very best solution. The theme of this chapter is the difficulty of meeting that expectation. But there is also the difficulty of approximating to it. Suppose that you have a good approximate solution to a particular strategic situation involving four people, each of whom has three choices. You are then faced with a new but similar situation involving five people, each of whom has four choices. Can you apply or adapt your heuristic? The considerations above suggest that you should not be optimistic. Maximization problems do not pass easily from n to n+1; the slightly augmented case may be very different in character. So if your aim is to have approximations to the exact game theoretical solutions to your problems, you may find that you need to discover a large inventory of routines, different ones for even slightly different situations.

Procedures that are easier to calculate present less need for approximation, and so present this danger less. So we might take the worry to be directed at the attempt to find equilibria and not at the search for good enough solutions. That might be too hasty, though. Procedures such as the Pareto-finding rule of the previous section are still quite expensive. Approximations are still called for. There are reasons to think these approximations could be more tractable. For one thing, the searches required by these rules less often need to look through all of any set of outcomes. That is, the characteristics that make an outcome significant for the procedure are more an intrinsic feature of the outcome and less a comparison with a large set of others. For another, the use of thresholds will often curtail a long search. And being on the right side of a threshold is an absolute characteristic, not one which can be true relative to one set of comparators and false relative to a larger one.

Common sense ways of thinking are not very sensitive to the number of agents involved or the number of acts each has. The reason is the form solution-thinking takes. We think ‘what might happen?’ and identify the salient outcomes, not labeling them as results of combinations of actions, but as physical situations bearing good or evil for the people involved. Then we think back to what combinations of actions could result in these outcomes. Up to this point, we may not even have determined how many people are involved and precisely what the options open to each of them are. In fact, we may never do this, since all we are interested in are combinations of acts that can lead to outcomes we have identified as worth thinking about. As a result, the transition from n to n+1, for agents and acts, is not a particular headache for common sense ways of thinking. The hard work is done in isolating the salient outcomes. And this is where common sense thinking can miss opportunities that a finer-grained analysis can find. No doubt the routines that we use for finding salient outcomes depend primarily on classifications of situations. We ask: is this a matter of division, shared effort, coordination, or what? Depending on the answer we look for different opportunities and dangers. But in each case our routines are fairly insensitive to the numbers involved. (These middle-grained classifications of strategic situations are discussed further in Morton 2001b.) Good enough thinking is like common sense in crucial respects here: it identifies significant outcomes in terms that are relatively intrinsic to them. (This is especially true of the threshold-based versions.) So it too may be expected to be insensitive to numbers.

The most frequent word in this section was ‘might’. Approximations to precise strategic thinking might be inconveniently sensitive to the numbers of agents and acts. Approximations to good enough strategic thinking might not be, and might share the robustness of common sense in this regard.

Summary

Where we've got to

The preceding five chapters have presented arguments for, as the preface put it, beneficial circularities between our capacities to attribute states of mind and our capacities to engage in shared activities. (It is hard to find a simple accurate label for the second of these. Coordination and cooperation have acquired special meanings, acting for mutual gain excludes situation where one is acting to minimize damage. I'd like to be allowed just to call this domain 'ethics'. I mean all situations in which it makes a difference to each agent what each other agent does.) There is a stronger and weaker form for these circularities. In the stronger form the attributive capacity works in part – never more than in part – because it is shared by the people whose actions it explains. The central example of this is solution thinking or plural cocognition (Chapters 1 and 5). A more complex example is the advantage, as described in Chapter 5, of using simulation routines that are congruent with those used by those with whom one is interacting.

In the weaker form the attributive capacity is part of an environment in which there is pressure on people to behave in ways to which the capacity applies. It pays to act intelligibly. Examples of this are the situations in which it is in an agent's interest to act in ways that relate to basic parameters of the attributions made by others: the problem categories that define virtues and vices (Chapter 2), the tunings of belief (Chapter 3), and the contrasts between attractors (Chapter 4). In all these cases shared projects and explanatory accuracy are both best served when the terms in which agents formulate and choose their actions are the same as those in which they explain those of others. In fact, agents in these cases do not just think of their actions in the same terms as they describe those of others; they act as they do in part because others around them are explaining as they do. In all these cases of coordination between explanation and action there is some element of the stronger form too. For action in a strategic situation is explained in part by the expectations of the agents about one another's actions, and these are shaped by the ways they have tuned the parameters in question. So wherever there are tunable parameters – as with practical reasoning, belief and the evolution of desire – there will be situations in which part of the reason that the expectations are true is that the agents have congruent tunings. Chapters 2 to 5 all contain examples of this.

I take it that this is a fundamental aspect of our everyday concept of mind. Our understanding of mind and action is, in part, shaped by its need to mediate shared activity, just as the shared activities we undertake are shaped by the need to rely on our capacities to gather and conceptualize information about one another. Several of the more interesting consequences that flow from this lead directly to hard questions, to which no one knows the answers. The remaining four items in the book engage with these questions. They do not present sharp arguments for definite conclusions. They describe possibilities, and reasons why we might take these possibilities seriously. To mark the difference between these items and the main body of the book I call them not chapters but explorations. Let me briefly list these exploration-inviting consequences. They do not presuppose one another, though there are connections between them, so you can choose which ones might interest you.

I Attribution biases and the statistics of cooperation

Strategic reasoning and mind-attribution are both specifically directed at other humans. Yet prediction and explanation are not. We predict and explain all manner of things. Moreover, in Chapter 5 the simulation of other people's thinking was linked to conditional thinking, itself an all-purpose tool. How specific to our thinking about other humans are the most fundamental factors on which the attribution-cooperation circularity depends? In the first exploration I investigate a very small corner of this question by wondering how some statistics-digesting processes that are prevalent when the data are social are related to more general statistical short-cuts. One possibility explored here is that the social reasoning in question is got by tweaking a more general pattern. Another is the opposite, that patterns of thinking that have specific advantages in social situations are also used in more general contexts. To investigate this I consider whether there can be an advantage in strategic contexts to information-managing heuristics that then generate some standard statistical fallacies.

II Interspection and expression

In attributing beliefs, desires, capacities, states of character, and other explanatory concepts we are beginning down a long road that leads to increasingly deeper appreciation of what it is to be a human person. And in progressing from plural cocognition to singular centered modeling we are moving further in the direction of appreciating what it is like to be a particular individual person. How far down this road does the link between understanding and cooperation hold? I explore this question by suggesting one way in which a richer grasp of what it is to be a particular person at a particular time – a conception of subjectivity – can be based on simulation procedures described in Chapter 5, whose rationale is essentially social. This exploration centers on the concept of 'interspection', of knowing what attributions to make to interacting people collectively, as a basis for later and more conjectural attributions to them individually.

III Ethos

Crucial to the argument in the five core chapters was the fact that important aspects of our understanding of others can be fine-tuned so that they are in harmony with the fine-tuning employed by those we are understanding and interacting with. It follows that there is room for a plurality of possible folk psychologies, given different combinations of tunings. These will correspond to different styles of strategic interaction. How large is the possible variation? I explore this by considering some imaginary combinations of explanation and evaluation. I call a stable complex of ideas about what is intelligible and ideas about what is valuable an *ethos*. I describe four invented ethos, which are abstracted from real human forms of life.

IV Moral progress

Learning to use a set of psychological ideas is, according to what I have been arguing, a part of learning to achieve good results in cooperation with others. It thus takes one in a direction in which the central moral concepts, such as those of obligation, duty, or rights, can arise. One might thus suspect that the more advanced levels of folk psychology, such as those concerned with intention, self-knowledge, varieties of desire, and the will, are also deeply interdependent with moral ideas. More complex and sensitive moral thinking might require more complex psychological thinking. And then, if there is anything to the ideas explored in III, different versions of folk psychology would be in symbiosis with different patterns of moral thinking. What would the ideal combination look like? Rather than tackle this directly I ask whether some progress in moral thinking has to be accompanied by changes in its background of psychological understanding.

Exploration I

Attribution biases and the statistics of cooperation

The problem, if it is one

When we attribute states of mind – particularly mood and character – to people in order to explain their actions, and when we use these attributions to predict what people will do, we are subject to biases which systematically distort our expectations. This is, at any rate, the conclusion of an influential tradition in social psychology. Work in this tradition, notably that of Nisbett and Ross, makes an extremely convincing case for thinking that there are patterns of false beliefs that permeate our thinking about one another's likely actions. (See Nisbett and Ross 1980, 1991.)

These conclusions are very relevant to a study of folk psychology. If folk psychology is a first attempt at science, aiming for true explanation and accurate prediction, then it would seem that it must be a pretty crude and unsuccessful attempt. And if folk psychology is not crude science but rather, as I believe, an instrument for social interaction, helping us live worthwhile lives together, then it would seem to be a flawed instrument, incorporating mistakes about what we are likely to do. A wrench with a twisted handle. Either way, there is a threat. And a puzzle: how could something we depend on for central aspects of our lives, drawing on psychological capacities which have evolved in response to our need for accurate expectations about one another, embody systematic distortions of those expectations?

I shall argue that some attribution biases can help the ultimate aim of attribution, which is, according to me, (of course, of course) coordinating with your friends, not being tricked by your enemies and telling the one from the other. This conclusion is of the same general character as that of Chapters 1 and 4: folk psychology is not best appreciated as a device for producing all-purpose causal explanations and predictions. It has its own aims, and is more effective at accomplishing them. First we must learn more about the biases.

The fundamental attribution bias

People are often very wrong about one another. For example we overestimate the likelihood that a person who has told a lie one day will lie in a similar context the next day, and we underestimate the likelihood that a social situation will coerce people into doing repugnant acts. (This latter is of course the famous Milgram experiment, in which subjects administer what they take to be severe electrical shock rather than defy a bossy and manipulative experimenter.) We are almost completely ignorant of the extent to which our judgments on many topics, from matters of pure opinion to simple perceptual judgment, are shaped by our awareness of the opinions of others. These effects have been intensively and rigorously studied, in a tradition of social psychology over the past forty years. (The classic piece is Nisbett and Wilson 1977. For philosophical context see Kornblith 1998.) Though it is almost beyond dispute that there are such effects in a variety of experimental and real-life situations, the general pattern of bias lying behind the particular effects is inevitably a matter for theory. Another topic for theory is the cognitive processes involved when people explain actions which conform to patterns about which they are ignorant. It seems likely that to some extent people simply confabulate, making up plausible stories to fit their own behavior and imbuing them with the flavor of first-person authority. We tell stories to fill out the gaps between the fragments of knowledge that our attributive and coordinative routines give us.

An influential line of theorizing points to one fundamental pattern in the data. (There may be other patterns too, of course.) That is what Nisbett and Ross call the ‘fundamental attribution error’. It is based on a very simple and general proposition: people assign to one another more definite individual propensities to specific types of behavior, independent of the effects of the environment, than is in fact the case or is justified by the evidence available to them. To put it very crudely, people take people to have more vivid and fixed characters than they in fact do. To put the idea more precisely, consider a population of individual people and some specific kind of behavior that they can perform to a greater or lesser degree (such as answering a question truthfully or obeying a rule.) There will be an average degree to which the behavior is performed, and a variation around this average in each individual’s behavior. Suppose that the average is roughly constant from one interval of time, long enough for all individuals to have contributed to it, to another. Then considering a series of moments in time we will have a long-term average A over all individuals and all times, and an average a_i for each person. There will also be an overall variance V of the distribution over all people and all occasions and an individual variance v_i for each person over all occasions. If the population were completely homogeneous in nature and subject to exactly the same environmental influences then A and the a_i would be very near to one another. The population is typically not homogeneous and the environment acts differently on different individuals, so that the question of the difference between A and the a_i and between V and the v_i arises and has practical importance. The fundamental attribution error consists in a tendency to exaggerate the correlation between the behavior a person exhibits at different times and thus to overestimate the difference between the a_i and A, or equivalently to underestimate the size of the v_i.

An analogy: it is as if one had a collection of fairly similar and fairly fair coins. The 'behavior' of a coin is the number of heads minus the number of tails that results from tossing it a given number of times in an hour. The coins are tossed for an hour and it emerges that the average result is rarely far from zero, but that usually there is considerable variation in the results, in that some scores are well above or well below zero. The analog of the fundamental attribution error would be to think that most of this variation was due to the fact that the long term average of some coins differed considerably from that of the rest. In effect, it is to think that most coins maintain a rough equality of heads and tails while coins in a smaller subset of particular individuals are biased either to heads or to tails. It would be to give coins definite characters of Fair, H-biased, and T-biased.

The hypothesis that such a bias operates in our expectations about our own and other people's behavior, can explain a number of systematic errors in our attributions and predictions. And in fact there is very direct evidence for it, in experiments eliciting people's expectations about the frequency with which particular others will perform particular acts. (For details see Nisbett and Ross 1991, Chapters 1 and 4.) Besides the direct experimental evidence for the hypothesis, and its general explanatory force, an additional reason for taking it seriously is its congruence with another large class of systematic human failings. For the general bias is clearly very similar to the well-known base-rate fallacy, our tendency when assessing the probability that a given A is a B to give more weight to the proportion of Bs that are As than to the proportion of objects in the population that are B. The attribution bias tends us to ask 'what proportion of this person's Actions exhibit B?' in a way that blinds us to the relevance of 'what proportion of individuals exhibit B?' It is essentially the same pattern. (For a simple exposition of the base rate fallacy see Chapter 10 of Morton 2002. For more context see Stein 1996, especially Chapters 3 and 7.)

In fact, we can bring a third major cognitive bias into the picture. The Gambler's Fallacy is the tendency to think that the more often in a series of independent repetitions of a process which can have a number of random outcomes one of those outcomes occurs, the less likely it is to occur as the series is continued. It thus tends us to think that a coin that comes down heads seven times in a row is more likely to come down tails the next time, or that a tree once struck by lightning is unlikely to be struck again. Fortunes and lives have been lost. There is a close connection with the attribution bias, which goes back to the tendency to trace the variance in a population to the presence of individuals with differing less-varying behavior. So with a population of coins we tend to think that there are numbers of heads- and tails-biased individuals as well as fair individuals, in order to make sense of the fact that the heads/tails average is sometimes far from unity. Consider the 'fair' individuals on this understanding. They cannot have the variance in their behavior that the coins of classical probability theory will have; they must stick more closely to a heads/tails parity, if their average combined with that of the biased individuals is to come out to the observed value. But as a matter of fact there can only be such individuals if for them successive trials are not independent. Or, to put it differently, if a string of Heads is strong evidence that the following toss will yield a Tails.

So if our thinking analyzes statistics in a way that will produce the fundamental attribution error it will hypothesize the individuals presupposed by the Gambler's Fallacy. If this explanation is right then the Gambler's Fallacy should be more tempting when the sequence of unlikely events is less than some threshold. Above that threshold the subject would have concluded that the object or process is biased, and would expect *longer* runs of events than is probabilistically correct. So my account predicts that at some point there will be an abrupt transition from expecting too few 'heads' to expecting too many.

Profiles

The attribution bias, the base-rate fallacy, and the gambler's fallacy, can be brought into a single pattern. The suggestion that they – or some instances of them – are the result of a single type of reasoning is of course speculative. But there is a fairly simple thread connecting the three. It is the disposition to classify things into classes on the basis of whether they satisfy the defining characteristics of those classes. Or put differently, the desire to divide the world up into kinds of things with stable properties. (That this is a basic and general tendency of our thinking is argued in Keil 1989.)

One is facing a question about the future behavior of an individual I, who may have exhibited behavior β on occasions where β was possible. One proceeds as follows:

step 1: construct a profile of an individual who exhibits β
step 2: match I against the profile. Record fit or non-fit.
step 3: if I fits, classify I as β-biased.

To fully specify the procedure we need to say how profiles are constructed and how individuals are matched against them. But first notice how the three 'fallacies' can fit the procedure. Assume that in attribution bias cases the profile is 'often exhibits β'. Then when an individual has often exhibited β one classifies it as β-prone and predicts more β. Assume that in base-rate fallacy cases the profile is 'has property A'. Then when an individual has property A one will classify it as β-prone and thus as falling into some class B. (Ignoring the number of individuals with A who are not B.) Assume that in gambler's fallacy cases the profile is 'biased towards β'. Then when an individual is not biased one will classify it as not β-prone and not expect β.

The attribution bias and the gambler's fallacy are apparent opposites, since the first leads to estimates of values that are higher than the statistical data suggests and the latter to expectations that are lower. (The bias leads us to too high an expectation that someone will lie in the future if she has in the past; the fallacy leads us to too low an expectation that the coin will come down heads yet another time.) So it is worth spelling out the construction of profiles to show how both can be derived from the same model.

For the attribution fallacy, suppose that I has exhibited behavior β a proportion p of the occasions where β was possible. Construct three profiles: β-bias if p is greater than or equal to a contextually set threshold θ, $0.5 < \theta <1$, β-contrabias if p is less than or equal to another threshold θ^*, $0.5 > \theta^* > 0$, and β-neutrality otherwise. (It would be natural to take θ^* as $1–\theta$.) Then if the question is 'will A exhibit β?' one answers Yes if one has attributed β-bias, No if β-contrabias, and 'Don't know' if neutrality. If the question is 'how likely is it that A will exhibit β?' one chooses an answer q between θ and 1 if one has attributed β-bias, between θ^* and 0 if one has attributed contrabias, and between θ and θ^* if one has attributed neutrality. (Just to be pedantic: if bias then $\theta \leq q \leq 1$; if contrabias then $0 \leq q \leq \theta^*$; if neutrality then $\theta^* < q < \theta$.)

(What is the point of 'forgetting' the fine structure of the data once one has coded it as bias? It saves memory and thinking space, for one thing. Suppose that you meet twelve people in one day and two weeks later have to assign them to tasks. When we were evolving we didn't have notebooks.)

This routine predicts the attribution error in that if q is picked from its permitted interval randomly then q will be above p more often than not when p is greater than θ, and below p more often than not when p is less than θ^*. (And even for some non-random selection algorithms, which keep q closer to the thresholds, this will be true.) The attribution error will be an error to the extent that q is different from the true probability that I exhibits β; this will depend on the underlying probabilistic set-up. But on most set-ups if there is any random element in the determination of I's behavior, even estimating q as p would be erroneous, so that an algorithm that set q near to the relevant threshold will produce over- and under-estimates as long as the sample from which B is taken has a mean around 0.5.

To derive the gambler's fallacy we must suppose that we have certain information that I's long-term actual or potential behavior is predominantly β (e.g. the coin is fair, or the lightning-producing process is random), together with short-term not-β behavior. Then we construct two profiles: β-tending if I's past and future behavior is predominantly β, and β-avoiding if I's past and future behavior is predominantly not-β. So at step 2 when we match I against the profiles we find that it does not match β-avoiding, since its long-term behavior is dominated by β. So A is β-tending. But future non-β behavior would be inconsistent with this profile, so we conclude that β is to be expected.

If this explanation is right then the contrast between the attribution fallacy and the gambler's fallacy is not intrinsically one between a routine that is employed when thinking about people and their dispositions and a routine that is employed when thinking about random processes. It thus bears on the deep and important question of how human-specific our cognitive routines in social situations are. (The connections could go either way, though. The attribution error could be based on a general profiling heuristic that was fundamentally non-social. Or that heuristic could be an extension of social thinking to non-social cases: in the gambler's fallacy we might be thinking of coins as trying to live up to their 'characters' as fair or biased.) If the explanation is right, then it ought to be possible to elicit gambler's fallacy-like counter-inductive expectations in social situations with questions that elicit suitable profiles. Suppose, for example, that subjects are told that a person A is an unbiased distributor of presents. Every day she gives a present to one of two children and distributes them equally between the two. In the past few days she has given presents only to Amie: will Amie or Billy get the next present? It seems likely to me that, contrary to the social psychology textbooks, people will expect that it is Billy's lucky day.

Why we might be this way

Here are two ways in which people might benefit from being subject to the general attribution error. (They are not completely different ways: resemblances later.)

Enforcing norms

Imagine a group of people interacting in some repeated situation. Suppose the situation to have the structure of a two-person two-option coordination problem. Each agent knows that if the other chooses one option, the 'better' one, they are best off also choosing it, and if the other chooses the other, 'worse', option they are best off also choosing it, but that both people are best off if both choose the better option. Suppose that although people generally choose the better option they have a hard-to-predict tendency to choose the worse option. (Perhaps they don't like to see the other get the good result, perhaps they tend to suspect that the other suspects that they will choose the worse option.) Then that tendency can spread, from people's fear that the other will choose the worse option. Suppose that when each pair of people interact each knows the behavior of the other in some few previous interactions with that partner or others.

In order to know whether she should choose the better option, a person needs to know how likely it is that the other will also choose it. That is easy if the other has usually chosen the better recently and if people have generally been choosing the better. But suppose that this is not the case. Suppose that people have been choosing the worse often enough to give serious pause to someone considering choosing the better. And suppose that the particular partner on this occasion has been choosing the better recently. Then if the person appreciated the tendency of all people to choose the worse she would too very likely choose it. That is, she would if she thought things out in the statistical terms with which I have been describing them. Things look different if she thinks in the terms suggested by the fundamental attribution bias. Then she thinks along these lines: there is a certain amount of worse-choosing about; it must be due to there being some inveterate worse-choosers among us; but this partner has been choosing better recently; so I'm safe choosing the better.

The result will be that in people with this typical attitude to the statistics, the point at which a tendency to choose the worse leads to a general collapse of the better-choosing outcome is shifted. The better-choosing convention becomes more stable, and less likely to be subverted by fluctuations in the numbers of people choosing the worse. Everyone is likely to gain, as double better-choosing, which is the best combination for both choosers, becomes more likely.

There is a darker side, of course. If the established pattern is worse-choosing, such reasoning will make that convention more stable too, making it harder to break through to a better one. For someone choosing in the context of a general swing towards more better choices but a partner who has happened to make worse choices recently will reason: there are some better-choosers around, I see, but this person is evidently not one of them. It is established patterns of choice that become more stable, whether they be for better or worse.

This example was of a coordination problem. With small changes the same considerations apply to other situations, notably to repeated prisoner's dilemmas as played by people with a tendency to conditional cooperation, but also a perhaps inexplicable tendency to defection. (Perhaps they are easily tempted by immediate gains; perhaps they just make mistakes.) They will cooperate as long as they can believe that the other person will too. Then if they are subject to the attribution bias they will often reason: this person has defected recently, and there is some tendency to defection around, so I suppose she is an inveterate defector, so I shall not cooperate with her. The result will be that the other person's recent behavior is taken as a reliable guide to their behavior in the current round. And given conditional cooperation, this will mean that one will defect if defection is predominant in the other person's record (and cooperate otherwise.) Thus conditional cooperation is pushed in the direction of Tit for Tat: the rule 'cooperate if and only if you expect the other to', generates the rule 'cooperate with recent cooperators, defect with the rest'. Defectors will be punished, and an added motive for cooperation is created.

Conservatism, optimism and pessimism that go beyond the evidence, a tendency to vindictiveness. These are familiar features of human life. The attribution bias partly explains them. And in a very general way this is a point in its favor, a reason for thinking that it does represent a deep feature of the way people think about people.

Self-fulfilling expectations

Imagine a crowd being addressed by a despot. (The example, taken from Russell Hardin 1995, is from one version of the fall of the Rumanian dictator Ceausescu.) If any individual reacts with hostility they will be singled out and dealt with. But if many do so, the thugs supporting the despot will be powerless, they will in fact realize that it is in their own interest to lose their uniforms in a hurry. You are in the crowd and you have an inclination to shout your feelings. But you have more sense than to do it unless many others do too. You look around and see how much muttering against the regime there is. Unless there is a lot you are going to be pretty careful. Perhaps a revolt does begin and the crowd surges forward in fury. You'll join it unless you suspect it might suddenly collapse. So you look carefully for people holding hats and newspapers between their faces and the cameras of the security people. Too many of them and you'll quietly slip away.

The attribution bias has a clear effect on such motives. If you hear less than a certain amount of muttering when the uprising is still just possible you will attribute it to a small number of malcontents and you will not show any visible discontent yourself. If after a riot has begun you see less than a certain amount of face-hiding is happening you conclude that it is all due to a few cowardly individuals who are taking good care not to be identified. You don't conclude that everyone is a little bit wary. If you did, and if most people did, the riot would collapse.

Though the effect here concerns a person's estimate of the number of people doing something, rather than the probability that a specific person will act in a specific way, the consequences are similar. Thresholds for coordinated behavior are shifted, so that established patterns become more stable. And, more generally, ambiguous middle grounds tend to lessen. An example of this more general consequence is the appropriate reaction to testimony. Everyone is mistaken some of the time, and everyone tells the occasional lie, so it makes sense to take much of what one is told with a pinch of salt. But if one thinks in accord with the bias one will suppose that there are a few fools and liars out there, whose testimony is not to be trusted, and everyone else's reports are generally OK. One will be generally trusting unless there is some reason for suspicion. This obviously eases coordination generally: when one person shouts that the buffalo are stampeding the others don't apply complex statistics to discover the likelihood that he is mistaken or misleading. And it obviously lightens the cognitive load: instead of a subtle balancing of opposing factors one has a default assumption which one holds on to until it is definitively overturned.

Bootstrapping and truth

Human cognition has systematic tendencies to falsehood. There are perceptual illusions, distortions of memory, and the like. Current orthodoxy has it that these are incidental effects of procedures that are generally efficacious. Given our limited cognitive powers we embody short cuts and heuristics which generally work well, but sometimes give us false or even dangerous information. The false information usually has no practical consequences and the dangerous information (most false, some true) comes typically in situations which would be very rare under the conditions in which we evolved. That is the price we pay for the ability to think our way through most of the situations nature thinks us likely to meet.

Attribution biases are slightly different, if what I have been arguing is right. They are *beneficial* false beliefs. They support coordination between individuals and give stability to patterns of social behavior.

While the generally beneficial qualities seem clear, it would be wrong to associate the attribution bias too deeply with error. Many of the false beliefs that it leads to are essentially notes made in the course of inferences that lead eventually to true conclusions. In seeing why this is so we can see versions of several of the themes of this book. The central point here is the self-amplifying, self-fulfilling nature of the bias.

Take the case discussed above of people estimating the likelihood that a current partner will make the better choice in a coordination problem. Suppose the pattern of choices is near the threshold at which people would start expecting others to make the worse choice. Enough people have made the worse choice recently that *if* you are thinking in straight statistical terms you might expect that your current partner will choose worse, so you might choose worse too, defensively, and thus tip the statistics a little further to the worse. As I argued above, if people approach the problem through the bias then they will tend to judge the current partner on the basis of that partner's individual record and thus be more likely to make the better choice. But there is more to it than this. Your estimate of the partner's likely choice will depend also on what you expect the partner to expect that you yourself will choose. If the partner is a normal human being (rather than a statistician) she will also be subject to the bias. As a result if you have a good record she will be more inclined to expect the better choice of you, and thus to make it herself. And you know she thinks this of you; and that she thinks you think she thinks this. At each iteration of expectations the bias adds a bit of assurance. The result is that it is in fact very likely that both people will make the better choice, and that the pattern in which people generally choose the better will continue. The bias has in the end led to a true conclusion.

In some contexts the expectations will definitely be false. We will expect people who have deviated from the average to deviate more often in the future than they actually will. To that extent we often apply character terms linked to such deviating behavior when the behavior is not forthcoming. We suppose that when people have acted bravely or dishonestly they can be labeled as brave or dishonest and will live up to the labels by future brave or dishonest acts. In a recent article Gilbert Harman (1999, see also Sreenivasan 2002) has concluded from this that most concepts of character and virtue are myths: they are the result of false assumptions and we would be best off avoiding them. What is clearly true about Harman's suggestion is that on the way to largely true conclusions about the actions that will result in mutual benefit we pass through intermediate thoughts, lemmas, which if taken in isolation from these patterns of motivation are misleading. The misleading quality is not too worrying, since they are usually not isolated from their natural patterns of motivation. More worrying is the suggestion that we'd be better off without the character concepts. Perhaps we would be, if we could find some other way of articulating reasoning such as the expectation-amplifying process described above. Or, better, perhaps we would be better off thinking of many terms of virtue and character as being like 'fair' or 'biased' as applied to coins. They are descriptions of rarely-found ideal patterns which are most useful as part of sets of general hypotheses considered on the way to forming expectations about particular cases.

Exploration II

Interspection and expression

A very mild expressivism

The theme of this book is ways in which the sharing of human activity can (in part) generate concepts of mind. The grasp of mind that we get in this way goes beyond what is needed for predictions of behavior. In fact it goes beyond what is needed for explaining behavior. It plays a role in the genesis of our sense of what it is to be like another person. (As a result, it can support moral ideas that go well beyond a simple social contract in which people work out the way of getting the most of what they presently want, given their present bargaining position.) In this Exploration I give *part* of a picture of subjectivity: of the sense people have that there is something which is what it is like to be someone else. The picture is in a very general way expressivist, as suggested by Wittgenstein's remarks in the *Philosophical Investigations* comparing the verbal self-attribution of a state to its expression in behavior. Wittgenstein's remarks amount to some suggestive analogies between describing one's own pain and pained behavior. To parody: 'It hurts' = 'ouch' = writhing. (See Wittgenstein 1953 sections 285, 304, 580, and passages surrounding these. I found Chapters 1 and 3 of Mulhall 1990 gave a useful perspective on Wittgenstein on this topic.) My claim is that when we consider interactions that are based on what in Chapter 5 I called 'plural simulation', we can see how some self-attributions have some of the characteristics suggested by Wittgenstein's examples. And, more generally, we can see a general route by which skills of social interaction can generate beliefs about one's own and other people's mental lives.

I am not aiming to reproduce in simulationist terms Wittgenstein's position. In spite of some effort with the secondary literature I am still very uncertain what Wittgenstein is trying to persuade us of about the states that we express: can they coincide with causes of our actions? Or about the meaning of verbal attributions: does a third person attribution say the same as a first person one? And I have no idea, not even a rough one, of the range of states for which the claims are being made. By calling the position I will derive expressivist I mean to stress two features of it. First, that the ascription of a state of mind to oneself or to another is part of a social interaction, in which the ascription of states by the participants to themselves and to one another plays complementary roles. *Some* of these roles can be served both by expressive behavior and by self-ascriptions. Second, that the self-ascription of the state is not based on a direct introspective awareness of it, but on a pretty complex cognitive process which the person is rarely aware of and which often uses information about others in order to yield an ascription to oneself. I do *not* mean to imply any irrealism about states of mind, or even about individual subjectivity. As far as I am concerned there are definite facts about what it is like to be a particular person at a particular time. But how these acts are to be understood, and how we come to know about them, are complicated and subtle matters.

Interspection

Let us begin with examples.

Construction

A rugby team is facing defeat. They are six points behind to a superior team with only ten minutes to go. The other team has the ball, very near to our team's end. Emerging from the scrum someone fumbles it; one of our players grabs it; he is immediately tackled but manages to pass it to a team-mate. A desperate series of passes and some very complicated running results in a try. They convert and from then on just hold on defensively for the remaining minutes to win against all the odds.

Later, in the pub, the captain proposes a toast to whoever first grabbed the fumbled ball. But no one knows who it is. Several players remember handling the ball, and they all agree on the path the ball took from one end of the field to the other. And they all see the improvised strategy that emerged and developed. But no one is sure who, until a few crises before the try, carried out which bit of it. It was all happening too fast, and they were, paradoxically, all too alert. 'Well, here's to us' says the captain.

They know what has been done, but not who did it. They know moreover what *they* have been doing, jointly, and its quality of smooth cooperation. This is not at all unusual: *often it is easier to know what has been done than who has done it, and often it is easier to know the manner in which several people have acted than the characteristics of any single person's action.*

Quarrel

You go to see your friend Carmen. You have sometimes quarrelled in the past and sometimes been very close friends. As you enter her flat something feels wrong to you. After a few exchanges of trivial information you know something is very wrong. ‘Why are you so hostile?’ you ask. ‘I just came through the door and you started to get at me.’ ‘Me?’ she replies incredulously, ‘Why you just come in here and start interrogating me like I’ve done something wrong.’ Now the two of you really do have something to be hostile about: who is it that is being angry? So you begin a row. The row is about who is feeling the anger: the first anger, that the other is reacting to. And you wouldn’t be having this row if it were clear to you which one of you it was. What *is* clear to you is that there is anger in the air. There is anger between the two of you, and individually or together you can attribute it (or perhaps distribute it) between you. The conclusion to draw is evident in many other cases: *What one knows immediately is often the nature of the shared situation; only after this does one infer each individual person’s state or situation.*

Discussion

You and Socrates are discussing the possibility that there is intelligent life on a nearby star. ‘Sure’, says Socrates, ‘there has to be; the basic chemical requirements are quite simple, and given them it is just a matter of time before evolution takes over.’ You tease out how widespread the necessary conditions might be, and what is included under evolution, and then you have a doubt. ‘Perhaps we are the slowest rather than the fastest to evolve’, you say. ‘Perhaps elsewhere intelligent life has reached the point where it is intelligent enough to see its pointlessness, and just stopped.’ Socrates thinks that this view is absurd and disparages your thesis of inevitable voluntary extinction in the face of your attempts to make it plausible. By the end of the discussion each of you has a well-defined position, and a set of arguments for defending it against that of the other.

Two features of the discussion are worth underlining, as they apply much more generally.

Self-attribution arises out of the interaction

As you argue you find yourself defending beliefs you did not know you had. Some of them were formed during the argument, and some of them you had earlier, and you cannot easily know which are which. Your reasons for thinking you have these beliefs are of two main kinds. There is your explicit assertion of some of them, provoked by the discussion. And there is your assertion of other beliefs, which would make most sense if you had the beliefs in question. In either case the best explanation of your pattern of assertions is that you have these beliefs. Thus in the case of your own beliefs as in the case of the other person’s, your ascription depends on what happens between the two of you.

The ascriptions to self and other are inseparable

The assumption that you believe that p is part of the best explanation of why you were saying q when the other was indirectly defending r in the expectation that you were going to argue for p. But you never said p and the other never said r. Each is ascribed as part of a process of making sense of what both people say, simultaneously. Neither set of implicit inferences to the best explanation would work in the absence of the other one. You construct a coherent set of beliefs and desires, a persona, for yourself and for the other, and the two processes are thoroughly entwined.

Pain

Now for the primal Wittgensteinian example of the sufferer and the sympathizer. You go to visit a friend in hospital. The friend has just learned that her disease is fatal. She tells you, straight, with no embellishment or self-pity. *You* burst into tears and she gives you a hug. You both cry.

Who is suffering and who is showing sympathy? You may not know. What you know is that there is suffering and sympathy between you. All the conclusions above apply: the attributions to self and other are inseparable, your primary knowledge is of how it is with the pair of you, each is better placed to know some things about each of you than the other is.

Expression to simulation

The argument-by-storytelling in the previous section was meant to bring out a sense of everyday interspection, that is, the frequent primacy of us-knowledge over I-knowledge. We regularly know basic facts about what is going on between people (usually ourselves and others) more firmly than we know how it is caused or distributed between them. That is my target; I want to describe some ways in which interspection can come about. My strategy is implicit in the first of the examples, *construction*. In that example the team are doing a job together; the knowledge that comes easily to them is of the job and how it is being done, and from this they have to reconstruct an account of what each has done. Taking that as a clue, consider accounts of shared activity.

Begin with the 'plural simulation' of Chapter 5. The basic idea there was that there are advantages to a decision-making procedure which in its initial stages considers the coordinations and interactions of the options open to agents acting in concert or in opposition. I argued that some specific procedures of this kind are less expensive of cognitive resources than procedures which separate the decision problem of the agent in question from that of the other agents. (They can give better results, too.) These are simulationist procedures in the specific sense that a prediction of other agents' actions results from the agent's own decision-making procedure. Let us call a way of choosing actions an instance of *plural simulation* when it has the following features.

(a) The procedure provides material relevant to the agent's own decision making.
(b) The same material is relevant to a prediction or explanation of what another agent may do.
(c) It will benefit each agent to act as if assuming that the decisions of the other will be based on the output of the process.

(There is no claim that this is the whole story about any decision. The output of the process is required only to be part of the basis for decisions, predictions and explanations.)

I shall assume that any process that provides predictions and decisions in a way that is economical with cognitive resources and which satisfies features (a), (b), (c) will have two other features.

(i) The choice of an action by each agent will be a consequence of the fact that some outcome or set of outcomes has some desirable property. (This is suggested by features a and b: since the same material plays a crucial role in the determination of the actions of several interacting agents, it is natural to take it as defining a set of outcomes which could be produced by combinations of their actions. For example in one of the models described in Chapter 5 the property would be that the benefit to all agents was above some threshold.)
(ii) In determining what action to choose given the preferred set of outcomes agents must distribute roles or tasks between them. One outcome may consist of one person doing A and another B, while another may consist of the first doing B and the second A.

Now imagine several people performing some activity together. The activity may have some general purpose, which we can presume they know about, but as the action proceeds they will generate many sub-goals of which they have no conscious awareness. For there is nothing in the description of plural simulation that requires that people have access to its component processes. Suppose that we ask one of the people why she is performing some action that is part of the shared activity, or why some other person is performing some action. She won't be able to answer, not just by virtue of what she is doing. She will have to reconstruct the structure of the shared action from what she can experience. She may not find it easy to say exactly which component actions she has done and which have been done by others (as in the first example.) There are two things that she will find it relatively easy to describe. They are, first, the general tendency of the activity: whether for example it is proceeding well or lurching into disaster. (Think of music or tennis.) Second, the kind of effort that is being asked of her at the moment: whether for example she is having to perform many frantic actions or a few uninterrupted ones. (Think of trying to meet someone in a crowded room.) These two kinds of knowledge obviously reinforce one another. They are relatively easily obtained because they are things the agent must bear in mind to choose particular sub-actions at the present moment. (So she can look at the tendency of her voice or hands at the present moment, consider the kinds of information she is needing to find, and so on.)

So the person often has knowledge of these two things: on the one hand the direction of the project and the kind of effort being required at the moment, and on the other hand the very general shared purpose of the activity. Inferring from these, she can reconstruct more details of the action. She can say, first 'this is what we are doing', next 'this is the result it is tending to', next 'here is the distribution of roles between us', and last of all 'here are each of our detailed aims'. Each of these is an inference to the best explanation of the previous conclusions. Each is less certain and requires more thought.

That is my derivation of interspection from plural simulation. It makes a case that in situations like my four examples the relative priority of the people's knowledge of plural over singular facts can derive from the relative ease with which a process of plural simulation can be reconstructed by its participants. Consider again the first three examples. In 'construction' the team members find it easy to tell that their actions have the general characteristics of cooperation. It is as if each thought what needed to be done, moment by moment, thought who then would best do what, and then did their assigned list. The order of reconstruction begins with the acts, then infers the general character of the interaction, and only lastly and speculatively comes to the distribution of roles. In 'quarrel' the two people find that they are acting as if in a pattern of attacks and defences. It is as if acts were chosen because they were responses to attacks and preparations for future attacks. If the two people assume that their actions resulted from a plural simulation then they can easily invert the process to see that these actions exhibited the peculiar coordinations of a quarrel: refusals to make peace, wariness to traps, openness to opportunities to wound. They can thus conclude that they are quarrelling. But it will take longer and more speculative thought to show them who was attacking whom.

In 'discussion' the point was the interdependence and inseparability of the attributions each person made to themselves and to the other. This too makes sense in terms of plural simulation. Having determined the general characteristics of their interaction, in this case cooperative disagreement with a view to becoming clearer about a hard question, they can then explore the motives behind each act. But they do not follow automatically from the plural characteristics of the interaction. They have to be conjectured, and the conjectural attributions to each must be consistent with the attributions to the other.

Figure and ground

Now consider the last example ‘pain’. The same conclusions apply. But what is the cooperative or rivalrous activity? That will depend on the details. In the story as I told it the one person is paying a hospital visit to the other. So the project of keeping the one person’s morale intact is evident. But below that there is a deeper and more universal human activity which we might just label ‘grooming’: where the essence of the activity is one person’s paying attention to another, and whose effects are solidarity in future activities. This activity, in all its variations, is clearly a ubiquitous source of human fellow-feeling. So there is a challenge to relate it to plural simulation. Especially since it looks at first as if a non-plural simulation along the lines suggested by Gordon or Goldman (what in Chapter 5 I called modeling) more naturally applies. You put yourself in the other person’s situation and you find that you simulate unhappiness.

The example brings to the fore one very basic fact about sympathy, though, which does not fit well with singular simulation. When you appreciate someone’s anguish you feel sympathy. (Normally, that is. Sometimes you feel delight or satisfaction.) Almost never do you get the other person’s anguish directly, unframed by your attitude to it. This is just the expressivist claim again: attributions come in pairs, or multiples, of what one person feels and what another person in interaction with them feels. In fact, it suggests a claim about how one comes to appreciate the situation of another person. *Very often, the experience of knowing what state of mind another person is in is that of experiencing some complementary state and knowing that the other is in a state to which your state is complementary*.

(Compare the point made in the appendix to Chapter 1, that attributions of preferences to yourself and to another are often interdependent.)

This is a claim about the experience of knowing another’s state. It does not deny that very often you know, propositionally, simply as one fact among others, a fact about someone’s state of mind, without there being any characteristic experience or fellow-feeling. But when there is such an experience it typically consists of taking an experience of your own and then adding a label: the other is in a state which fits with this in the following respect. An analogy may help: your own experience and that of the other person are like figure and ground in one of those familiar gestalt drawings in which, for example, a pair of faces in black form the outline of a vase in white. The vase shape is the state of mind of the other, the framing faces are your own state. If you ask what the state of the other is the answer is that it is the shape that is framed by your own state, if you stop thinking of it as figure and start thinking of it as ground.

This is a rudimentary account of subjectivity. As a first stab we might say that – sometimes, for some experiences – to think of someone as having a subjective life is to take a part of your own experience and to see it as the figure to which their experience is ground. This is *not* to take your own experience as part of your subjectivity. One way to do that is to take your experience as complementary to another person's, that is, as complementary to the complement of an experience of yours. Being sorrowful is one thing; thinking of your sorrow as part of what it is like to be you is another. The latter thought is more sophisticated and potentially more perverse, akin to feeling sorry for yourself, since it is to experience, directly or in simulation, the experience which would complement someone else's sympathy for you. (Crudely: one way of taking your own sorrow from the inside, to have a subjective perspective on it, is to imagine bravely refusing the consolation that someone might offer you. The thing you simply have is the refusal; this complements and thus gives a subjectivity to the sympathy; and this in turn complements and makes subjective the sorrow.)

I shall leave this as merely suggestive, or perhaps as irritatingly rhetorical. I do not think there is a single right way of making the suggestion clear and unmetaphorical. (But see Chapters 4 and 5 of Humphrey 1984 and Chapter 13 of Dennett 1991, both of which I take to be on the same wavelength as what I am saying.) I do not think that there is a single concept of subjectivity that will survive clarification. I do suspect though that there is a variety of things that can fairly be called conceptions of subjectivity, which may fit in a variety of ways into patterns of psychological explanation.

The conclusion I want to draw is that the experience of how it is with another person, that one gets from simple communing with them, grooming, is *of the same kind* as those that arise from plural simulation. They consist in having an experience that gives you the other person's experience by the same kind of figure–ground reversal. Does this mean that it is actually the result of a plural simulation, even when there is no specific shared practical activity?

Here is a conjecture, that it is very natural to make given all this. Imagine creatures, proto-humans perhaps, whose social lives have two features. On the one hand they engage in complex shared practical activities, of a kind that require them to form conceptions of ends to be jointly achieved and then work backwards to divisions of tasks that best achieve them. And on the other hand they have practices of communing with one another, perhaps literally grooming or perhaps simply demonstrative visual attention, eye-play, or social vocalising, whose function is to create bonds between members of a group and to form special alliances that will be useful politically within a group (see Dunbar 1996.) The shared activities would lead them to have a conception of one another's experiences, in ways that I have been trying to describe. The grooming would not; it would result just in bonding. But if grooming could come to result in knowledge of the states of the groomers, there would be clear advantages. Individuals would come to know things about the general dispositions of others: their likelihood to cooperate or be obstructive in various ways, their mood of the moment which – as one might learn – are related to their dispositions to make various energetic efforts on one's behalf, and so on. In general, individuals would have to acquire a capacity for transforming the kind of experience they have while grooming into useful facts about the dispositions of others.

In order for such a capacity to get going we would have first to have suitable experiences while grooming, indicating the state of the other. And on the account I am developing that requires shared activity. In fact, it suggests a special kind of shared activity whose purpose is to find out how it is with the other. But we do have such a special shared activity. It is called conversation. By talking in a general aimless 'Hi, how you doing?' way we aim to acquire just the sense in question, of how it is with the people we are talking to. And, significantly, we also aim at establishing ourselves as potential allies, or at the least as non-enemies. In a way it is misleading to describe this kind of talk as aimless, for it has a very specific purpose: to allow people to make conversational moves that to an empathetic listener reveal their states. And for this it is often much better to talk about the weather or computers than about one another's soul. So there *is* after all a shared practical activity. It has a high level practicality but it still has a definite shared aim, roles in achieving which have to be extrapolated backwards: that of finding out about one another.

Beyond expression

The processes I have been describing are just some among the many ways in which we attribute states to one another. They can be seen as occupying a middle ground. On one side of them there are fully conceptualized attributions that are not intrinsic to any social process. Typically these are made by one person on the basis of another person's behavior as guided by beliefs about states of mind. And on the other side of them there are the ascriptions of simple qualitative feelings. You have an itch, and it occurs to you that someone else might have an itch that feels like *that*. (This is the home of the classic argument from analogy. A very small part of our concept of mind, but a real one.) In real life attributions from across this spectrum, and others, coexist and mingle. This makes it very hard to be sure that any particular case fits any specific pattern. There are many cases that seem intuitively to be mixtures. Consider for example someone who is in extreme pain, and another person who is struggling to understand what she is going through. He listens and watches and experiences a very specific sympathy, which constitutes part of his sense of her situation. But he also gathers from what she says comparisons with the time he had an inflamed gall bladder, and he thinks 'the pain must be a little like what I experienced then'. And he also remembers reading in a medical book that pain from her condition is typically of a deep pervasive kind, which the sufferer cannot locate at any specific part of her body. So he assumes that this too is true of her, though he cannot easily imagine what intense pain with that quality would be like.

How are these very different elements combined? In many ways, no doubt. One way is to exploit the link between social emotions and moral judgments. Hostility is linked to blame, sympathy to the belief that what is happening to the person is wrong, gratitude to a sense of obligation. The links are rough ones; they support very few perfect generalizations. But there are reliable imprecise facts about humans here. Someone who thinks that what he has done is very wrong will usually feel shame. Someone who thinks that what someone else has done is very wrong will usually find it harder to feel affection for them. Someone who feels sympathetic to someone's plight will often think that they ought to be helped (see Greenspan 1988.)

These links generate a bridge between expression and cognition. The connections go both ways. From cognitive to affective, the connection can go from a belief that someone has been wronged to sympathy for them. From affective to cognitive, the connection can go from a sense that there is an unevenness in the distribution of some burden or benefit between oneself and another to feeling that the other has been treated unfairly, to a belief that one should compensate them. Very often the link goes both ways at once. And very often it is mediated by social roles. For example, suppose that you are parenting a child. That goes with a set of beliefs about your duties to the child and a set of beliefs about how the child should behave to you. It also goes with a set of standard ways of sharing tasks between you and the child, in which, typically, you either take the lead or stand ready to catch the consequences of a mistake. So in any shared activity you will be primed to understand yourself and the child in terms of paired states of mind – reproving and rebellious, advising and learning, example-giving and observing. Each of these gives you a pair of labels for your own experience and that of the child. But it also engages with a set of normative and explanatory beliefs, which give you an alternative handle on the interaction. So you can as easily take the child to be, for example, in a rebellious mood because its action is one that contradicts a norm, as because the pattern of activity between you is taking the overall form of control-and-resistance (see the paradigm scenarios of de Sousa 1987).

There are many patterns of attribution here, but one general pattern runs through them. Very often when a person knows what state another person is in their knowledge consists of a mixture of an experience whose complement with respect to some particular interaction is attributed to the other plus a belief that the other is in some particular state. These can combine easily into a single piece of information because the belief is a normative one, it concerns some normative attribute which is closely associated with a characteristic emotion. When this is so the most accurate way for the person to describe the state the other is in is often in terms of moral attributes. One person says that what another feels is remorse, or the outrage of injustice, or a sense of the claim made by some person's need. Since our moral beliefs tie our instinctive cooperative and defensive impulses to our complex worked out social conventions they are well suited to mediate a two-way traffic between interspection and belief-based attribution. There really isn't any substitute for them in this respect.

When interspection is linked to both ascription of qualia and explanatory beliefs, questions of accuracy arise. Interspection has to be true to both its neighbors. It has to be the case that when one person takes the situation of another to be a certain way, by the process I have been describing, that 'way' is conceivable as a 'like this' pointing to that person's own experience. And it has to be the case that when the process leads to an ascription of a specific state and the person believes that people in that state act in a given way, then, often enough, this expectation is satisfied. Neither constraint is automatically satisfied, but each can be.

Consistency with the ascription of qualia is possible as long as when one person is in state A and knows of another person that they are in state B, characterizing it as 'the state in reaction to which it is appropriate to be in state A', that first person can think 'that state' of it, and also think 'that state' of it when they are themselves in B. State A serves as a kind of pointing gesture towards state B, which must later be also pointable to with a different gesture from a different direction. (Indeed, as I argued in the previous section, the second of these demonstrations is actually sometimes harder than the first, since it can require pointing to a pointing.) You feel sympathy for someone while taking them to be in distress, and later when you are in distress you identify what you feel with the state that evoked your sympathy, to some extent thinking of yourself as a potential object of sympathy. There need be no obstacle to this identification, unless some other feature of a felt experience stands out as inconsistent with it. Indeed, the identification gives a point to saying 'that experience'. It links it via the interspection to explanatory and predictive beliefs.

Consistency with explanatory beliefs is harder to achieve. The interspective process results in a pair of attributions: one person is in state A and the other is in B to which the appropriate reaction is A. If these states have conceptual labels then there are beliefs linking them to actions. So the attributions may be falsified if too often the actions are not forthcoming. A person feels abashed in the presence of another and so thinks of them as angry, and also thinks of anger as a state that tends to produce aggressive behavior. But the other acts in a friendly unaggressive way, as do others to whom that person's abashedness leads them to attribute anger. So something has gone wrong. That is a welcome result, interspection does not sit in a closed system of reactions and reactions to reactions: it plays a role in the larger network of explanations, predictions and causes. Since it can be wrong, there is some point to taking it, sometimes, to be right.

Empathy and evil

Most of the discussion has been of attributions arising out of cooperative situations. I have argued that a natural way of thinking through such situations can give rise to an experience of what it is like to be the other person. But not all interactions are cooperative. Some are competitive or hostile. Can these be sources of insight into others, and oneself?

A denial that they can be is found in an article by John Deigh.

> the pleasure a sadist gets from, say, assaulting someone is typically increased by his imagining his victim's pain. The sadist thus exhibits empathy inasmuch as he shows that he is taking in his victim's suffering, imagining, say, its course and intensity. Yet the sadist's empathy does not count against his having an egocentric view. The reason is that his empathy . . . does not imply a recognition of his victim as an autonomous agent. The sadist, one might say, in getting pleasure from another's pain, fails to take in the whole person. He revels in the pain and suffering he has produced in that person but does not see beyond these particular feelings and emotions to the life his victim is living or the purposes that give it extension and structure. He does not see those purposes as worthwhile, as purposes that matter.
>
> Deigh 1998, p. 201

Deigh is arguing that although something very close to what I have called interspection can occur between a sadist and his victim, the sense of the other that results is in specific ways limited and inferior. The morally worse attitude gives the psychologically shallower attributions. Deigh's reasons for believing this are interesting. The sadistic attitude is not consistent with grasping the whole person: one cannot link it to beliefs about the person's wider aims or the structure of their life. In the terms of the previous section we might say: the sadistic attitude is confined to a basic interspection, and cannot be linked to general beliefs about the person.

There is something persuasive about this argument. Much of the vocabulary we have for describing our states of mind is meant to help cooperative people avoid conflict or work out the smoothest routes to mutual benefit. When things get nasty we fall back on a smaller and cruder sub-vocabulary. There is a similar contrast in the explanatory principles available to us: a rich and subtle body for dealing with friends and a simpler more guarded body for dealing with enemies. And, plausibly, our acquired and innate routines of simulation, attribution and expectation are more geared to working out what is to people's advantage than what will hurt them.

But there is a lot this angle leaves out. There are specific kinds of psychological acuteness that abusive, manipulative, or deceptive people have. And there are attitudes deeply embedded in the vocabulary and explanatory principles of many, probably all, cultures, whose function is in various ways less than admirable. For their function is to preserve the local cooperative equilibrium, even if it is built on arbitrary distribution of power and an irrational assignment of rights. (Sometimes it is in most people's interest if some people's feelings are not taken seriously.)

The first point – the acuteness of bad people – is my concern now. Could there be a way of understanding others that served purposes that were evil, and which also gave real and deep insight? Specifically, could there be a way of framing the states of others with one's own cruelty or exploitativeness, which generated an accurate sense of what it is to be those other persons?

I see no reason apriori why there could not be such an evil interspection. However this acceptance must be qualified in two ways.

First, there is the effect of the bias in human culture towards action in the common interest, with antagonistic interactions largely restricted to outsiders or the powerless. A culture devoted to mutual harm would not evolve or survive. As a result, the deeply and universally evil person is largely on their own in terms of concepts and vocabulary. They will have to make up their own words and attitudes to describe their states, those of their victims, and the interactions between them. This would be very demanding. In real life, evil attitudes exist at the margin, exploiting opportunities created by a general folk psychology and way of life.

Second, and more profoundly, there is a question of moral-psychological fact: are the motivational states associated with the experiences framed by such attitudes as fundamental or as explanatorily relevant as those framed by more benevolent ones? Can a sadist, for example, have a good understanding of why his victim acts as she does, by using his particular point of view? Consider for example the primitive evil described by Colin McGinn:

> It can be a primitive fact about someone that their own pleasure . . . is reinforced by the pain of another . . . This is, as it were, simply how they *are*. Of course, a person with the evil disposition might well contrive some sort of rationalization of his psychology . . . But this is really *ex post facto*; the evil disposition comes first
>
> McGinn 1998, p. 82

Suppose, to give the idea some scope, that there were many such persons, and that this fundamental desire to do harm played a wide role in their motivation. Many of their actions are directed at the harm of others for its own sake. If we try to imagine how people like this would imagine what it is like to be one another there seem to be two clear possibilities.

The first possibility is that the desire to harm is generally known and has a simple rationalizing relation to their actions. Then the psychology is basically Humean: this is what people want, and so they will do what they need to get it. (I suspect this is the case McGinn is imagining.) Then the interspection of harm-doer and victim does not lead to any greater understanding of what each is likely to do. It just adds a qualitative frisson: the harmer can experience the squirm or terror of the victim framed by his complementary anticipation or glee. (Indeed, can experience the victim's experience of that anticipation or glee.) Neither harmer nor victim gains any extra insight into the other's motives.

The alternative possibility is that the desire to harm is not apparent on the surface but is a hidden factor in the people's psychology, emerging in the effect it has on the long term evolution of more graspable desires. It is an attractor, in the terms of Chapter 4. Then there is some scope for the characteristic dialectic of interspection. One or both people in some social situations can realize first that there is aggression and fear in the atmosphere; someone is potentially harmer and someone is potentially victim. The next step is to work out who is which. (This is particularly hard in sexual contexts, intuitively.) And the conclusions of this next step may be surprising. So the attribution of individual motives based on dissecting the feel of social interactions, could be a valuable source of information to people about their own and other people's hidden motives.

In fact we do learn about possible aggression and victimhood in this way. And it is very likely that frustrating or dominating others does play some role among the qualities towards which our desires are attracted. We're no angels. We then approach something more like the actual human condition, in which a mixture of people with varying degrees of benevolence and malice to one another try to discern good opportunities from traps. The interesting possibility then is that agents might make use of interspective capacities to intuit possibilities for exploiting others. And the hard question that goes with this possibility is whether these could be the same capacities that are used when the frame in which subjectivity appears is cooperative. Could this be just as accurate a window through which to see the same qualities of our experience? If the analysis of this Exploration is along the right lines, then such exploitative capacities have to build on skills whose primary function is to mediate shared projects. To anticipate the vocabulary of the last section of this book, the exploiter understands nothing that the cooperator cannot also understand. The cognitive components of interspection-based understanding are all potentially usable in interactions which aim at finding and achieving the good of the people involved.

Exploration III

Ethos

No one can rule unless he can also submit to rule. Consequently empires have belonged to peoples who enjoy milder climates. Peoples inclining towards the frozen north have savage characters.

Seneca, *On Anger*, 2.15.4–5

'And what sort of a young man is he?'

'As good a kind of fellow as ever lived, I assure you. A very decent shot, and there is not a bolder rider in England.'

'And is that all you can say for him?' cried Marianne, indignantly. 'But what are his manners on more intimate acquaintance? What his pursuits, his talents and genius?'

Sir John was rather puzzled.

'Upon my soul,' said he, 'I do not know much about him as to all that. But he is a pleasant, good humoured fellow, and has got the nicest little black bitch of a pointer I ever saw. Was she out with him to-day?'

Jane Austen, *Sense and Sensibility*

Ethos

Crucial aspects of our understanding of one another work because they are shared. Some of these aspects and some of the ways in which they are shared were explored in the first five chapters of this book. Because our grasp of one another's motives, character and actions is essential to our capacity for co-operative action, a shared conception of how to understand, and how to act so as to be understood, is closely linked to a shared conception of cooperation: what its aims are and how it is achieved. And such shared values will reinforce the shared psychology. So we should expect there to be stable combinations of explanatory devices tuned in specific ways, values both acknowledged and implicit, and norms of action. And we should expect these combinations to vary from one time, place and social group to another.

The aim of this chapter is to grasp, as well as I can, the stability and variability of these combinations of explanation and valuation. I shall refer to a local style of explanation and evaluation, which has a certain degree of permanence and is shared by a number of people interpreting one another's actions and cooperating to common ends, as an *ethos*. The term is loaded with assumptions, as I use it. The main assumption is that there is something to talk about, that the component bits of apparatus for explaining action, deciding what one should do, and forming conclusions about what others have done, work together. Different elements of the whole support and influence other elements. And in particular, there is support and influence between explanatory and normative elements. There are stable ways of thinking about people.

A warning, though. The obvious way to investigate ethos would be to apply well-confirmed social and developmental psychology to a compendious account of the values and forms of explanation found in the various forms of the human condition. In the end, that is what must settle the issues. Philosophy often provides a way of thinking carefully about a topic in advance of the kinds of evidence or theoretical techniques that would give a scientific understanding. But on this topic I have got as far as I can with the techniques of my subject. And nothing will be gained by armchair social anthropology. So my technique will be more imaginative than argumentative. The center of gravity of the chapter is 'Toy ethos', in which four toy ethos are described. Each is a schematic description of an extreme possibility, presented with an appeal to our capacity to imagine what it would be like to live in a different complex of ideas. The contrasts between the four are more important than the details of any one: if they do indeed represent anything like four extreme possibilities then the space for variation between them should describe possible variability of actual realizeable ethos. It is for this reason that I have named them North, South, East and West, rather than by allusion to the locations of any real cultures. But, to repeat, the whole exercise is not intellectually responsible unless it is taken as one of imagination rather than persuasion.

Four crucial factors

We are looking for factors shaping the ways people put together the resources they have for thinking about one another. These resources are very varied, as varied as the ways that people can make decisions, coordinate their activities, imagine one another's state of mind, and conform to general norms. I see no reason to believe that there are not many ways to put the pieces together, though most of the pieces are parts of the common inheritance of humankind. Are there general types of pattern that result when the pieces are assembled? To begin, here are four kinds of pieces.

Basic explanatory patterns

When saying ‘P did A because of . . .’ what fills the ‘. . .’? A ‘Humean’ folk psychology would supply only beliefs and desires here, or will insist that any other motivational factor must be shorthand for something that includes beliefs or desires. Such a folk psychology would also understand the relation between beliefs and desires and the actions they explain as one of means–end rationality: the action represents the best way to satisfy the desires given the truth of the beliefs. This itself is not a univocal thing: there are many plausible ways of understanding what the best way to satisfy a set of desires is, and there is no reason to believe that we always focus on the same one. (Think for example of the different possible attitudes to the management of risk.) But there are also non-Humean ways of understanding what the basic explanatory factors are. We can appeal to the attractive powers of individually and socially desirable outcomes, as described in Chapter 4. We can emphasize the virtues and vices that determine which beliefs and desires are acted on and which means are chosen, as described in Chapter 2. There are also more extreme possibilities. We can explain by reference to states of character and social position, combined with knowledge, rather than states of desire, combined with belief. We can explain directly by norms of right action, perhaps augmented with conceptions of rights and duties. (We doubtless do this as a time-saving heuristic even if our official psychology is Humean. Our default expectations are that people will act roughly as they should.) And, as argued in Chapter 3, the concepts of belief and desire are not fixed quantities. We fine-tune them to bring our concepts of belief and desire into line with those of others.

Simulation procedures

Often instead of representing someone else’s thinking we think, ourselves, in some parallel way whose outcome is informative about the other person. This is simulation. It is unlikely that any explanation by motive can work except by appealing to background conditions provided by some sort of simulation. There are many different ways of simulating. In Chapter 5 I described three contrasts between kinds of simulation: cocognition versus modeling, singular versus plural, centered versus non-centered. These may not encompass all the simulation procedures we use. The important point for this chapter is just that there are many possible ways of using your head as an analog tool for dealing with what could be going through another’s, so that different such ways may combine differently with other elements of a strategy for thinking about others. I argued in Chapter 5 that simulation procedures present a coordination problem: it is generally best to adopt the procedure that others around one are also using. So though we cannot expect that there is only one kind of simulation that people use to anticipate one another’s actions, we can expect that there is a rough congruence between the kinds that people sharing a common social life employ.

Kinds of expected interaction

Some people may live under conditions that mean that most of their interactions are coordination problems. Others may find, or expect, many prisoner's dilemmas. And so on. It may be part of a society's beliefs about social life that the general character of social life is cooperative, or exploitative, or grounded on deception, and this belief may be as influential as the actual facts may be. Or it may be self-fulfilling: the belief that all interaction is zero-sum competition has a notorious capacity to create its own truth.

Conception of subjectivity

This is the hardest to describe. In every person's development the innate capacity to react to others as persons and to imagine what it is like to be those persons develops into an individual take on the human world. The end result is not fixed: there are many ways of conceiving of The Other, and no doubt every one of us is to some extent deficient in her intellectual and intuitive grasp of what it is to be another person. (If Exploration II is right, this means that every one of us is deficient in her grasp of what it is to be herself.) The values and explanatory practices of people around us must have a large influence. So, as a result, in part of the culture around us, we can think of people as: centers of sensory experience, bearers of propositional attitudes towards the common environment, seekers after the good, partial participants in the divine, semi-blind stumblers towards satisfactory accommodation with others. I would not want to reject any of these as possibly sensible ways to think of other people. But I am sure that these labels do not go very deeply into the fundamental differences and dimensions here. In the next exploration I discuss a view of Thomas Nagel's, according to which at the heart of the deontological conviction, that one cannot balance possible evil done to one person against possible benefits to others, there is a recognition of the reality of the other person's point of view. I am sure that this is right in the following sense: there are very subtle, but deep and pervasive, ways of grasping what it is to be another person, which are reflected in very hard-to-systematize patterns of explanation and of moral judgment.

Toy ethos

Here are four toy examples, vastly simplified models of stable combinations of the factors I have been discussing. If the four have each a plausible coherence then we have an indirect and artificial grasp of something not completely different from the real things. I call them *South*, *North*, *East* and *West*. The names are just labels.

South: honor

We can start with any element of an ethos, and follow the threads to the others. With South I shall begin with concepts and follow their connections to strategy. (With North I shall do the opposite.) In a South ethos basic normative and explanatory concepts describe people's relations to other people: social status, admiration, contempt, honor, shame. People are described in terms that fix them in relation to others, generating both expectations about their patterns of action and also obligations centering on their treatment of those others. Alliances are important: besides relatively fixed long term social status an individual is understood in terms of the bonds he has formed with others, sometimes temporary and for special purposes. Almost any interactions between two people can generate an understanding that there is a bond between them which creates obligations on each.

Patterns of explanation concentrate on the achievement of the social ends to which the individual is committed. There is room in explanation for individual calculation, but this calculation takes for granted ends that are defined in terms of the relations between people. The modes of simulation which back up these explanations tend to be plural, working back from standard paradigmatic relations between people occupying the roles of those concerned to a distribution of actions to them. There is a close dovetailing between the roles that feature in these end-states for simulation and the roles that determine whether a plural simulation is applied. Strangers do not fit in easily.

These ways of predicting and deciding are well suited for situations in which the main aim is coordination. That is, we assume that in most situations there is an assignment of acts to agents such that each person will be best off if everyone performs the act assigned to them. Then we will find that plural simulation will reliably assign the right acts to the right agents. (For example the routine RC of Chapter 4 works well for coordination problems, less well for more devious situations.) The resulting actions will be explainable in terms of desires for the resulting equilibria. And in general belief/desire reasoning is appealed to either to confirm that someone is related as expected to publicly recognized objects and objectives or to explain errors in terms of factual mistakes and whims, still with respect to what is public. In effect, belief is *de re* of a world constituted by shared acceptance.

There is a self-fulfilling quality to the belief that social life is a series of coordination problems. If, assuming it, people think of their choices in terms of finding the act that conforms to the pattern that others are conforming to, then they will find that they will simply overlook options which would create non-coordination problems. Suppose for example that two people are considering a deal where one will send an item to the other and the other will send money to the first in payment. Then thinking coordinatively the first will see their options as acting as 'seller' or 'refuser' and the second will see their options as 'payer' or 'refuser'. There are two coordinative equilibria here, where one sells and the other pays or they both refuse the deal. But this way of thinking makes invisible the option of setting up the deal and then not sending the item or the money, which is in fact the dominant option for either if all options are considered and there are no repercussions in terms of reputation or revenge.

It is important to make a good impression. Each person will of course try to be seen as trustworthy and cooperative, but in a South ethos display will have a particular role. In coordination problems the pay-off is not the same for all coordinators: some assignments of acts to agents are more in the interest of a given agent than others. Thus it is good to make one's choice of role very visible to others. If you can do this saliently enough then it will be in their interest to choose complementary roles, for they will assume that others will also have noticed your choice of role. And then all will know that conformity to any assignment that is chosen by others is better than falling through the ice off the safe diagonal. South social display thus says 'Trust me: act in the way my act suggests and we'll all be alright.' Trust, pride and self-respect have a role in any human ethos, but the particular forms they take will vary in characteristic ways. In a South ethos to trust someone is to follow their coordinative lead, pride is based on the kind of display that establishes such a lead, and self respect is based on successfully leading or following in coordination with others.

North: sincerity

In a North ethos the strategic focus is on situations in which possible benefits and losses depend very delicately on the exact choices that individuals make. Typical are prisoner's dilemmas and free riding situations, and more complex situations such as the loonie game of Chapter 5, where there are potential benefits to all parties that can only be gained if they are moved by one kind of reasoning rather than another. The general presumption is that agents are in complex situations to which they can respond with a variety of lines of reasoning, the differences between which are of great importance for those interacting with them. As a result basic normative and explanatory concepts describe the ways in which individuals think out ways of satisfying their desires, and ways in which their desires and the acts that satisfy them relate to the interests of others. People are described in terms of: impulsiveness, care, intelligence, consideration, selfishness. Patterns of explanation state explicitly the individual's ends and their relation to their beliefs. There is thus frequent reliance on complex forms of individual simulation in which the point of view of the other must be taken into account. Agents' actions are explained in terms of the interaction of beliefs and desires unique to those agents. In effect, belief is a relation to a belief-world that may be unique to that individual.

Given the range of possible desires and the infinity of possible patterns of reason, there is an enormous range of types of situation for which agents' behavior is explicable. Since the range includes situations in which the combination of individually rational actions leads to an undesirable outcome, there is a particular set of character concepts which describe aspects of thinking which relate to an agent's behavior when there are conflicts between their interest and that of others: (ir)responsibility, (dis)honesty, (un)trustworthiness. Sincerity is an absolutely central attribute.

The good impression that a person strives to make consists in producing signs of a deeper earnestness. People will generally avoid too-easy display, since that is what is to be expected from someone either trying to lure others into a deception or trying to set up a coordination on an assignment of roles from which they would benefit. So the signs generally consist in some fairly delicate form of expressive behavior that will lead another person raised in the culture to ascribe sincerity. Sometimes a suitably nuanced renunciation of obvious display will constitute the necessary non-ostentatious display.

North social display thus says 'Trust my sincerity: my unique beliefs and commitments are such as to make it safe to deal with me.' In a North ethos to trust someone is to be willing to take a risk that only that person's invisible attributes can allay. Pride is based on one's ability to estimate these attributes of others in a way that results in successful interactions, in particular in well-judged risks. Self-respect is based on the conviction that one has the attributes upon which trust depends. Lack of these attributes can remain undetected, and it can result in activities which, while successful, work to the detriment of others. In either case the realization leads to self-respect-eroding guilt.

North beside South

Social life is very different in South and North. Pride, status and fear of humiliation; or self-respect, invisible attributes and the possibility of guilt. To be a member of a North culture is more demanding cognitively, since one has to consider complex chains of reasoning to understand the actions of others. Even simple cooperative behavior can be correspondingly more delicate, and can break down because people have more access to patterns of thought whose presence corrodes cooperation. On the other hand North does not suppose that the social relations between people are fixed.

Different philosophers' descriptions of moral and psychological life seem to describe South- or North-ish possibilities. Aristotle is clearly South-ish; Hume is clearly North-ish. Plato writes like a member of a South-ish culture trying to nudge it in a more Northerly direction. Wittgenstein writes like a member of a North-ish culture trying to nudge it in a more Southerly direction. South-ish cultures will be friendly to religions which emphasize observance and conformity – I think of traditional Catholicism or Mahayana Buddhism – while North-ish ones will be friendly to religions which emphasize invisible individual attributes – Protestantism or Hinayana. All of these characterizations are crude, impressionistic and in detail misleading. But they do give the imagination a handle that it can use to see the different facets of each ethos as complementing one another.

When North meets South North is puzzled that South can trust such obvious displays of insincerity. And South is puzzled that North can coordinate around such unimpressive leaders. They do not realize they mean different things by trust, as they do by belief and self-respect. South thinks that North is wrapped in needless worry which stifles fellow-feeling. North thinks that South is doomed to superficiality in both social arrangements and moral understanding. (While the North/South contrast is shaped by my specific purposes, I have been influenced by many writers who have tried to articulate their sense of how it might be different to belong to different cultures. In the case of Dodds 1951, Adkins 1970 and Williams 1993 the impetus is contrasts between different classical and modern cultures.)

East: suspicion

East is paranoid territory. There can be a paranoid ethos on a variety of scales: family, tribe, nation. Such an ethos represents a cognitively simple, self-fulfilling, evidentially closed complex of values and explanations to which almost any ethos can collapse or retreat. The ethos is easily presented in terms of its basic strategic image: all interactions are zero-sum; whenever one person gains another loses. It is a fact about zero-sum games that equilibrium solutions to them can be found by the minimax method: choose that option which exposes you to the least risk of a bad outcome, given all the choices the other agent may make (see Luce and Raiffa 1957, Chapter 4). This method is non-strategic: it does not require agents to consider one another's reasoning, since all they need do is to choose the option that has the least bad possible consequences. It does not require even that the agents know one another's preferences. But the zero-sum assumption that justifies it does allow one to infer another person's preferences in a two-person situation from one's own. It is thus a natural method to use when you think the person you are dealing with is irrational, or when you have no idea of their preferences. It is minimally psychological.

The zero-sum assumption makes most cooperation for mutual gain impossible. It denies that there are such situations. On the other hand it creates a strong pressure to some kinds of collective thinking. That is, thinking in terms of coalitions seems inevitable if minimax heuristics are applied to situations of more than two people. For suppose that we have a situation with three people, A, B, and C. See it from A's point of view. Since the situation is taken to be zero-sum each outcome can be labeled with the benefit to A and the benefit to B and C together. Then thinking in minimax terms A will choose an action which minimizes his possible losses, expecting pessimistically that B and C together will choose the complementary joint action which will minimize their losses, and against which A's choice is a best reply. But this is not in general going to be the case: the combination of B and C's actions which minimizes the sum of their losses need not minimize the losses of each one of them. (It may not distribute these losses evenly between B and C.) So in order for the minimax strategy to make sense A will have to think of B and C as in league, as having formed a coalition to achieve their best combined result.

East thus thinks in terms of Us and Them, with frequently shifting boundaries between the two. The acts of others are understood by grouping them into coalitions. And the threat of coalitions by Them enforces a degree of cooperation between Us. The concept of a coalition here is different from that in a South ethos. It centers neither on the trade between present and future favors nor on reliance on those who rely on you, but on the absolute commonality of interests. It is thus unstable: apparent allies can show themselves to be potential antagonists by wanting something different, even when so wanting presents no immediate problems of coordination or cooperation.

Only a very rudimentary range of psychological and moral concepts can take root in this ground. The nearest we get to belief is the characterization of agents as reliably or unreliably (maliciously) passing on knowledge. The nearest we get to moral worth are the concepts of loyalty and treachery. There is also a tendency to a rough-hewn vocabulary of distributive justice: *fair* and *deserves*. For the zero-sum presumption entails that human actions do not change the size of the cake. So when a non-ally is distributing you want to urge fairness, to get an at least equal share; and when you are distributing you want to urge desert, to minimize the amount that is wasted on the bad. (There is a standing danger in the concept of justice, that it collapses to that of distribution of a fixed resource, which encourages the zero-sum frame, which generates the paranoid ethos. Among the best-intentioned people, as Wolff 1998 points out.) The nearest we get to simulation is a kind of defensive plural conditional thinking, in which the actions of others are calculated in terms of their best responses to one's own minimax choice. In effect one simulates only to predict devious ploys and to catch others in traps.

It is far from obvious that a thoroughly applied East ethos will support the kinds of interaction needed to acquire many concepts needed in other domains. It is not clear for example that a child could learn how conversation proceeds, in an environment in which someone's saying something does not set up a presumption that they believe it. And it is hard to see how someone could have a conception of the desires of others without a grasp of what was good for them: what the attractors of their longer term projects might be. East seems much more readily to support the special cases of malice (wanting a result because another is avoiding it) or encouragement (wanting a result because an ally is striving towards it). There is a serious question whether all the interactions of any sustainable culture could be governed by the East ethos. Perhaps it has to be a passing phase or a dead end. (A dead end that is easier to turn into than get out of. Since given a fuller characterization of human motives minimax choice is often irrational, and since minimax choice is a good way of dealing with an irrational opponent, minimax breeds minimax. A trap.)

West: layers

West is Manichean. It bifurcates social interactions: either they are coordination problems or prisoner's dilemma-like situations in which the other is inclined to cooperate, or they are zero-sum or prisoner's dilemma-like with an uncooperative other. In either case it is easy to decide how to act and think, but it is not clear how to classify the cases. At the most basic level West says: most interactions are unproblematic and mutually beneficial, but some are subtle and treacherous, and you can be wrong about which is which. Moreover, since evidence that not all interactions are of the easy kind is inevitable, it becomes indubitable that there are subtle and treacherous interactions to be detected. As a result West will develop either or both of two ways of dealing with the uncertainty. It can divide the world into good guys and bad guys, either by some visible feature or by some subtle characteristic that is only revealed in interaction. Or it can classify people and interactions in terms of very subtle attributions of attitudes and reasoning.

The result is a surface of straightforward helpfulness above a hidden depth of uncertainty and complication. The tension between these can lead to xenophobia, racism, or ideological paranoia (things are not simple because of foreigners, heretics, communists). Or to psychobabble (things are not simple because people have unconscious desires, bad karma, messy childhoods.) Whatever the resolution, it becomes important to demonstrate that one is not one of these complication-making people, to display oneself as straightforward. So there are carefully planned displays of spontaneous helpfulness. (And a kind of prisoner's dilemma arises at a higher level: if all your neighbors are showing the flag, then failure to do so makes you an Other. All would prefer that none did, but if even some do, all must.)

The attribution of propositional states in West maintains the appearance of straightforwardness while allowing for devious reasoning. Thus beliefs and desires are directed at public objects – myths must be shared to be objects of thought – but the syntactical patterns of these propositions are allowed to become very complex, in order to support the tangled web needed to explain the frequent failures of the project of a simple cooperative life.

There are three fundamental contrasts between the four ethos. The first is the understanding of propositional thought. The import of simple attributions of the form 'she believes/wants that p' can vary from *de re* attitudes to a shared environment to *de dicto* attitudes to private belief worlds. From 'she believes *of* a and b that they stand in relation R', where R is some commonly understood relation between things, to at the other extreme 'she believes that there are objects with properties A and B that stand in relation R' (or even 'she believes that there are properties A and B such that A has characteristic Φ and B has characteristic Ψ and she believes that some objects satisfying A and B stand in some relation with characteristic Σ'!) South and (less simply) West are more comfortable with the first understanding, North with the second.

The second contrast is between strategic paradigms. South takes it that most manageable situations are coordination problems. North takes it that they are cooperation problems. East takes it that they are zero-sum interactions. West takes it that they all call for either cooperation or wariness (choice of a Pareto-optimum or of a minimax strategy.) The strategic paradigm influences fundamental aspects of an ethos. It describes the likelihood of mutual benefit. And it determines the likely complications of strategic thinking, the degree to which the motives of others will have to be followed out in order for an interaction to have a satisfactory outcome. The complication of strategic thinking is also affected by the degree of strategicality of the paradigm. East's zero-sum assumption keeps psychological thinking at a minimum, while West's superpositions of incompatible paradigms makes room for extremes of attribution.

The third fundamental contrast is in the simulation routines employed. In Chapter 5 I described a progression from simple plural cocognition towards singular centered modeling. Different stages along this progression provide suitable means for managing different strategic situations, as I argued in that chapter. These will correlate with the other two factors. *De re* propositional attitudes are correlated with a paradigm of coordination and with a reliance on cocognitive simulation. *De dicto* attitudes are correlated with a paradigm of cooperation and with a reliance on centered modeling. Moreover, a low estimate of the possibilities of mutually beneficial interaction is correlated with a tendency to rudimentary simulation and to a narrow range of attributed beliefs, while an expectation of a rich variety of possibly exploitable situations is correlated with a similar richness of attributed beliefs and of simulatory capacities. To put this last correlation in extremely simplistic terms: the more varied your conception of potentially beneficial interaction the richer your conception of subjectivity.

Right idea, wrong tools?

These four toy examples are meant to show how there can be a mutual pressure between explanatory, normative and strategic thinking. They are also meant to show how from a few fragments and examples we, being human and equipped with normal ethos-constructing capacities, can reconstruct an intuitive sense of a whole attitude to life. This capacity plays a role in fiction, when an author generates a sense of the values and social thinking of a country, a clan, or a single household, with a few well-chosen fragments. Some novelists are particularly good at this. (Ursula LeGuin is a fine example. A remarkable example is Peter Carey's *The Unusual Life of Tristan Smith*, in which two completely imaginary cultures, each with an ethos different from that of any human culture and from the other, are conjured with amazing vividness.) In philosophy this capacity is used invisibly: we describe the constitutions of imaginary Republics and suppose that we have a grasp of what it would be like to live in them. We imagine that we know, for example, whether children growing up under those conditions would have dispositions that made it likely that the conditions would continue.

But there are illusions of imagination. We can be very wrong about what is possible, and, more subtly, we can be wrong about what it is that we are imagining. So when I say 'now imagine what conception of mind people might have if their central conception of social interaction was of a coordination problem' what you then conjure up may depend on many other implicit assumptions. And although we may take these pages, taken with the exercises of political philosophers and novelists, to make it more likely that there is a mutual pressure between values, conception of strategic interaction, and forms of psychological understanding, which results in the existence of stable combinations of all three, we should be much less confident that the factors I have listed really are the central ones. It may be that child-raising habits, gender roles and family structure are much more basic. It may be that one factor, mode of strategic interaction or choice of goods to take as common ends, perhaps, generates the others. An imaginative exercise such as this cannot resolve these questions. The most important idea, though, is that shared resources for explaining action and shared conceptions of the ends of action pressure each other. In adopting a folk psychology we make it harder not to have some values rather than others, and by adopting patterns of social life we make it easier to think of motivation in one way rather than another. That idea, though certainly not proven, would hold even if there were deeper factors shaping the features I have mentioned. There would still be such a thing as a stable and workable ethos.

Exploration IV

Moral progress

> I could only get on at all by taking 'nature' into my confidence, and my account, by treating my monstrous situation as a push in a direction unusual, of course, and unpleasant, but demanding, after all, for a fair front, only another turn of the screw of ordinary human virtue.
>
> Henry James, *The Turn of the Screw*

Moral tectonics

No doubt there are ways of thinking about action, character and motive that humans could learn and use profitably, whose existence is quite unimaginable to people immersed in a folk psychology such as ours. And no doubt there are features of our mind-ascribing that will appear in any context and any culture, superficial though they may intuitively seem. At any rate, unless the arguments of this book are not completely mistaken, there is considerable possible variation. We can appeal to different virtues, to different fine-tunings of 'belief', 'desire' and other concepts closely related to them; we can combine different kinds of simulation in different ways. Where there is variation, there can be better and worse. Some folk psychologies might allow us more understanding of our characters and motives, or be better tools for mediating the business of everyday life. Some systems of values and patterns of moral thinking might allow us smoother ways through moral dilemmas, suggest more humane solutions to our problems, or help us to imagine lives that are, simply, better. As you might expect, I suspect that these two kinds of better-and-worse go hand in hand.

Thomas Nagel has defended a compatible view. Nagel writes

> It is evident that we are at a primitive stage of moral development. Even the most civilized human beings have only a haphazard understanding of how to live, how to treat others, how to organize their societies. The idea that the basic principles of morality are *known*, and that the problems all come in their interpretation and application, is one of the most fantastic conceits to which our conceited

> species has been drawn . . . And the idea of the possibility of moral progress is an essential condition of moral progress. None of it is inevitable.
>
> Nagel 1986, p. 186

There are many ways of interpreting what Nagel says here. He could be expressing a faith in the general improvability of human society. He could be saying that we could develop values and moral theories that better capture what is really valuable in human life, whether or not possessing them actually gave us better lives. He could be saying that creatures with more imagination and self-control than we have could avoid many of the horrors of human histories. He could also be saying what I want to suggest, namely that there are combinations of explanatory and evaluative ideas and practices that would have better effects. People subscribing to them would lead lives which are singly and collectively better than those which we, with our psychological ideas and our values, can manage. That is, that there is a kind of moral progress that requires and carries with it progress in our ways of understanding ourselves.

No one could claim that this is the only kind of possible moral progress. People can invent new institutions, such as representative democracy or companionable marriage, which can potentially allow us to lead very different lives. People can formulate new values, as when they come to think of racial prejudice as wrong or when they come to think of children as having a right to self-expression. (These and other aspects are discussed in Rorty 1991 and Moody-Adams 1999.) And these new values can deeply affect their lives and their relations with others. There is no reason why either of these has to involve changes in the ways we understand and explain what we do. Still, I want to see many such changes as superimposed on a much longer term development of what in the previous exploration I called an ethos, a balance between psychological and evaluative ideas which allows a particular variety of both explanatory stories and normative rules. An ethos emerges from deep underlying aspects of human thinking, which change very slowly. As a result, developments that may seem straightforward in retrospect can take a long time. Ever since late antiquity we have been in a position to see what is wrong with slavery, but it was not until the nineteenth century that the institution became indefensible in terms of commonsense moral intuition. Liberal democracy focussed on individual rights may be facilitated by a specific north-European ethos (perhaps, perhaps), but that ethos can exist for centuries without democratic institutions developing. Ethos moves slowly from one relatively stable position towards another, like the slowly moving continental plates whose accumulated tensions and re-placements are invisible behind the sudden eruptions of earthquakes and the slow growth of mountain ranges.

First easy argument: bloodless theory

One central claim, that moral progress must build on explanatory progress, is very plausible. We won't get a lot further with our lives together unless we understand ourselves better, and can express that understanding in terms that we use easily in everyday life. So that's the thin end of the wedge. Beginning here will make it easier to state versions of the claim that are much less obviously true, and which need more adventurous arguments.

First try to imagine moral progress without corresponding psychological understanding. This might be purely intellectual, philosophers' stuff. Suppose for example that by assiduous work with generalizations and examples, proofs and intuition-pumps, we succeeded in getting a lot clearer about the conditions under which it is permissible to kill one person in order to save others. We could then see clear differences between – to use the famous examples stemming in part from Thomson 1976 – deflecting a runaway trolley towards a single person when otherwise it would run down three, and stopping the trolley by shaking a bridge-support in a way that will inevitably cause someone to fall off the bridge to their death. (Or, it could go the other way round. The analysis could show why our intuitions about these two cases are misleading, and why they should be assimilated to the same category or in the opposite directions to our present intuitions.) But we are to suppose that this is a kind of moral progress that brings no psychological insight. Then there would be something hollow about any claim that we had come to understand *why* these acts are permissible or impermissible. Consider some things that would not follow.

It would not follow that we *understood* the actions of someone who did what, according to the analysis, was right. There are two possibilities. The first is that a person reads the arguments, comes to an intellectual conclusion about what is right and then acts from a conviction that what is right must be done. What we are left in the dark about is how this person would negotiate the tension between this bloodless conviction and the motives that would arise from before she had met the arguments. Faced with a trolley that can be stopped by shaking a pole that throws someone to their death, does the person decide to shake and then decide she cannot do it, or have no problems shaking, or take a deep breath and steel herself to do what is deeply repugnant? Presumably different people fitting the description would react in different ways. But which people would react in which ways; which ways would be intelligible for which people?

The other possibility is that the person does not need the arguments. Their sense of what they should do is in accord with the analysis without the argumentative push. Now suppose that this person's action puzzles some other person. How can we make it make sense to him? We cannot do it by giving the abstract arguments, for they have no bearing on why this person acted as she did. And we cannot say something analogous to 'she realized that if she pushed the pole she would be bringing death onto someone who had nothing to do with the situation, just gratuitously.' For the supposition is that the moral analysis does not give anything that can be cited as an intuitive reason for action.

What is at work here is a very mild internalist principle: a full understanding of why a category of actions is right will allow us to understand a person's reasons for performing an action of that category. (Similarly for actions that are permissible, obligatory, wrong.) This principle is much weaker than internalist principles that are defended by some philosophers, which claim that a full understanding of why something is right must give one a motivation to perform the act. And it is consistent with the externalist insistence that a person could understand very well why something was right and have no inclination to do it. It requires only that understanding why something is right be relevant to understanding why someone who understands its moral character performs the action. (See Smith 1994. Svavarsdottir 1999 gives a powerful and stimulating resistance to Smith. The appendix to Scanlon 1998 warns against seeing too-easy contrasts between internalism and externalism. My mild internalism requires only that the link with intelligibility be so as a matter of contingent human psychology – though one whose absence we cannot easily imagine.)

Knowing *just* that something is 'right' or 'permissible' should not have much persuasive force for anyone whose commitment to doing the right thing is not a mere fetish, as Smith puts it. An abstract analysis that changes the boundaries of a more substantive term, like 'duty' or 'promise' or 'right' (as in 'human right') will have more effect. For then we have some idea how to integrate the thought that the act has the characteristic in question into our existing decision-making processes.

Someone might object that any moral theory that justifies the conclusion that an act is permissible, or obligatory will automatically give a reason *for* doing the action, even though it will not generally give a reason *why* someone acted that way. It will justify though it may not explain. And, the objection might continue, it is perverse to expect explanation from a justificatory enterprise.

But justification and explanation can never be this distinct (as Dancy 2000 urges). When we justify a belief, for example, by saying how it could have been grounded on real evidence in terms of sensible reasoning, although in fact it was learned by rote or prejudice or wishful thinking, we give a route by which someone *could* have acquired the belief. If someone did explain their acquisition by citing this evidence and this reasoning we would understand it. Similarly when we give good reasons for performing an action we say things that could explain why someone did actually perform that action. We might imagine a moral system that used categories and patterns of reasoning that no real person ever could use to explain their actions. It would be a peculiar and perverse system, at best deeply incomplete and at worst deliberately obscurantist. That's the point.

Second easy argument: decision making

When we plan what to do we balance, juggle and dodge between many competing considerations. We balance between the immediate and the distant, the prudential and the impulsive, the selfish and the generous, the principled and the particular. This is hard. No one does it perfectly. Perhaps no one does it even moderately well, compared to the solutions we could achieve if we were a little bit more reflective, a little bit clearer about what we want, and guided by slightly more enlightening ideas.

Our ideas about right and wrong are central in our decision making. But in most people's deciding, justifying and advice giving, there is no clear line between conclusions about what one ought to or should do, because it is the best solution to the problem one is facing, and conclusions about what one ought morally to do. In most people's way of thinking there are specifically moral considerations, concerning rights and duties and principles of fairness, and there are mysterious ways in which these considerations enter into competition with the claims of present and future self-interest, personal commitments, whims and all the rest. If you can sensibly resolve all these competing claims and see a satisfactory adjudication between them then that is what you ought to do. ('You're letting your conscience ruin your life, just because you come from a rich family in a rich country doesn't mean you *have to* give all your money to charities; you *ought* to keep a bit back for your own interests.')

What this suggests is that moral decision making is a special case of decision making in general. And perhaps not a very special or distinguished case, in terms of the abstract structure of the problems it faces. (I have argued this at greater length in Morton 1991, Chapter 11, and Morton 1996b). In both moral and non-moral decisions we have to face problems about incomparable wants, about conflicts among our beliefs and our desires, about expected changes in our desires, and with risk. No one knows all the good ways of getting safely through these problems. No one can say all the ways, good and bad, that people find their way around this territory. So in both normative and descriptive terms our grasp of decision making is extremely incomplete. And this incompleteness is found both in explicit philosophical accounts and in folk psychology. We can only articulate a few of the ways we have of making decisions. So there is a profound imperfection in all existing folk psychologies. There might be better folk psychologies that do not incorporate better ways of describing decision making, just as there might be better systems of moral ideas that do not. But such improvements would be very partial: large serious improvements in either folk psychology or moral ideas would have to include ways of describing the possible routes a decision can take.

This does not amount to a claim that if we became better at describing decisions we would inevitably become better at solving moral problems. More the other way around: in order to become better at solving moral problems we will have to have better ways of describing decisions. And this means both better criteria for satisfactory solutions to decision-making problems and better ways of relating the decisions people actually make to norms of good decision. Decision expertise will always be compatible with rotten motives; but noble motives will also always be compatible with moral incompetence, with well-intentioned bungling that leaves situations worse and creates more dilemmas than it solves.

The most general conclusion here can be put in terms of reflective equilibrium (in the Rawlsian sense elaborated in Daniels 1996). The ideal mesh or harmony of ideas involves a three-way fit. A system of general moral principles and categories will only mesh with our judgments and intuitions about particular cases when it is associated with a suitable set of decision-making norms. For it is decision that links the other two. We should want the link between the two to be as strong, flexible, and accurate as it can be. A wide reflective equilibrium should include a central part of folk psychology: the description of the thinking that leads from facts, values and preferences to judgments, intentions and actions.

A harder argument: turning theory into practice

Now we come to arguments that depend on premises that are harder to state, and when stated, harder to be sure of. The first concerns the particular features of present day folk psychology and present day moral thinking that we might expect to be changed by better ways of thinking. Suppose we could see links here, ways in which clearing up moral questions on which we know ourselves to be deeply confused would plausibly require thinking more clearly or more usefully about motives, emotions and character. Or links in the other direction, making connections between areas in which our present folk psychology seems not adequate for our needs and areas of moral controversy. Then we could make a much more practical argument for the connection between moral and explanatory progress: in order to move from here to an obviously desirable there we would have to take this specific route.

Our present system of moral ideas presents conceptual puzzles for philosophers and practical problems for responsible people. There are many topics on which practical uncertainty and philosophical puzzlement coincide. I discuss just one, which centers on deontic intuitions, intuitions that under some circumstances one should not choose the action which has the greatest balance of desirable over undesirable consequences. The cruder intuitions concern acts, such as killing innocent people, which we have enormous inhibitions against allowing, and the subtler ones concern sets of situations such as the trolley examples already mentioned in which superficially similar cases seem to call for very different ways of evaluating the consequences of acts. Medical ethics would hardly be a subject if we did not have such intuitions, and did not find them puzzling. The general consensus among philosophers is that we should not try to find reasons why some actions should in no circumstances be performed. Rather, we need to re-think what is going on in the particular cases which evoke the intuitions. There are a number of suggestions abroad about how we might do this.

One suggestion, due to Thomas Nagel, traces deontological intuitions to the need to acknowledge the individual point of view of people affected by an action. The consequentialist attitude, on the other hand, represents the demands of objectivity and impartiality. For example if an action will save several people from death by causing a single person to die, the outcome with only one death is objectively better, but from the perspective of the person whose death will be caused it may be a much worse outcome. Another suggestion, due to F. H. Kamm, traces the contrast to the closeness of the connection between an act and its bad effects. Here is a simplification of Kamm's complicated account. People have rights that prevent certain kinds of reason from justifying their being made victims of various kinds of harm. As a result an act of causing one person to die in order to save five has to be justified in a way that overcomes the one person's right to life. This can be done if the aspects of the act that cause the one death also save the five lives. But if the act causes the one death in one way and saves the five lives in another accidentally and inessentially connected way, then it violates the person's right to life and is not permissible. In this way some balancings between the greater and the lesser numbers of deaths are allowed and others are not. (See Kamm 1994, 1996, Kamm and Harris 2000. Rakowski 1998 gives a very helpful exposition of Kamm. It seems to me that Kamm's basic argumentative strategy has a lot in common with Scanlon's approach to issues of freedom of speech. See Scanlon 1972. Also very relevant here is the suggestion of Horowitz 1998 that deontic intuitions are illusions which can be explained in a way that undermines their validity. See also Chater and Oaksford 1996 and Cummins 1996.)

There are other suggestions, of course. And these two need not be rivals. Patterns of justification satisfying two superficially opposed explanations of what lies behind our intuitions could both be right. For example, agent-grounded rights – requirements that a person's own view of the situation be taken account of – and victim-grounded rights – requirements that acts bringing harm to individuals have certain kinds of justification – could both exist, both covered by our present vague concept of a right. What matters for present purposes is that both kinds of right if incorporated in a development of our present patterns of moral thinking would necessitate a corresponding development of our resources for explaining human action.

Suppose that Nagel's way of understanding deontological considerations became a salient feature of moral discourse, rather than the underlying thread reconstructable with a bit of imagination and insistence that it is now. In many situations people would then ask themselves about the priorities of some person affected by the act they are considering. (This would make a particular difference in cases involving the valuing of life.) They would, in effect, invite an imagined voice of that person into their decision making and allow that person to suggest valuations of outcomes and to recommend or argue for the exclusion of options. They would be acting in accordance with Kant's requirement that '[reason's] verdict is always simply the agreement of free citizens, of whom each must be permitted to express, without let or hindrance, his objection or even his veto.' (*Critique of Pure Reason*, A738–9, B766–7.) Or as Christine Korsgaard (1996, p. 348) puts it after applying this passage from Kant to the case of lying, 'the deceived person has not had a chance to agree or to cast his veto.' It does not matter that the injured person is not actually party to an actually shared decision; what makes the act wrong is that this person would have vetoed the action were they party to a decision. The most basic right is the right to be listened to, even in one's absence. (Compare this with the guiding thought of Scanlon 1998, that an act is permissible when it can be justified to others on grounds that they could not reasonably reject.)

If this were a distinctive feature of our practice some explicit patterns of argument would have to be developed, probably with suitable vocabulary. For example, the eligibility of the implicit participant would have to be disambiguated: is she someone who merely has an interest in what is being planned (as often with rights to be kept informed) or someone who will be directly affected by it (as with rights to life)? Linked to this, different ways in which a person can share in a decision would have to be distinguished. Some implicit participants would add options to be considered, others would rule options out absolutely, and yet others would contribute a valuation of outcomes that had in some way to be weighed against those of the decision maker herself. Thus we would be able to say 'person p is involved in this situation in way W and as a result our decision making is constrained to take the following form'. One possible vocabulary for expressing this, but certainly not the only possibility, would be to have terms like 'veto-right', 'commenting observer-right', 'expert witness-right', and so on. And so in thinking out what to do we might have thoughts which we could express along the lines of 'Now since we will all have to take over some of Ellen's students while she is in hospital, the obvious solution is to increase the size of the seminars. She wouldn't approve and as a member of the course team she's implicitly here. So there's one vote against the measure, anyway.'

But this vocabulary and the associated patterns of reasoning would not be confined to resolving moral problems. Whenever people actually made a decision together they would use the terminology to specify which person was occupying which role in the decision. They could do this whatever the topic and whatever their motives. It could apply to committees of robbers. And the terminology and patterns of sharing could be applied where the decision is shared with an implicitly present person when there is nothing traditionally moral about the purpose. 'One way in is to come in through the side door and shoot the guards guarding the front, but Buggsy has a say in this. He's a commenting observer even if he is in prison. He'd want to avoid the side door because he promised his old mother he'd never be associated with shooting anyone in the back. So I guess we'd better come in through the front and hope we can shoot the guards before they see us.' In fact, such extension from what seems to be a matter of moral decision, by our present intuitions, to what does not, would be essential. For suppose that one is making a decision which, as we now think of it, considers the rights of an affected person, and while one is taking into account the interests of that person, according her an appropriate role, she herself turns up. She could demand that she play that role for herself. That request could hardly be denied, and the role the actual person plays will surely be the same as the role the representation of her interests plays if her presence is only implicit.

So this moral change, a candidate for moral progress, would bring changes in the way we perform and describe decisions. And hence changes in the way that we explain and predict actions and changes in the thoughts that we are disposed to attribute to people. That is, changes in folk psychology. There would also be more diffuse consequences for our styles of explanation. For example, we might develop ways of explaining someone's actions or state of mind in terms of an internal dialogue between that person and an implicit other with their own perspective on a situation. These could turn out to be the deeper and more pervasive consequences. And in fact there is a very fundamental reason for expecting this. For the real bite of deontological ethics is not in the appeal of rules that must never be broken. It is in the force of the idea that there are some things that you just cannot do to someone, an idea that can strike the most resolute consequentialist in particular cases with dogma-overwhelming force. And since this idea is as Nagel says a particular way of grasping the perspective of another person, it is by its nature psychological. It is a sense of a way of grasping another person, and thus inevitably suited to be common to folk psychology and ethics.

Kamm's alternative way of thinking deontology suggests a different set of psychological ideas. I shall be briefer, though. The conceptual core of a Kamm morality might be a set of predicates describing different ways in which an action can cause or prevent an event. They would differ in the ways and extent to which the causation or prevention could be taken to be an essential part of the action. For example, we might speak of an action 'causing E *while* preventing F' or, contrasting with this, 'causing E *in* preventing F'. (Perhaps more vividly: 'causing E while at the same time resulting in F's not occurring' versus 'causing E as part of preventing F'.) I am supposing that we would have a clear understanding of the difference between these, as we don't from our present day perspective. But a consequence would be that pulling a switch redirecting a trolley from a track headed for five people to a track headed for one would cause the one death while preventing the five deaths. On the other hand redirecting the trolley by juggling a bridge support in a way that causes a person to fall to their death would cause that one death in preventing the five. These classifications would be used as input to two sets of explicit moral principles of the form: inflicting a harm of kind H on a person can or cannot be justified by greater benefit B, where the infliction causes the harm while (or, in the other set, in) causing B. The result might be a set of clear and usable verdicts coinciding in many cases with the confusing patchwork of intuitions from present commonsense morality that are evoked in Kamm's writings.

(It is important to gesture towards practices that we cannot now fully understand. If the aim is to show how there could be drastic changes in our thinking which link folk psychology and ethics then it is important to use some examples that point into the fog. For it wouldn't be real conceptual change if we could make perfect sense of it now. It is a bit like arguing that changes in physics might lead us to changes in mathematics: one has to gesture towards a path we don't yet know how to follow.)

The psychological aspect of a Kamm-inspired development of our morality would consist in the fine distinctions between kinds of causing and preventing. There would be obvious work for these in ways that have nothing to do with morality. For example in describing people's intentions we could use the Kamm-inspired terminology quite literally. 'Did you mean to show everyone you are back in town *in* coming to chat with the secretaries?' (Because in that case I think it won't work tomorrow when things will be different.) 'I hope that the most you wanted me to do was to distract the director *while* making his wife think I was flirting with her'. (Because if I had done more than that I would have been putting my job at risk.) And these refined classifications of intentions could play a role in the kinds of plans people formed and the judgments they made of one another's characters. The roles would have to be helpful and enlightening ones. If not, the part of the moral enterprise that was bound up with the classification of intentions would appear as arbitrary and pointless.

Another use would be in the delegation of tasks. Suppose that one person, the agent, agrees to accomplish some aim for another, the principal. The agent will agree to bring about some situation A, perhaps also agreeing not to bring about some other situation B. She is to be a fairly free and initiative-taking agent, and will thus make her own evaluation of means towards A. If she finds that in order to achieve A it would help to have brought about C she will consider bringing about C, and will consider the pros and cons of ways of doing this. But she has to avoid means to C that produce B. And of course if some situation D is such that A would be more probable if it occurred, but that producing D would not tend to bring about A, then there is no point in considering ways of achieving D. (This could be because A and D were both effects of a common cause.) The result is that the agent will consider means to A as long as they do not fall into various excluded categories. There will be a play between the two excluded categories just mentioned. Sometimes considering how to bring about an end A the agent will become aware of possibilities that will make A more likely but which will also make B more likely. The tension can be resolved if such a possibility can be split into a means C which advances A and does not itself bring about B and a by-product D which does not itself bring about A but which does bring about B. Then the agent can consider ways of accomplishing C but not ways of accomplishing D. She makes A more likely *in* bringing about C but *while* bringing about D.

Consider a standard example. A group of people trying to escape a burning building has picked you as its leader. The way is blocked by a large person frozen with panic. A means to getting the group out is to push this person out of the way. That will make it likely that he will die. But his death is not itself a means to the group's escape. So you can consider different ways of pushing him, but not different ways of causing his death.

The upshot is that when an agent is commissioned to accomplish a task for a principal the commission also binds her to the kinds of means she can consider. A criterion of causal relevance obviously applies, as does a requirement not to investigate ways of achieving forbidden ends. These two are likely to be combined, in the form of criteria that determine the admissibility of means in ways that combine their causal relevance and their relation to desired and forbidden ends. The very simplest case is that in which the commission says 'do A, and under no circumstances do B'. It is easy to see how in accepting such commissions we both make it possible to share tasks without having to worry about what others may do in carrying them out, and make our actions intelligible to one another. We make it easier to predict what one another will do by undertaking not to do whole classes of acts. So in this simple case it is clear how a practical need for mutual intelligibility leads to deontology. But people leading complex lives generate much more complex cases.

The deontological strand in moral philosophy tends to presuppose, in effect, that there is one inevitable or most natural way of formulating these conditions of admissibility. The presupposition is that when an agent engages to accomplish some things and avoid others it is obvious – on enough reflection – what means the agent is thus allowing herself to consider. Perhaps this is so; I don't think I have to take a stand on the question. The important point is that when, case by case, agents do take on tasks they do also take on restrictions on the means that they will consider. Indeed it would be impossibly risky to delegate a task were this not the case. Agents will very often draw on subtle learned routines which package complex restrictions on deliberation – *in* and *while* – which principals assume to be part of what is understood when commissioning others to perform tasks.

The links between moral and psychological progress are clear here. At present when we understand the actions of someone who has performed a task on someone else's behalf we appeal to an implicit understanding of the restrictions that went with the commission. And when we deliberate about our obligations in the context of such a commission we struggle to see what those restrictions entail. Were we to understand better, more explicitly, how the causal and the evaluative interact in these restrictions, we would both be able to formulate clearer analyses of moral dilemmas and their solutions, and be able to articulate more clearly some of the reasons why people acting in implicit contact with others choose as they do. As always, moral thinking is continuous with all the other thinking we do when we work out, individually and together, what to do.

(This second folk-psychologizing of Kammian moral theory makes it more like Nagel's version. For it ties deontic reasoning in another way to shared decision making. It also suggests connections both with issues about the relative roles of cause and probability in rational decision – as in Joyce 1999 – and with issues about the reasoning that is involved in carrying out an intention – as in Bratman 1987, 1992.)

I believe that similar exercises can be carried out for a number of topics where our present day moral ideas are clearly in a state of controversy. Examples are issues about the limits of obligation, about the relation of the good life to the moral life, and about reconciling moral sincerity with tolerance for the convictions of others. In all of these, philosophers have made suggestions that can be turned into speculative possible moralities, which need a corresponding possible folk psychology. And I believe that there are examples running in the opposite direction, areas where it is our present folk psychology that seems inadequate, and more smoothly running variants on it would generate different moral possibilities.

The most convincing such area is change of desire. Although commonsense is not hostile to the idea that people change their desires in intelligible ways, it does not provide very many explicit procedures for describing or explaining such changes. Yet there clearly are systematic ways in which desires change. Two philosophers who have described such ways are Henry Richardson (1994) and Elijah Millgram (1997). Millgram distinguishes between 'real desires' and various desire-like motivational states. Real desires are grounded in experiences of what will bring satisfaction or what will satisfy basic needs, while desire-like states seize people in various ways without connecting to anything that would lead the person to endorse them. Or, as Millgram puts it, real desires can coexist with knowledge of how they were obtained, while pseudo desires are undermined by such knowledge. Richardson on the other hand argues convincingly that moral insight often requires that we understand when it is reasonable to change our desires. Combining the two ideas, we get a sense of how we could better understand the ways in which we do in fact change our desires, and as a result acquire the resources to describe which among them are the ones that a responsible moral agent will try to follow. (The link between rational change of desire and moral obligation is frequently re-discovered. See Grice 1967, Bond 1983, Stocker 1989, Chapter 9 of Morton 1991, Richardson 1994 and Schick 1997.)

None of these imaginary developments – those inspired by Nagel and

Kamm or those inspired by Richardson and Millgram – may be realizable. And if they are they might not be usable additions to common sense. But that isn't the main point. The point is rather that when you look at the problems of what we have now with a revisionary eye you begin to imagine changed psychologies hand in hand with changed moralities.

Evil

One might wonder whether this is an argument that changes in folk psychology could bring *progress* in our moral thinking. What makes the changed situation better? One answer starts from the uncontroversial fact that we are all confused about resolutions of the deontological-consequentialist tension, about the limits to obligation, about the relation between desire-satisfaction and the good, and about resolving conflicts between first and second order desires. Anything that makes us less confused on these topics counts as progress. Take that as a stipulative definition if you will. And if you really suspect that resolving moral dilemmas leads to superficiality or dogmatism, accept only that this progress is not trivial or easily achieved, and would mean having intellectually and intuitively satisfying answers to questions that now trouble us deeply.

But there is another reply, which focusses less on the need to understand. Our culture's ethos has left us in no doubt of its moral failings. People sharing fundamental values and explanatory procedures with the author and readers of this book have participated in genocide, been blind to racial prejudice, and allowed millions to starve. As Nagel (1979) puts it, it is simply a matter of moral luck for most of us that our grasp of our own motives has not been tried. Any ethos that grappled more effectively with the factors that permit evil would be an uncontroversial improvement.

Given a suitable diagnosis of the sources of evil, the Nagel- or Kamm-inspired folk psychological developments discussed in 'A harder argument' above would grapple with exactly these factors. The diagnosis derives from Arendt and Serenyi, and is surely at the very least on the right track. (See Arendt 1964, Arendt 1994, Sereny 1995 and Chapters 5, 15, 16, 19 of Sereny 2000.) It assumes that mass evil such as the Holocaust requires three factors. The first is an ideology that allows a few deluded or ruthless people to wrap themselves in principles that seem noble and in fact justify horrors. There may be a great variety of religious or political doctrines which can serve this purpose. The second is a supply of sadistic or psychopathic individuals to carry out deeds requiring either an intense lack of imagination or a positive devotion to the suffering of others. And the third, the most important, is a social ethos. This ethos leads a large number of people's efforts to be cooperative and responsible, indeed to act on moral principle, and allows them to participate in acts which were it not for the presence of the ideologists and the sadists they would not dream of.

The first two factors may seem to fit the traditional demonic image of evil, but it is the third, on this diagnosis, that is essential. (I am sure, convinced by Arendt and Sereny, that the demonic image diverts us from what we really need to understand. But for an eloquent evocation of that image that connects it with themes in the philosophy of mind see McGinn 1998. See also Morton 1998 for reservations about McGinn.) For large scale horror the cooperation or at any rate complicity of a large number of people is required. These people do, or allow, what they do because they are good citizens and obedient employees or soldiers, and because the consequences for them of not being malleable are dire. It is this last that makes the connection with the issues about deontology.

Imagine a person who is ordered to participate in killing innocent people. (My description is based on Sereny's account of Stangl.) Imagine that though this is not work that he would have freely chosen, he knows that if he does not his own life or those of others will be at risk. He gives in and participates, finding some psychological accommodation to his initial revulsion. His reasons for participating are easy to understand. If he does not, he or others he cares about will suffer, and the killing will not be seriously impeded. Now imagine someone in the same situation who refuses to participate, accepting the likely consequences for himself or others. His reasons for refusing are also easy to understand. He does not want to participate in murder.

Though the reasons of the participator and the refuser are both intelligible, without any expansion of our motive-deciphering capacities, the *difference* between the two people is not easy to understand. Both were, we can assume, in the same situation and facing the same risks. Why did the one but not the other participate? What was the relevant difference? The refuser followed through some reasoning that the participator did not, which not only led him to the belief that it would be wrong to participate, but made him not want to participate. The refuser could understand this reasoning, but something in his problem-solving dispositions did not lead him to adopt it for his own decisions. What?

It is important to see that the moral problems the participator and the refuser face are very hard ones. For refusal brings enormous risks not only to the refuser, but also to his family and associates, and does not necessarily save any lives. Would you refuse to participate in a crime if refusal meant the death of your children, and would only prevent the crime if it were joined by the very unlikely refusal of many other people? If you did refuse might you not be guilty of saving your own peace of mind at the expense of the lives of others? (So one of the things that makes the problem hard is that the acts in question are one-by-one defensible, but taken in series, and multiplied by many participants, they amount to something from which anyone whose vision had not been darkened by participation or by ideology would recoil.) In Kamm-ish terms the problem involves asking whether in participating in the atrocity one would be involving oneself just by protecting one's loved ones, or whether one would be involving oneself while managing to protect them. In the former case it might be defensible and in the latter not. In Nagelian terms the problem involves asking whether the potential victims of the atrocity are eligible to have an implicit say in whether one should sacrifice their interests to protect those of one's loved ones. (Obviously neither of these captures the whole problem.)

If we want to understand why two people adopt different solutions to one of these hard moral problems, we need to see the reasoning they follow as continuous with general facts about their thinking and character. We need to see how the characteristics that led to these solutions reveal themselves in many other acts and decisions, including ones where the problem would not usually be described as moral. So, we need the explanations to emerge from a folk psychology that makes some enlightening articulation of the major moral problems continuous with the explanation of innumerable little everyday acts. But our culture does not give us such explanations, and as a result we are incapable of explaining the difference between the two people.

A culture that did afford these explanations would be very different from ours. The difference would not just consist in the presence of some subtle vocabulary for describing intentions, or some novel decision-making procedures. For the visible continuities between the major moral issues and the little everyday acts would make an enormous difference. From a person's way of, for example, deciding which appointments to keep and which minor annoyances to take seriously, others would be able to project to her behavior when subjected to ultimate tests. Then, presumably, the patterns of thought that were continuous with resistance to evil would be encouraged, long before the testing situations arose. These differences would permeate everyday life. They would, that is, if such a culture were possible, and we certainly do not know that it is.

The knot: moral competence

The link between everyday psychological explanation and moral thinking was at its clearest at the level discussed in Chapter 1. There a primitive stratum of moral thinking – what I called microethics – had an easily described overlap with a primitive stratum of psychological attribution. The connections were less direct with the topics of later chapters, where the emphasis was on coordinated understanding rather than on better and worse action. Still, the examples often had a moral flavor, one reason for which was that a coordination is unlikely to be sustained for long if it is not in the interests of those concerned. Now is the time to bring together some of these free-floating strands, in a definite if speculative knot, to end this chapter and indeed this book.

Consider a suggestion about the relation between folk psychology and ethics. It is that a fully developed moral competence – one that approached that of the ideal moral agent – would require ideal understanding of motive and action. And that a person who had ideal understanding of motive and action, enough to make her an ideal participant in human affairs, would not lack any *understanding* in order to be an ideal moral agent. Such a person might be morally numb, or in the service of the devil, but would know all she needed to know in order to do right. The suggestion is that the link between everyday moral and psychological understanding that we find at their elementary forms returns at an advanced level. It is only in-between, where we live, that their relation is subtle.

The plausibility of the suggestion comes from the lack of a special subject matter for morality besides human interaction. The deeper your capacity to deal with people, the more likely you are to be able to do the right thing by them. Or so it might seem. At the same time, it is clear that psychological insight alone is not enough to make moral excellence. Leaving aside the case of the person who has deep understanding of others and also wishes them harm, if only because it is so hard to think about how deep a malevolent understanding could be, there remain many other cases. For example a person whose intentions were often admirable but who often failed to carry them out because of a failure of nerve, an inability to carry through under pressure or threat, would not be morally admirable. She would not be someone one could trust with anything important. But she might not be unusually deficient in *understanding*, either of human motives or of right action.

With examples like this in mind, the best way to formulate the suggestion is as follows: folk psychological understanding is the only general conceptual requirement for acting well. (Though there may also be many requirements of motivation and character.) Or, with a different emphasis: the more understanding a person has of the factors relevant to interaction with others the more she has of the understanding needed for morally admirable action. And there is no morally relevant understanding that would not find its place in a steadily increasing understanding of how to interact with others. This latter formulation has the advantage that it constrains the content of folk psychology more narrowly than it might be taken by some, for instance by someone who thought that all of psychoanalysis or of neuropsychology was potentially folk psychological. To box the suggestion's content in from another direction, take 'understanding' in a fairly wide way, to include all uses of propositional knowledge, and of information-exploiting cognitive routines such as simulation. Take 'moral' in a fairly wide way too, to include all considerations about better and worse in shared activities and individual lives. The suggestion then is that on a wide construal of understanding and of the moral, a full understanding of moral choice would fit within a narrow construal of the folk psychological.

The suggestion is not going to get more definite than that. It is still not sharp enough to sustain a definite proof. It is sharp enough to be wrong, though. And there is a natural and formidable objection. Deontology again. Suppose that there are situations in which moral competence requires that one not weigh up good against bad consequences but proceed differently, sometimes avoiding an action though its balance is very positive. Couldn't a person have all the understanding of motive and action that a person could have, and not know when or how to do this deontic thinking? It is not immediately evident that the intellectual resources needed to solve the kinds of moral problems that evoke deontic intuitions consist just in people-understanding. Some specifically moral understanding may be needed. (If deontic intuitions are illusions then there is no problem here to handle. I shall assume they are not, and I suspect that my defense against the objection from them will add to the plausibility of the suggestion even for hard-core consequentialists.)

For all that, I think the suggestion stands. Consider what a person would be like who could see her way easily through such problems, on either the Nagel or the Kamm understanding of them. On the Nagel account such a person would know when and how to incorporate another person's actual or simulated vote into her own decision making. She would be an expert in implicitly shared decision making. That is certainly a folk psychological skill that mediates social interaction. It consists in being able to imagine the claims that another person would make in a deliberative context, and it allows a person to carry out, for example, a task that another has entrusted to them without referring back to the other when unexpected developments occur. It gives the intellectual skills to be trustworthy. And it consists entirely of such skills.

Consider now the Kamm account of competence at deontic problems. A person who had these skills would be able to differentiate in a specific way between possibilities that are a means to an end and acts that are anticipated consequences of a means. As I argued in 'A harder argument' above, if these distinctions are morally relevant they are also psychologically relevant. They arise out of the capacity to plan the details of a project whose general aims have been determined in advance. So knowing one's way around the distinctions in moral cases would entail knowing how to apply them to the actions of others in moral and non-moral contexts. Enough explicit moral understanding would be a step towards psychological insight.

It may seem that on both the Nagel and the Kamm versions there remains an irreducible moral residue. A moral agent has to know when a potential sharing of decision making is appropriate, or when to regard himself as bound by an implicit principal–agent commitment. But consider what it is that the moral agent knows in knowing these things. The criteria that make it appropriate to share a decision – criteria that for example give the potential victims and not potential bystanders of an act a say in whether it is to be performed – are not arbitrary. They must be criteria that would have been accepted by sufficiently reflective people given sufficient experience of life. So a fully competent moral agent would be able to reproduce the reasoning that leads to this acceptance and come to the conclusion that, given the situation as it is, the decision is to be shared with some individuals and not others. And there is nothing special about this reasoning, except its difficulty; it also gives insight into why people do share some decisions and not others, by completely or partially following the reasoning. A similar point applies to restrictions on deliberation about means. A fully competent moral agent would be able to understand the reasons that people in his culture would accept these restrictions, and thus expand the reasoning that presupposes them into reasoning from prevailing conditions to the results of the restricted deliberation. In both cases the theme is: for a fully competent moral agent specifically moral considerations would be just a short-cut, expandable into wider reasoning shared with other such agents about how to achieve common ends. And the capacity to engage in this wider reasoning would be part of what was required in order to understand principal–agent behavior in general, with instances labeled 'moral' as a special case.

I think an honest summing up of the situation is this. The problems that produce deontic intuitions are among the hardest that we face. Hardest to find solutions to that we will not regret, and hardest to understand the solutions others stumble onto. We cannot describe straightforward routes through them. But it is reasonable to suppose that they focus on ways that people can set priorities and restrictions on shared activities. So when the trolley of your life is hurtling towards a juncture from which various consequences for various other people will flow if you do or refrain from various actions, you first consider how each of these consequences are not merely consequences, and then in evaluating ways in which they could be brought about you simulate the contributions the affected people might make to the discussion, weighing all the competing desiderata as best you can. This is hard enough, but it is only the beginning of the more general explanatory task of seeing when patterns of reasoning like this, but bungled, incompletely performed, inappropriate, or based on mistaken assumptions, lie behind particular people's actions. But in the end all of this must be faced by both moral and practical psychological competence.

There is another, less cautious, way of putting this conclusion. The competence I have just been describing is a description of a wise person. The claim has been that the wisdom demanded by problems that are normally classified as moral is the same as the wisdom demanded by problems of understanding other people for the sake of practical human interaction. The wise person – the ideally wise person – would know how people individually and collectively make choices shaped by needs they cannot articulate and regrets that they do not acknowledge. Such a person would at the same time have ideal moral and psychological competence. For every moral capacity there is a folk psychological capacity of which it is a special case, and for every folk psychological capacity there is a moral problem to whose solution it is essential. But it does not follow that in actual agents of limited wisdom psychological and moral competence are so easily related. With many moral capacities the folk psychological capacity may be undeveloped; the conditions for its realization may not have developed. In these cases we learn the moral routines by rote; they remain intuitions not backed up by reasoning that we can apply more generally. And with many psychological capacities the moral problems they could solve may remain un-tackled. The fusion of the moral and the psychological may require progress that is yet to come, or that will remain forever beyond us.

Notes

1 Microethics

1 I would prefer to use 'folk psychology' as meaning 'whatever beliefs, skills, or other cognitive processes human beings use in everyday life to predict and understand actions'. Certainly we need a term with this wide non-question begging range. (And I would rather that the term had been 'vernacular psychology' or the like, but that battle has been long lost.) However, many writers use the term in a much narrower way, to apply only to prediction and explanation in terms of beliefs. (The beliefs may be implicit, and not easily stated in terms available to the person who uses them, but the assumption is still that the thinking involved is the kind that applies to beliefs: inference.) In effect they commit folk psychology to what in Morton (1980) I called the 'theory theory'. The psychologist's term 'theory of mind' is even more obviously loaded. I'd like to see 'folk psychology' used for the wider meaning and 'theory of mind' for the narrower one. But, since I cannot command this usage I shall rarely use 'folk psychology' in this book without some accompanying gloss.

2 Dennett also comes out as a realist, but for different reasons. On his view it is impossible for beliefs and desires, at any rate, not to apply to people, as long as ascribing them is successful in predicting what they will do. But in most of Dennett's writings the success of these predictions is not at all the result of the fact either that the intentional stance is being taken towards the people in question, or that they themselves are taking it. At least in none of the essays in Dennett (1987) is this possibility suggested, though in recent conversation Dennett has raised the possibility that the intentional stance may have self-fulfilling properties.

3 Our intuitions about the application of our familiar psychological concepts are clearly very suspect in this context. If the second, conventionalist, position were correct then these intuitions would be the by-product of the socialization process that generated the illusion of inherent applicability. Empirical evidence is no more likely to settle the question in the short run. Anthropologists studying vernacular psychologies try to get a grasp on how widely the human understanding of human beings can vary. So far the variety of opinions is as wide as the spectrum we are considering. Developmental psychologists studying the acquisition of concepts of mind work largely with children subject to the forming influences of a culture's explanatory practices and microethics. And most of their work, partly because of an obsessive concern with the concept of belief and the false-belief test, is done with children who have already been initiated into particular patterns of shared activity. (But see 'Joint attention and mutual knowledge' in this chapter in this

connection.) Eventually we may hope for evidence that will help resolve the question. As it trickles in we should try to see it in as unprejudiced a fashion as we can.

4 Saying that the filter must have this kind of output does not go very far to say what forms the filter can take and what outputs it can have in specified situations. I take this to be an instance of the frame problem: we can apply usable classifications to situations but we do so in ways that are so sensitive to details as to resist easy formulas. See Boden 1990, Chapters 7–9. Thanks to Peter Goldie for making this connection.

5 Why is language-use strategic, when only one person at a time speaks? Because the effect of what a speaker says depends on what the hearer makes of it. So contributions to a conversation alternate, each pushing to different practical and communicative outcomes in part by each shaping the interpretation of what has come before. Children's developing grasp of strategy in conversation is as far as I know completely unstudied. This is not surprising given that we do not have an enlightening analysis of the kinds of strategic situation that arise in language use, as opposed to strategy in language formation – convention – which we do dimly understand.

2 Motives and virtues

1 In correspondence Zagzebski accepts these as friendly re-formulations of her intentions. A further issue concerns the 'enduring'. One could argue, with Harman 1999, that there are few dispositions that a person exhibits over a long time in a range of contexts. In fact, most of the explanatory uses of virtues I discuss could easily be adapted to a concept of virtues-on-an-occasion.

2 The distinction between contained and uncontained is relative to a conception of human capacity. Though it is not needed for the argument here, I would defend a conception according to which real time and memory limitations should be taken into account, so that for example the problem of determining whether one arbitrarily chosen sentence is a truth functional consequence of another is uncontained. See Cherniak 1986.

3 In ignoring the hard meta-normative issues here I hope not to appear to be dismissing them. See Raz 2000, Hookway 2000. My own view is that in calling a solution better than another we are simply reporting or predicting that a considered and shareable reflection on the matter would endorse it. If this is a context-sensitive and ambiguous business then so is rationality. See Gibbard 1990. Whether or not this is right, in exploring or creating the structure of solutions we are developing a framework against which our present and future actions become intelligible. See also Cohon 2000.

4 This is a very very mild analog of internalism about moral motivation. To understand a virtue is to see that in some circumstances you would wish to have the virtue. The connection need not hold as a matter of meaning: it can just be a matter of psychological fact. For full-blown internalism about the moral see Little 1994.

3 Belief and coordination

1 There is no single concept of almost-knowledge. One such concept is given by Williamson's technical term 'opine' (Williamson 2000, pp. 44–7) which holds when a person has an attitude to a proposition that she cannot discriminate from knowledge. When you opine that p, then for all you know you know it. Another is

'in situations that could easily have occurred the person would have known that p'. And there are others. This is a basic reason why inasmuch as we understand 'believes' as 'in a state much like knowledge', belief is a very slippery concept.

2 When the factors that best illustrate intensionality – identity, nonexistence – enter, knowledge itself acquires some of the deviousness of belief. I would like to think in terms of a basic very fact-linked concept of knowledge which then gives rise to a concept of belief which then adds twists and complications to result in our full concept of knowledge. Knowledge of identities is a good example. On a basic externalist and extensional concept of knowledge no one could be ignorant of a true identity between objects they are acquainted with. For the truth about Superman and Lois Lane see Saul 1997, Barber 2000.

3 There are two points here. The familiar one is that belief comes in degrees, and that there is no definite threshold when suspicion does or should become belief. The more important point now is that there are cognitive processes leading to all the intermediate levels of confidence. A small initial segment of the evidence produces your first glimmering suspicion that a repeatedly tossed coin might be biased. That process is a representation-producing process just as the one that produces belief is. And if the coin is in fact biased the result is knowledge. Some systematic processes lead to the states of: taking seriously the possibility that the coin may be biased, having an inclination to think it actually is biased, thinking that it is worth checking further for possible bias, taking it as a working hypothesis that it is biased, and so on. All of these are processes that lead to definite states, which represent something like knowledge. So you have something like knowledge that the coin is biased long before you believe it.

4 The example comes from Veyne 1988, a book full of fine examples and stimulating problems, wrapped up in a naivety about what it is to address a hard problem that will drive any philosopher wild. A similar line with very different examples is taken in Shapin 1994, which will also both stimulate and annoy philosophers. Abimelech, incidentally, was a son of Gidon who murdered several of his brothers to become king. See Judges.

5 The idea of the 'same' action can be understood in a general way here. Whenever there is a 1–1 function from one person's options to those of another such that the resulting pattern of outcomes has the structure of a coordination problem, we can regard the argument and value of the function as the same action. This is worth pointing out because it shows the range of situations that are really coordination problems.

6 A more radical view is that when we ascribe beliefs we are not representing believers' states of mind at all, but making assertions on their behalf. 'George thinks there are buffalo over the hill' is taken as 'Reporting for George: there are buffalo over the hill'. This alternative is explored by Christopher Gauker, most recently in Gauker 2002.

5 Learning to simulate

1 It is not even essential that the procedure be used to make decisions as well as to anticipate those of others. For one person can think through what would be sensible for a group of people to do, knowing all along that she herself would not perform her component action (from weakness of will, perversity, or imminent death.) Or one person can simply think as if she were a member of a group of interacting people. This doesn't require any complex assumption of beliefs or desires that she does not have. She simply has to think through the course of action that would be in the best interests of any one of the people concerned,

using some procedure that she can expect them to share. Then she forms expectations about what that person, and the others, will do.

2 Take ‘simulation’ to be a helpful label when one person gets information about another person by using some process that works because of some relevant similarity between the two persons. There is then a contrast with the use of verbally encoded information learned from others. There is also a contrast with the use of any propositional information whose manipulation is largely a matter of inference from learned facts. But these are separate contrasts, as propositional information is often not encoded in any public language. To this extent the standard contrast between simulation and ‘the theory theory’ is misleading. (The sinners here include my past self in Morton 1980 and Gopnik and Meltzoff 1997. For wisdom on the topic see the introduction to Davies and Stone 1995b. For an interpretation of ‘theory’ that takes it well away from verbal doctrine see Churchland 1998. The standard psychologist’s term ‘theory of mind’, as for example in Premack and Woodruff 1978, is similarly misleading, as it can mean either ‘systematic way of thinking about motives’ or ‘conceptualized beliefs about mind’. See also note 1 of Chapter 1.) After all, one relevant similarity that understanders can exploit is that both they and those they are understanding have learned some of the same theories.

3 This is just a price that has to be paid for using procedures that are not prohibitively expensive in time, or working memory. (One reason for the phenomenon is that an option which it is reasonable to reject in a large list of options may have advantages which only become apparent when the situation is simplified enough for them to become discoverable.) The loonie game, and others like it, also have another form of instability. Their equilibria are not always equilibria of subgames. This must be related to the instability described here, but I do not know how to make a precise connection.

4 As far as I can see this will only occur when the sub-situations that develop as the situation is played out are themselves concave. But this seems to be the case in the situations of interest.

5 We tend to ignore quite how much cooperation we pass up. Consider any one of your acquaintances, who possesses something they would happily exchange for something of yours, which you would happily exchange for their possession. You both know that. But neither spontaneously sends their thing to the other, knowing that each knows that there is a transaction from which both would gain. In fact, it would be mad to. The prisoner’s dilemmas, and the like, in which you do cooperate, though frequent, are a tiny subset of those that surround you and which you have the information to negotiate but ignore.

6 What about the propositional information that the other will choose whatever you will? I take this to be simulation shrunk down to a point. The information is that the other will choose *this*, where this is whatever you happen to be in the process of choosing.

7 A way in which the simulation here might be taken as cocognitive: you may know, for reasons of cultural similarity, that the other will take the situation as one presenting a moral problem, and there may be a familiar space of problems and solutions to the relevant kinds of problems (along the lines discussed in Chapter 2.) Then if you cocognize tackling the moral problem as presented you may confidently come up with the cooperative solution. You do have to be assured that the other will stick to it, though.

8 The development would be even less inevitable if, as may be the case, there is no best or normal form of conditional thinking. But supposing that conditional thinking is a basic ingredient of psychological thinking, we get another grasp of an issue from Chapter 2. There patterns of rational explanation often seemed to

be asking to be reformulated in a more general probabilistic form, though that form was not plausible as any part of commonsense psychology. The resistance of commonsense psychology to probabilistic rationality would be less mysterious if we tie it closely to conditional thinking, which is itself not easily given a probabilistic rationale. Thinking conditionally we look for the proposition p such that if q were true p would be true, while thinking probabilistically we look for the distribution over all propositions which q's truth would most strongly support.

Bibliography

Adkins, A. W. H. (1970) *From the Many to the One: A study of personality and views of human nature in the context of ancient Greek society, values and beliefs*, London: Constable

Arendt, Hannah (1964) *Eichmann in Jerusalem: A report on the banality of evil*, New York, NY: Viking

Arendt, Hannah (1994) *Essays in Understanding*, *1930–1954*, New York, NY: Harcourt, Brace and Co.

Axelrod, R. (1984) *The Evolution of Cooperation*, New York, NY: Basic Books

Bach, Kent (1998) 'Wide and narrow content', *Routledge Encyclopedia of Philosophy*

Baier, Annette (1985) 'Trust and anti-trust', *Ethics*, 96, 271–96

Barber, Alex (2000) 'A pragmatic treatment of simple sentences', *Analysis*, 60, 300–8

Barkow, Jerome H., Cosmides, Leda and Toby, John (1992) *The Adapted Mind: Evolutionary psychology and the development of culture*, Oxford: Oxford University Press

Baron-Cohen, Simon (1995) *Mindblindness*, Cambridge, MA: MIT Press

Bermúdez, José (1998) *The Paradox of Self-consciousness*, Cambridge, MA: MIT Press

Bermúdez, José (1999) 'Rationality and the backwards induction argument', *Analysis*, 59, 243–8

Bertram, Christopher (1994) 'Self-effacing Hobbesianism', *Proceedings of the Aristotelian Society*, 94, 19–34

Binmore, Ken (1998) *Game Theory and the Social Contract*, vol. II: *Just Playing*, Cambridge, MA: MIT Press

Boden, M. A. (ed.) (1990) *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press

Boghossian, Paul (1989) 'The rule-following considerations', *Mind*, 98, 507–49

Bond, E. J. (1983) *Reason and Value*, Cambridge: Cambridge University Press

Bratman, Michael (1987) *Intention, Plans, and Practical Reason*, Cambridge, MA: Harvard University Press

Bratman, Michael (1992) 'Shared cooperative activity', *Philosophical Review*, 102, 327–42

Braun, David (1998) 'Understanding belief reports', *Philosophical Review*, 107, 555–96

Brentano, Franz (1995/1874) *Descriptive Psychology*, London: Routledge

Broome, John and Rabinowicz, Vlodek (1999) 'Backwards induction in the centipede game', *Analysis*, 59, 237–42

Burge, Tyler (1979) 'Individualism and the mental', in Peter A. French, Theodore E. Uehling, Jr. and Howard K. Wettstein (eds) *Midwest Studies in Philosophy*, vol. 4, Minneapolis, MN: University of Minnesota Press, 73–122

Burge, Tyler (1986) 'Intellectual norms and foundations of mind', *Journal of Philosophy*, 83, 697–720

Byrne, Richard (1995) *The Thinking Ape*, Oxford: Oxford University Press

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

Carpenter, Malinda, Nagell, Katherine and Tomasello, Michael (1988) *Social Cognition and Communicative Competence from 8 to 15 Months of Age*, Chicago, IL: University of Chicago Press (Monographs for the Society for Research in Child Development, 63, no. 4)

Carruthers, Peter (1987) 'Russellian thoughts', *Mind*, 96

Carruthers, Peter (1988) 'More faith than hope: Russelian thoughts attacked', *Analysis*, 48, 91–6

Chater, N. and Oaksford, M. (1996) 'Deontic reasoning modules and innateness: a second look', *Mind and Language*, 11, 191–202

Cherniak, Christopher (1986) *Minimal Rationality*, Cambridge, MA: MIT Press

Child, William (1994) *Causality, Interpretation and the Mind*, Oxford: Oxford University Press

Churchland, Paul (1978) *Scientific Realism and the Plasticity of Mind*, Cambridge: Cambridge University Press

Churchland, Paul (1998) 'Folk psychology', in Paul Churchland and Patricia Churchland, *On the Contrary: Critical essays 1987–1997*, Cambridge, MA: MIT Press, pp. 17–24

Cohen, L. J. (1992) *An Essay on Belief and Acceptance*, Oxford: Oxford University Press

Cohon, Rachel (2000) 'The roots of reasons', *Philosophical Review*, 109, 63–86

Crimmins, Mark (1993) *Talk about Beliefs*, Cambridge, MA: MIT Press

Crutchfield, James P., Doyne Farmer, J., Packard, Norman H. and Shaw, Robert S. (1986) 'Chaos', *Scientific American*, 255, 46–57

Cummins, Denise (1996) 'Evidence for the innateness of deontic reasoning', *Mind and Language*, 11, 160–90

Currie, Gregory (1995) *Image and Mind*, Cambridge: Cambridge University Press

Currie, Gregory (1995b) 'Imagination and simulation: aesthetics meets cognitive science', in Martin Davies and Tony Stone (eds) (1995b) *Mental Simulation: Evaluations and applications*, Oxford: Blackwell

Dancy, Jonathan (2000) *Practical Reality*, Oxford: Oxford University Press

Daniels, Norman (1976) 'Wide reflective equilibrium and theory acceptance in ethics', *Journal of Philosophy*, 76, 256–82

Daniels, Norman (1996) *Justice and Justification: Reflective equilibrium in theory and practice*, New York, NY: Cambridge University Press

Danielson, Peter (1992) *Artificial Morality: Virtuous robots for virtual games*, London: Routledge

Davidson, Donald (1969) 'How is weakness of the will possible?', reprinted in *Essays on Actions and Events*, second edition, Oxford University Press, 2001, pp. 21–42

Davidson, Donald (1970) 'Mental events', reprinted in *Essays on Actions and Events*, second edition, Oxford University Press, 2001, pp. 207–24

Davidson, Donald (1976) 'Hempel on explaining actions', reprinted in *Essays on Actions and Events*, second edition, Oxford University Press, 2001, pp. 21–42

Davidson, Donald (1984) 'Belief and the basis of meaning', in *Inquiries into Truth and Interpretation*, Oxford: Oxford University Press, pp. 141–54

Davidson, Donald (2001) *Essays on Actions and Events*, second edition, Oxford: Oxford University Press

Davies, Martin and Stone, Tony (eds) (1995a) *Folk Psychology: The theory of mind debate*, Oxford: Blackwell

Davies, Martin and Stone, Tony (eds) (1995b) *Mental Simulation: Evaluations and applications*, Oxford: Blackwell

Deigh, John (1998) 'Empathy and universalizability', in *Mind and Morals: Essays in ethics and cognitive science*, Larry May, Marilyn Friedman and Andy Clark, (eds) Cambridge, MA: MIT Press, pp. 199–219.

Dennett, Daniel C. (1978) 'How to change your mind', Chapter 16 of *Brainstorms*, Hassocks: Harvester Press

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

Dennett, Daniel C. (1991) *Consciousness Explained*, Boston, MA: Little, Brown

de Sousa, Ronald (1971) 'How to give a piece of your mind, or the logic of belief and assent', *Review of Metaphysics*, 35, 52–79

de Sousa, Ronald (1987) *The Rationality of the Emotions*, Cambridge, MA: MIT Press

Dewdney, A. K. (1989) *The Turing Omnibus*, Computer Science Press

Dodds, E. R. (1951) *The Greeks and the Irrational*, Berkeley, CA: University of California Press

Dretske, F. (1988) *Explaining Behavior*, Cambridge, MA: MIT Press

Dunbar, Robin (1996) *Grooming, Gossip, and the Evolution of Language*, Cambridge, MA: Harvard University Press

Dunn, Judy (1988) *The Beginnings of Social Understanding*, Oxford: Blackwell

Edgington, Dorothy (1995) 'On conditionals', *Mind*, 104, 235–330

Evans, Gareth (1983) *The Varieties of Reference*, Oxford: Oxford University Press

Fane, Carlos (1995) 'Wittgenstein's rule-following considerations, an argument against naturalistic reductionism in semantics', PhD dissertation, University of Bristol

Fodor, J. A. (1990) *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press

Freeman, N. and Lacohée, Hazel (1995) 'Making explicit 3 year olds implicit competence with their own false beliefs', *Cognition*, 56, 31–60

Garey, R. and Johnson, D. S. (1979) *Computers and Intractability*, San Francisco, CA: Freeman

Garfinkel, Alan (1980) *Forms of Explanation*, New Haven, CN: Yale University Press

Gauker, Christopher (1994) *Thinking Out Loud*, Princeton, NJ: Princeton University Press

Gauker, Christopher (2002) *Words without Meaning*, Cambridge, MA: MIT Press

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

Gergely, Gyorgy, Nadasdy, Zolty, Gergely, Csibra and Biro, Szilvia (1995) 'Taking the intentional stance at 12 months of age', *Cognition*, 56, 165–93

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

Gigerenzer, Gert (1997) 'The modularity of social intelligence', in Andrew Whiten and Richard W. Byrne (eds) *Machiavellian Intelligence II*, Cambridge: Cambridge University Press, pp. 264–88

Gigerenzer, Gert and Golstein, Daniel (1996) 'Reasoning the fast and frugal way: models of bounded rationality', *Psychological Review*, 103, 4, 650–69

Gigerenzer, Gert, Todd, Peter and the ABC research group (2000) *Simple Heuristics that Make Us Smart*, Oxford: Oxford University Press

Gleick, J. (1988) *Chaos*, London: Heinemann

Godfrey-Smith, P. (1994) 'A continuum of semantic optimism', in S. Stich and T. A. Warfield (eds) *Mental Representation: A reader*, Oxford: Blackwell

Goldie, Peter (1999) 'How we think of others' emotions', *Mind and Language*, 14, 394–423

Goldie, Peter (2000) *The Emotions: A philosophical exploration*, Oxford: Oxford University Press

Goldman, Alvin (1992a) 'In defence of the simulation theory', *Mind and Language*, 7, 104–19

Goldman, Alvin (1992b) 'Empathy mind and morals', *Proceedings and Addresses of the APA*, 66, 17–41

Goldman, Alvin (2001) 'Desire, intention, and the simulation theory' in Bertram Malle, Louis Moses and Dare Baldwin (eds) *Intentions and Intentionality*, Cambridge, MA: MIT Press, pp. 207–24

Goodman, Nelson (1973) *Fact Fiction and Forecast*, third edition, Indianapolis, IN: Bobbs-Merrill

Gopnik, Alison (1993) 'How we know our minds: the illusion of first person knowledge of intentionality', *Behavioral and Brain Sciences*, 16, 1–14

Gopnik, Alison and Meltzoff, Andrew (1997) *Words Thoughts and Theories*, Cambridge, MA: MIT Press

Gordon, Robert (1992) 'The simulation and the theory theory', *Mind and Language*, 7

Gordon, Robert (1995a) 'The simulation theory: objectives and misconceptions', in Martin Davies and Tony Stone (eds) (1995a) *Folk Psychology: The theory of mind debate*, Oxford: Blackwell

Gordon, Robert (1995b) 'Simulation without introspection or inference from me to you', in Martin Davies and Tony Stone (eds) (1995b) *Mental Simulation: Evaluations and applications*, Oxford: Blackwell

Greenspan, Patricia (1988) *Emotions and Reasons*, London: Routledge

Grice, G. R. (1967) *The Grounds of Moral Judgement*, Cambridge: Cambridge University Press

Hacking, Ian (1995) *Rewriting the Soul: Multiple personality and sciences of memory*, Princeton, NJ: Princeton University Press

Hammond, Peter (1976) 'Changing tastes and coherent dynamic choice', *Review of Economic Studies*, 43, 159–73

Hardin, Russell (1995) *One for All: The logic of group conflict*, Princeton, NJ: Princeton University Press

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

Harman, Gilbert (1999) *Reasoning, Meaning, and Mind*, Oxford: Oxford University Press

Harman, Gilbert (1999a) 'Moral philosophy meets social psychology: virtue ethics and the fundamental attribution error', *Proceedings of the Aristotelian Society*, 99, 315–31

Harris, John and Kamm, Frances (2000) 'The doctrine of triple effect' (symposium), *Proceedings of the Aristotelian Society*, supplementary volume 74, 21–57

Harris, Paul (1989) *Children and Emotion*, Oxford: Blackwell

Harris, Paul (1996) 'Desires, beliefs, and language', in Peter Carruthers and Peter Smith (eds) *Theories of Theories of Mind*, Cambridge: Cambridge University Press, pp. 200–22

Heal, Jane (1978) 'Common knowledge', *Philosophical Quarterly*, 28, 116–31

Heal, Jane (1995) 'How to think about thinking', in Martin Davies and Tony Stone (eds) (1995b) *Mental Simulation: Evaluations and applications*, Oxford: Blackwell

Heal, Jane (1998) 'Cocognition and off-line simulation', *Mind and Language*, 13, 348–64

Heal, Jane (2000) 'Other minds, rationality, and analogy', *Proceedings of the Aristotelian Society*, supplementary volume 74, 1–19

Hitchcock, Christopher (1996) 'The role of contrast in causal and explanatory claims', *Synthese*, 107

Hollis, Martin (1991) 'Penny pinching and backward induction', *Journal of Philosophy*, 88, 473–88

Holton, Richard (1994) 'Deciding to trust, coming to believe', *Australasian Journal of Philosophy*, 72, 1, 63–76

Hookway, Christopher (2000) 'Epistemic norms and theoretical deliberation', in Jonathan Dancy (ed.) *Normativity*, Oxford: Blackwell, pp. 60-77

Horowitz, Tamara (1998) 'Philosophical intuition and psychological theory', in R. Michael DePaul and William Ramsey (eds) *Rethinking Intuition*, Lanham, MD: Rowman and Littlefield, pp. 142–60

Humphrey, Nicholas (1984) *Consciousness Regained*, Oxford: Oxford University Press

Hurley, Susan and Bacharach, Michael (1991) *Foundations of Decision Theory*, Oxford: Blackwell

Jackson, Frank and Pargetter, Robert (1986) 'Oughts, options, and actualism', *Philosophical Review*, 95, 233–55

Jackson, Frank and Pettit, Philip (1990) 'In defence of folk psychology', *Philosophical Studies*, 31–53

Jeffrey, Richard (1983) *The Logic of Decision*, Chicago, IL: Chicago University Press

Johnson, Christine and Keil, Frank (2000) 'Explanatory knowledge and conceptual combination', in Frank Keil and Robert Wilson (eds) *Explanation and Cognition*, Cambridge, MA: MIT Press, pp. 327–60

Jones, Karen (1999) 'Second-hand moral knowledge', *Journal of Philosophy*, 96, 51–78

Joyce, James (1999) *The Foundations of Causal Decision Theory*, Cambridge: Cambridge University Press

Kahneman, D. and Tversky, A. 1979: 'Prospect theory: an analysis of decision under uncertainty', *Econometrica*, 47, 263–91

Kamm, F. M. (1994) *Morality, Mortality*, vol. I, Oxford: Oxford University Press

Kamm, F. M. (1996) *Morality, Mortality*, vol. II: *Rights, Duties, and Status*, Oxford: Oxford University Press

Kamm, Frances and Harris, John (2000) 'The doctrine of triple effect', *Proceedings of the Aristotelian Society*, supplementary volume 74, 21–58

Kaplan, David (1989) 'Demonstratives', in J. Almog, J. Perry and H. Wettstein (eds) *Themes from Kaplan*, New York, NY: Oxford University Press, pp. 481–614

Keil, Frank (1989) *Concepts, Kinds, and Cognitive Development, 1987: Fact and meaning*, Cambridge, MA: MIT Press

Kinderman, P., Dunbar, R. and Bentall, R. (1998) 'Theory of mind deficits and causal attributions', *British Journal of Psychology*, 89, 191–204

Kornblith, Hilary (1998) 'Epistemology of introspection', in Edward Craig (ed.) *The Routledge Encyclopedia of Philosophy*

Korsgaard, Christine (1996) *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press

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

Kreps, David M., Milgram, P., Roberts, J. and Wilson, R. (1982) 'Rational cooperation and the finitely repeated prisoners' dilemma', *Journal of Economic Theory*, 27, 245–59

Kripke, Saul (1979) 'A puzzle about belief', in A. Margalit (ed.) *Meaning and Use*, Dordrecht: Reidel, pp. 239–83

Kripke, S. A. (1982) *Wittgenstein on Rules and Private Language*, Oxford: Blackwell

Kusch, Martin (1999) *Psychological Knowledge: A social history and philosophy*, London: Routledge

Lennon, Kathleen (1990) *Explaining Human Action*, London: Duckworth

Lewis, David (1968) *Convention*, Cambridge, MA: Harvard University Press

Lewis, David (1973) *Counterfactuals*, Oxford: Blackwell

Lewis, David (1986) 'Causation', *Philosophical Papers*, vol. II, Oxford University Press

Lewis, David (2000) 'Causation as influence', *Journal of Philosophy*, 97, 182–97

Lipton, Peter (1991) *The Inference to the Best Explanation*, London: Routledge

Little, Margaret (1994) 'Recent work on moral realism', *Philosophical Books*, 35, 145–52, 225–32

Longino, Helen E. (1999) 'Feminist epistemology', in John Greco and Ernest Sosa (eds) *The Blackwell Guide to Epistemology*, Oxford: Blackwell, pp. 327–53

Luce, R. Duncan and Raiffa, Howard (1957) *Games and Decisions*, New York, NY: Wiley

May, Larry, Friedman, Marilyn and Clark, Andy (eds) (1996) *Mind and Morals: Essays on cognitive science and ethics*, Cambridge, MA: MIT Press

McCulloch, Gregory (1988) 'Faith, hope, and charity: Russellian thoughts defended', *Analysis* 48, 85–90

McDermott, Michael (1995) 'Redundant causation', *Journal of Philosophy*, 97, 235–56

McDowell, J. (1984) 'Wittgenstein on following a rule', *Synthèse* 58: 325–63

McGeer, Victoria (2001) 'Psycho-practice, psycho-theory, and the contrastive case of autism', *Journal of Consciousness Studies*, 8, 109 –32

McGeer, Victoria (2002) 'Developing trust', *Philosophical Explorations*, 5, 21–38

McGinn, Colin (1998) *Ethics, Evil, and Fiction*, Oxford: Oxford University Press

McKelvey, Richard D. and McLennan, Andrew (1996) 'Computation of equilibria in finite games', *Handbook of Computational Economics*, vol. I, Amsterdam: Elsevier, 87–142

McKelvey, Richard D. and McLennan, Andrew (1997) 'The maximal number of regular totally mixed Nash equilibria', *Journal of Economic Theory*, 72, 411–25

McLennan, Andrew (1997) 'On the expected number of equilibria of a normal form game', *Journal of Economic Literature*

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

Miller, Richard W. (1985) 'Ways of moral learning', *The Philosophical Review*, 44, 507–56

Millgram, Elijah (1997) *Practical Induction*, Cambridge, MA: Harvard University Press

Millikan, R. G. (1984) *Language, Thought, and Other Biological Categories*, Cambridge, MA: MIT Press

Moody-Adams, Michele (1999) 'The idea of moral progress', *Metaphilosophy*, 30, 169–85

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

Morton, Adam (1989) 'The inevitability of folk psychology', in R. Bogdan (ed.) *Mind and Common Sense*, Cambridge: Cambridge University Press, pp. 93–122

Morton, Adam (1991) *Disasters and Dilemmas: Strategies for real-life decision-making*, Oxford: Blackwell

Morton, Adam (1993) 'Heuristics and counterfactual self-knowledge', *Behavioral and Brain Sciences*, 16, 1, 63–4

Morton, Adam (1994) 'Game theory and knowledge by simulation', *Ratio*, 7

Morton, Adam (1994a) review of Crimmins' 'Talk about beliefs', *Philosophical Books*, 35, 1, 47–9

Morton, Adam (1996a) 'Folk psychology is not a predictive device', *Mind*, 105, 1–19

Morton, Adam (1996b) 'The disunity of the moral', in Jan Bransen and Marc Slors (eds) *The Problematic Reality of Values*, Dordrecht: Van Gorcum, pp. 142–55

Morton, Adam (1998) review of Colin McGinn, *Ethics, Evil and Fiction*, *Times Literary Supplement*, 30 January, 28–9

Morton, Adam (2000) 'The evolution of strategic thinking', in Peter Carruthers and Andrew Chamberlain (eds) *Evolution and the Human Mind: Language, modularity and meta-cognition*, Cambridge: Cambridge University Press, pp. 218–37

Morton, Adam (2001a) review of Binmore, *Game Theory and the Social Contract*, vol. 2*: Just Playing*, *Mind*, 110, 168–71

Morton, Adam (2001b) 'Psychology for cooperators', in Christopher Morris (ed.) *Practical Rationality and Preferences: Essays for David Gauthier*, Cambridge: Cambridge University Press, pp. 153–72

Morton, Adam (2002) *A Guide through the Theory of Knowledge*, third edition, Oxford: Blackwell

Mulhall, Stephen (1990) *On Being in the World: Wittgenstein and Heidegger on seeing aspects*, London: Routledge

Myerson, Roger B. (1991) *Game Theory*, Cambridge, MA: Harvard University Press

Nagel, Thomas (1979) *Mortal Questions*, Cambridge: Cambridge University Press

Nagel, Thomas (1986) *The View from Nowhere*, Oxford: Oxford University Press

Neander, K. (1995) 'Misrepresenting and Malfunctioning', *Philosophical Studies*, 79: 109–41

Nisbett, Richard E. and Ross, Lee (1980) *Human Inference: Strategies and shortcomings of social judgement*, New York, NY: Prentice Hall

Nisbett, Richard E. and Ross, Lee (1991) *The Person and the Situation*, New York, NY: McGraw-Hill

Nisbett, Richard E. and Wilson, T. D. (1977) 'Telling more than we can know', *Psychological Review*, 84, 231–59

Papineau, D. (1987) *Reality and Representation*, Oxford: Blackwell

Paul, L. A. (2000) 'Aspect causation', *Journal of Philosophy*, 97, 235–56

Peacocke, Christopher (1981) 'Demonstrative thought and psychological explanation', *Synthese*, 49, 187–217

Pears, David (1984) *Motivated Irrationality*, Oxford: Oxford University Press

Perner, Josef (1991) *Understanding the Representational Mind*, Cambridge, MA: MIT Press

Perner, Josef (1995) 'The many faces of belief: reflections on Fodor's and the child's theory of mind', *Cognition*, 57, 240–69

Perner, Josef (1996) 'Simulation as excitation of predication-implicit knowledge about the mind: arguments for a simulation-theory mix', in Peter Carruthers and Peter Smith (eds) *Theories of Theories of Mind*, Cambridge: Cambridge University Press, pp. 90–104

Perner, Josef and Howes, Deborah (1995) 'He thinks he knows', in Martin Davies and Tony Stone (eds) (1995a) *Folk Psychology: The theory of mind debate*, Oxford: Blackwell

Perner, J., Gschaider, A., Kühberger, A. and Schrofner, S. (1999) 'Predicting others through simulation or by theory? A method to decide', *Mind and Language*, 14, 57–79

Perry, John (1977) 'Frege on demonstratives', *Philosophical Review*, 86, 474–97

Pettit, Philip (1986) 'Free riding and foul dealing', *Journal of Philosophy*, 83, 7, 361–80

Pettit, Philip (1991) 'Folk psychology and game theory', in Susan Hurley and Michael Bacharach (eds) *Foundations of Decision Theory*, Oxford: Blackwell

Pettit, Philip and Sugden, Robert (1989) 'The backward induction paradox', *Journal of Philosophy*, 86, 169–82

Pietroski, P. (1992) 'Intentionality and teleological error', *Pacific Philosophical Quarterly*, 73, 267–82

Platts, Mark (1991) *Moral Realism*, London: Routledge

Premack, D. and Woodruff, G. (1978) 'Does the chimpanzee have a theory of mind?', *Behavioral and Brain Sciences*, 4, 515–26

Quine, W. V. (1960) *Word and Object*, Cambridge, MA: Harvard University Press

Rabinowicz, Wlodek (1995) 'To have one's cake and eat it too', *Journal of Philosophy*, 92, 1–32

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, pp. 107–26

Railton, Peter (1986) 'Moral realism', *Philosophical Review*, 95, 163–207

Rakowski, Eric (1998) review of Kamm *Morality, Mortality*, vol. II, *Mind*, 107, 492–8

Ramsey, F. P. (1931) *The Foundations of Mathematics*, London: Routledge

Rawls, John (1971) *A Theory of Justice*, Cambridge, MA: Harvard University Press

Raz, Joseph (2000) 'Explaining normativity: on rationality and the justification of reason', in Jonathan Dancy (ed.) *Normativity*, Oxford: Blackwell, pp. 34–59

Richard, Mark (1990) *Propositional Attitudes: An essay on thoughts and how we ascribe them*, Cambridge: Cambridge University Press

Richardson, Henry (1994) *Practical Reasoning about Final Ends*, Cambridge: Cambridge University Press

Rorty, Richard (1991) 'Solidarity or objectivity', in *Philosophical Papers*, vol. I: *Objectivity, Relativism and Truth*, Cambridge: Cambridge University Press

Ross, Lee and Nisbett, Richard E. (1991) *The Person and the Situation*, New York, NY: McGraw-Hill

Rubinstein, Ariel (1998) *Modelling Bounded Rationality*, Cambridge, MA: MIT Press

Rudder Baker, Lynne (1982) 'De re belief in action', *Philosophical Review*, 91, 363–88

Rueger, Alexander and Sharp, W. David (1996) 'Simple theories of a messy world: truth and explanatory power in nonlinear dynamics', *British Journal for the Philosophy of Science*, 47, 1, 93–112

Russell, J. (1997) *Autism as an Executive Disorder*, Oxford: Oxford University Press

Salmon, Nathan (1986) *Frege's Puzzle*, Cambridge, MA: MIT Press

Saul, Jennie (1997) 'Substitution and simple sentences', *Analysis*, 57, 102–8

Scanlon, Thomas (1972) 'A theory of freedom of expression', *Philosophy and Public Affairs*, 1, 204–20

Scanlon, Thomas (1998) *What We Owe to Each Other*, Cambridge, MA: Harvard University Press

Scheffler, Samuel (1994) *The Rejection of Consequentialism*, revised edition, Oxford: Oxford University Press

Schelling, Thomas C. (1960) *The Strategy of Conflict*, Cambridge, MA: Harvard University Press

Schelling, Thomas C. (1978) *Micromotives and Macrobehavior*, New York, NY: Norton

Schick, Frederic (1997) *Making Choices: Recasting decision theory*, Cambridge: Cambridge University Press

Schiffer, Stephen (1972) *Meaning*, Oxford: Oxford University Press

Schmitt, Frederick F. (1998) 'Social epistemology', *Routledge Encyclopedia of Philosophy*, London: Routledge

Schmitt, Alain and Grammer, Karl (1997) 'Social intelligence and success: don't be too clever in order to be smart', in Andrew Whiten and Richard W. Byrne (eds) *Machiavellian Intelligence II*, Cambridge: Cambridge University Press, pp. 86–111

Schuster, Heinz Georg (1984) *Deterministic Chaos*, Weinheim: Physik-Verlag

Sellars, Wilfrid (1963) *Science, Perception, and Reality*, New York, NY: Humanities Press

Sereny, Gitta (1995) *Albert Speer: His battle with truth*, New York, NY: Knopf

Sereny, Gitta (2000) *The German Trauma: Experiences and reflections 1938–2000*, London: Penguin

Shapin, Steven (1994) *A Social History of Truth*, Cambridge: Cambridge University Press

Shepard, Roger N. (1982) *Mental Images and Their Transformations*, Cambridge, MA: MIT Press

Simon, H. A. (1982) *Models of Bounded Rationality*, vol. 2, Cambridge, MA: MIT Press

Skarda, Christine and Freeman, Walter (1987) 'How brains make chaos in order to make sense of the world', *Behavioral and Brain Sciences*, 10, 161–95

Skinner, B. F. (1971) *Beyond Freedom and Dignity*, New York, NY: Knopf

Skyrms, Brian (1992) 'Equilibrium and the dynamics of rational deliberation', in

Christina Biccieri and Maria Luisa Dalla Chiara (eds) *Knowledge, Belief, and Strategic Interaction*, Cambridge: Cambridge University Press, pp. 93–106

Slote, Michael A. (1989) *Beyond Optimizing: A study of rational choice*, Cambridge, MA: Harvard University Press

Smith, Michael (1994) *The Moral Problem*, Oxford: Blackwell

Sperber, Dan (1996) *Explaining Culture: A naturalistic approach*, Oxford: Blackwell

Sperber, Dan and Wilson, Deirdre (1986) *Relevance*, Oxford: Blackwell

Sreenivasam, Gopal (2002) 'Errors about errors: virtue theory and trait attribution', *Mind*, 111, 47–68

Stalnaker, Robert (1994) 'On the evolution of solution concepts', *Theory and Decision*, 37, 1, 49–76

Stein, Edward (1996) *Without Good Reason*, Oxford: Oxford University Press

Sterelney, Kim (1995) 'Understanding life: Recent work in philosophy of biology', *British Journal for the Philosophy of Science*, 46, 155–83

Stevenson, Leslie (2002) 'Six levels of mentality', *Philosophical Explorations*, 5, 120–32

Stich, Stephen P. (1983) *From Folk Psychology to Cognitive Science*, Cambridge, MA: MIT Press

Stich, Stephen P. (1996) *Deconstructing the Mind*, Oxford: Oxford University Press

Stich, Stephen P. and Ravenscroft, Ian (1996) 'What *is* folk psychology?', in Stephen P. Stich (ed.) *Deconstructing the Mind*, Oxford: Oxford University Press

Stocker, Michael (1989) *Plural and Conflicting Values*, Oxford: Oxford University Press

Svavarsdottir, Sigrun (1999) 'Moral cognitivism and motivation', *Philosophical Review*, 108, 161–220

Taylor, M. (1987) *The Possibility of Cooperation*, Cambridge: Cambridge University Press

Thomson, Judith Jarvis (1976) 'Killing, letting die, and the trolley problem', *Monist*, 59, 204–17

Trianoski, Gregory (1987) 'Virtue, action, and the good life: towards a theory of the virtues', *Pacific Philosophical Quarterly*, 68, 124–47

Turiel, Eliot (1983) *The Development of Social Knowledge*, Cambridge: Cambridge University Press

Tversky, A. and Kahneman, D. (1992) 'Advances in prospect theory: cumulative representation of uncertainty', *Journal of Risk and Uncertainty*, 5, 297–323

Tye, Michael (1995) *The Imagery Debate*, Cambridge, MA: MIT Press

van Fraassen, Bas (1980) *The Scientific Image*, Oxford: Oxford University Press

Vayne, Paul (1988) *Did the Greeks Believe in Their Myths?* Chicago, IL: University of Chicago Press

Velleman, David (2000) 'How to share an intention', in *The Possibility of Practical Reason*, Oxford: Oxford University Press, 200–20

von Neumann, John and Morgenstern, Oskar (1944) *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press

Warren, Dôna D. (1999) 'Externalism and causality: simulation and the prospects for a reconciliation', *Mind and Language*, 14, 154–76

Wellman, Henry (1990) *The Child's Theory of Mind*, Cambridge, MA: MIT Press

Whiten, Andrew and Byrne, Richard W. (eds) (1997) *Machiavellian Intelligence II*, Cambridge: Cambridge University Press

Wiggins, David (1998) 'Universalizability, impartiality, truth', in *Needs, Values, Truth*, third edition, Oxford: Oxford University Press, pp. 59–86

Williams, Bernard (1993) *Shame and Necessity*, Berkeley, CA: University of California Press

Williams, Bernard (1995) 'Formal structures and social reality', in *Making Sense of Humanity*, Cambridge: Cambridge University Press, pp. 111–21

Williamson, Timothy (1995) 'Is knowing a state of mind? *Mind*, 104, 533–65

Williamson, Timothy (2000) *Knowledge and Its Limits*, Oxford: Oxford University Press

Wilson, Mark (1982) 'Predicate meets property', *Philosophical Review*, 9, 4, 549–90

Wittgenstein, Ludwig (1953) *Philosophical Investigations*, Oxford: Blackwell

Wolff, Jonathan (1998) 'Fairness, respect, and the egalitarian ethos', *Philosophy and Public Affairs*, 27, 97–122

Wollheim, Richard (1984) *The Thread of Life*, Cambridge: Cambridge University Press

Wright, C. (1989) 'Wittgenstein's rule-following considerations and the central project of theoretical linguistics', in A. George (ed.) *Reflections on Chomsky*, Oxford: Blackwell

Zagzebski, Linda (1996) *Virtues of the Mind*, Cambridge: Cambridge University Press

Zalta, Edward (1988) *Intensional Logic and the Metaphysics of Intentionality*, Cambridge, MA: MIT Press

Index