JUSTICE, RESPONSIBILITY, AND ACQUIESCENCE

Christopher Woodard

Submitted for the degree of PhD

University of Warwick

Department of Politics and International Studies

July 1997

For Annabel

CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Susan Hurley, for her extensive and penetrating comments on drafts of this thesis, and preliminary work, over the last four years. I have tried to respond to her critical comments, though I fear I have not always succeeded.

I should also like to thank Andrew Reeve, Andrew Williams, Greg Hunt, Michael Luntley, Steven Lukes, John Halliday, Ed Page, my parents and my brothers for help and encouragement.

Most of all I thank Annabel Kiernan, for cheering me up, putting up with me whilst I've been preoccupied, and otherwise helping me along.

All mistakes are mine.

DECLARATION

Some of the arguments of chapters Two, Three and Four have appeared in a different form, in "Responsibility, Desert, and Liberal Theories of Justice", in J. Stanyer and G. Stoker (eds.), *Contemporary Political Studies 1997*, volume two (Nottingham: Political Studies Association, 1997), pp. 1161-1168.

SUMMARY

This thesis investigates the relationship between the concepts of *justice* and *responsibility*. It is important to decide what the relationship is, because the details of a theory of justice will depend on it. Four possible views of the relationship are outlined, and arguments are canvassed for and against one of them, which I call *naturalism*.

Naturalism is appealing because it offers to make theories of justice independent of troubling agency-implicating judgements. But I argue that naturalism is false, because political argument, including theories of justice, cannot do without such judgements. They play an essential role in determining which range of possible actions or arrangements is relevant to a political argument.

The argument against naturalism is in two parts. The first part analyses the concept of *benefit*, underlining the feature of that concept which makes agency-implicating judgements necessary for those who employ it. This first anti-naturalist argument is directed to arguments in *ideal theory*, in Rawls's sense of that term.

The second part of the argument against naturalism is directed to *deliberative arguments*. Naturalism is, I claim, a much more plausible doctrine if it is understood to apply to such arguments in particular. But I argue that it is nevertheless false, because it leaves us unable to account for some of the reasons persons have for resisting *acquiescence*.

Discussion of the rationality of acquiescence leads into discussion of the nature of deliberation. I argue that a feature of some consequentialist models of deliberation, which I call the *hard-nosed view*, must be rejected. I end with a comparison of the resulting view with Kant's ethics, and some variant forms of consequentialism.

Chapter One

Introduction

Political arguments often turn on judgements about what is possible. You think that we could introduce a statutory minimum wage *and* reduce unemployment, but I disagree; she thinks that communism is possible, whereas he does not. Such judgements do not arise only in large-scale ideological confrontations. They are involved in more mundane arguments too. We might disagree about whether the local council could keep the library open, if it had the will to do so. In fact, disagreements in political matters are as often about what is possible, as they are about what is desirable. But this is not reflected in academic discussion of political argument.[1]

This fact explains why a certain view of the conceptual relationship between justice and responsibility has gained currency. Theories of justice differ according to the role they give the concept of responsibility. Most theories find some relationship between justice and responsibility, but there is room for a great deal of disagreement

[1] Academic discussion of normative political argument is usually oriented to disagreement about values or principles, and so the range of possibilities in cases under discussion is usually simply stipulated: 'Suppose an actor is faced with three options . . .'. There is nothing wrong with such stipulation. But the lack of discussion of judgements of possibility is striking, once it is noticed. One partial exception to the norm is T. Nagel, *Equality and Partiality* (New York and Oxford: Oxford University Press, 1991).

about the nature of the relationship.[2] Common-sense depicts justice as conceptually dependent on responsibility – so that, for example, whether someone earned their fortune or inherited it, is important for determining whether it is just that they should keep it. Judgements of this kind, about the aspects of their circumstances persons are responsible for, play an important role in thoughts about justice, according to the common-sense view.

But the common-sense view has not gone unchallenged. Some recent discussion of justice has sought to avoid making justice dependent on responsibility, and for good reasons. For judgements of responsibility are notoriously controversial and subject to deep philosophical problems. If we could describe what justice requires without invoking any judgements of responsibility, we could avoid a fertile source of disagreement.[3] And, according to one version of this strain of thought, we might even be able to exploit the conceptual connection between justice and responsibility in an unusual way, to explain the latter in terms of the former. Where common-sense finds justice to be dependent on responsibility, and so uses judgements of responsibility to explain what justice requires, this alternative view finds responsibility to be dependent on justice, and seeks to explain our judgements of responsibility *in terms of* our understanding of justice.

I call this revisionist idea, *naturalism* about responsibility. It is naturalist (in a sense of that word) because its chief attraction is that it offers to provide a way to finesse problems of agency. A naturalist reconstruction of judgements of responsibility says something like this: "'Smith is responsible for what she does'

---

[2] I shall explain the various possible views in Chapter Two, section 1.

[3] It is probably no accident that the leading exponents of this view of the relationship between justice and responsibility, Scanlon and Rawls, are contractualists. I shall discuss their views in Chapter Two and Chapter Three.

means 'it is fair or just to let Smith bear the costs or enjoy the benefits of what she does'".

Since common-sense thinks that, when certain other conditions are satisfied, the second expression is an *implication* of the first, it is committed to explicating the concept of *responsibility* in such a way that it has explanatory power. In practice, this lands the common-sense view with all of the controversy and philosophical problems of concepts of agency. But since naturalism claims that the second expression is *synonymous* with the first, it need not endow the concept of responsibility with any explanatory power. On this view, responsibility is, in effect, simply a derivative *moral* concept, parasitic on our understanding of fairness or justice, and need not involve us in problems of agency. Usually the fairness of letting Smith bear the costs will depend on what Smith has done; but the explanatory emphasis is on the concept of fairness, or the concept of justice, rather than on the concept of agency.

This dissertation is about naturalism, and the view about political possibility which underlies it. Chapters Two and Three examine naturalism and other possible views about the relationship between justice and responsibility. It is important to decide whether naturalism or the common-sense view is correct, but it is difficult to find conclusive arguments for either view. In Chapter Two I consider the suggestion that the common-sense view alone is able to account adequately for our intuitions about justice, but conclude that our intuitions are ambiguous in their message. Chapter Three tries a different approach, examining Rawls's remarks on desert and responsibility, which are arguably naturalist, to see whether they provide evidence for or against naturalism. The examination alerts us to the different ways in which a

naturalist theory may be constructed, but it does not tell decisively for or against naturalism.

I find naturalism attractive in certain respects, but I think it is false. It is false because we cannot do without agency-implicating concepts and judgements – and the reason for that is that such judgements are essential in determining the range of possibilities relevant to a political argument. I make this argument in Chapter Four, drawing on G. A. Cohen's remarks on incentives.[4]

Cohen argues that a common putative justification of inequalities fails, because it doesn't take account of a certain desirable form of community. I agree with Cohen's conclusion, but I diagnose the problem somewhat differently. Like many other arguments, Cohen's target relies on the concept of *benefit*. Now, since the concept of benefit is inherently comparative, asking us to compare one arrangement with some others, arguments which rely on it implicitly invoke some characterisation of the range of relevant arrangements. I claim that the argument Cohen criticises fails, ultimately, because it doesn't consider the right range of alternatives. It is too narrow, arguing the merits of systems of inequality against an unduly restricted range of alternatives.

What, then, should determine the appropriate range? The answer, I think, is judgements about *what persons could do*. For example, whether or not a strongly egalitarian regime is a relevant possibility – against which other arrangements should be compared – depends on what persons could do. The claim that *Smith could do such-and-such* is an *agency-implicating judgement*, and we can't do without

[4] See Chapter Four for citations of Cohen's relevant works.

judgements like that, so naturalism is false because it tells us that we can do without them.

This is in some ways a discouraging conclusion, since it suggests that political philosophy is at the mercy of developments in our understanding of agency, whilst problems of agency remain apparently intractable. Traditionally, one large class of such problems has centred around the attempt to describe the special causal character of actions.[5] One of the things agents do is cause events. It has often been thought that the best way to shed light on the nature of agency would be to isolate the special causal character of this process. Advocates of this approach begin by saying: it must be something about the way agents cause events that makes them agents. So they go on to ask what the peculiarity is.

Naturalists do not deny that agents cause events, they simply deny that we should look for an explanation of agency in terms of the special causal character of actions. Instead, they say, we should look at the *institutional context* of action. Being an agent, on this view, is still a matter of causing events, but what distinguishes agents from other objects with causal powers is the moral and institutional background of action, not the way in which events are caused.[6] Hence we can finesse metaphysical problems of agency by explaining the institutional conditions which make it fair to let persons bear the costs of their actions. But this naturalist strategy cannot work. It

---

[5] I do not claim that no other aspects of agency are problematic. Nagel, for example, thinks that a large part of the problem lies in the irreconcilability of agent-centred and objective views. T. Nagel, *The View From Nowhere* (New York and Oxford: Oxford University Press, 1986), Chapter VII. But I think that what I've called the problem of specifying the special causal character of agency has been, traditionally, a fundamental preoccupation.

[6] See Scanlon's discussion of the significance of choice. His answer to the question, 'what makes choice morally significant?', is not: 'the fact that it has a special kind of cause', but: 'the fact that it is made in favourable conditions'. T. Scanlon, "The Significance of Choice", in S. Darwall (ed.), *Equal Freedom. Selected Tanner Lectures on Human Values* (Ann Arbor: University of Michigan Press, 1995).

cannot work because we need to rely on agency-implicating judgements in order to say what fairness or justice requires.

The argument to this point is relatively straightforward. Naturalism goes wrong by overlooking the essential role which agency-implicating judgements have in determining the range of relevant possibilities. We need such judgements to discriminate between relevant and irrelevant possibilities. Once we reflect on the issue about the range of relevant possibilities, it is fairly clear that we need to be able to discriminate between things persons could and could not do, in an agency-implicating sense.[7] But the argument gets more complicated when we try to specify the appropriate grounds for making judgements about which possibilities are relevant.

At this point, we have to make a distinction between two types or modes of normative political argument. *Ideal theory* evaluates arrangements according to very generous assumptions about the range of relevant possibilities. In order for a possibility to be irrelevant to ideal theory, there must be some very general, persistent fact about the world which it is inconsistent with. The fact that persons have certain motivations currently is not sufficient to make irrelevant an arrangement which presupposes different motivations, for example. Only deep and persistent facts about humans and their circumstances – which mark the boundaries of what we could do, or how we could live – suffice in ideal theory to make a possibility irrelevant. Demanding criteria of irrelevance are appropriate to thought about ideals.

In *deliberation*, however, we seek to guide the actions of some agent. To that end, the criteria of *relevance* (not irrelevance) are demanding. For something to be a relevant possibility, it is not sufficient that it is possible for humans in a very broad

[7] I explain what I mean by 'agency-implicating' in Chapter Two, section 1.

sense; it must be possible given the circumstances at hand. There is much more information for deliberative argument legitimately to rule-out possibilities, information about the actor's circumstances. The conclusions of ideal theory, in contrast to the conclusions of deliberation, must apply generally, not specifically to a certain actor at a certain time.

The anti-naturalist argument I sketched earlier assumes that we are interested in ideal theory, not deliberation. It is quite clear, I think, that in ideal theory we must rely on agency-implicating judgements about what persons could do. But naturalism is much harder to refute if it is considered as a doctrine about deliberation. For it is not so clear that we must rely on judgements about what persons could do, in an agency-implicating sense, in deliberation. We must, it is true, suppose that the actor in question could do a range of things – but with respect to other persons (third parties), it seems that we should simply rely on our best predictions of their behaviour, which need not be agency-implicating, and are quite consistent with naturalism.

Discussion of this issue is helped by the use of some terminology. In Chapter Four I suggest a simple model of deliberative argument. In terms of this model, the crucial issue is how we evaluate the *expected effectiveness* of the actor's options. It seems plausible that we should do so on the basis of our best predictions about the future behaviour of the actor's environment, including other persons, conditional on the actor's behaviour. Rejection of naturalism about deliberation (via this route of discussion of issues of possibility) requires rejection of this plausible view. That is the aim of Chapter Five. There I argue that the plausible view is inconsistent with some of our intuitions about when it is rational to resist *acquiescent* courses of action. When acquiescence is an issue, typically, the actor's otherwise most favoured option

involves going along with another person's unreasonable intentions or dispositions. I try to isolate a type of reason not to acquiesce, which I claim cannot be accounted for if we adhere to the plausible view about expected effectiveness. And I try, further, to develop an alternative account of the assessment of expected effectiveness, which (it is hoped) underscores the argument from intuition.

The view of deliberation which results has some similarities with other moral views, which I explore in Chapter Six. All of the views discussed may be contrasted with one form of consequentialism, which I call individual act-consequentialism, but they differ also amongst themselves. It is useful, in gaining a clearer picture of the method of evaluating expected effectiveness which I propose, to explore these contrasts.

The chapters which follow range over a number of topics – starting with the conceptual relationship of justice and responsibility, moving onto general considerations about the nature of political argument and judgements of possibility, and finishing with the rationality of acquiescence and the view of deliberation required to explain it. What unites them is a concern with the role and importance, in political argument, of judgements about possibility – judgements, that is, about which things are possible, and which possibilities are practically relevant. My central argument is that the proper basis for such judgements is an understanding of what persons could do, which lamentably saddles political argument with the unresolved problems of our understanding of agency.

However, the chapters do not attempt to analyse the role and importance of judgements about possibility head-on, to produce a general theory of their grounds. That would be too difficult, and in any case might not produce significant results

(their grounds may resist codification, being too context-dependent). So instead I shall take an oblique route, starting on the surface with the idea of naturalism, then later touching some narrow aspects of the deeper issues. I shall not propose a general account of judgements of political possibility: I claim only that agency-implicating judgements are essential to discriminate between relevant and irrelevant possibilities.

Studies of the nature of normative political argument usually concentrate their attention on its most obvious distinguishing marks – so they discuss the nature of political values, including their cognitive status, issues of value-conflict, varieties of trade-off or other rules for adjudicating conflict, and the nature of rights-claims. Such studies greatly improve our understanding of political argument, but they leave something out. Normative arguments rely on descriptions of circumstances – descriptions at a very general level of the kind of world we live in, and at a very particular level, of the circumstances facing a certain actor. Much of the time dispute turns on these descriptions, with their often hidden suggestions and hints at what is possible for us.

Finally, I want to add a brief word about theoretical models of practical reasoning. In recent years there has been a reaction against (what is seen as) over-formalisation in conceptions of practical reasoning.[8] It is felt that theoretical models of practical reasoning do not describe what agents do, when they deliberate, *and* that this is to the models' discredit, not the agents'. For what the agents have, and the models do not succeed in codifying, is sensitivity, or judgement, or virtuous habits. In what follows I debate the merits of different theoretical accounts of practical (political) reasoning, so it may be thought that my arguments are vulnerable to this charge.

There is something in the criticism, but it is difficult to say exactly what it is. It is possible to view practical reasoning in too formal a way, and to think that there must be some mechanical decision-procedure capable of guiding our actions in every situation. But arguments about the relative importance of principles as against judgement seem often to be arguments at crossed-purposes: anyone with a remotely sophisticated understanding of general rules knows that judgement is required to apply them.[9] If the question is *at what point* we appeal to the idea of judgement, rather than make recourse to some more abstract principle, in explaining our thoughts, the issue is a real one, but very tricky to decide. I take the view that the burden of proof lies with the critic, in this case, in explaining exactly why we should not appeal to an abstract principle on a particular occasion.

It should not be assumed that those who propose theoretical models of practical reasoning suppose that deliberators pass consciously through the stages they describe; nor that they suppose that deliberators ought to do so. The criterion of realism which governs this kind of theorising is more subtle than these criticisms allow. A theory can explain or account for rational processes without either describing the conscious experience of those processes or implying that the conscious experience ought to be like that. We almost certainly do better not to try to pass consciously through the stages described by theoretical models of practical reasoning, because to

---

[8] This complaint is often voiced. See, for example, Charles Larmore, *Patterns of Moral Complexity* (Cambridge: Cambridge University Press, 1987), Chapter One. The issues are discussed in S. Scheffler, *Human Morality* (New York and Oxford: Oxford University Press, 1992), pp. 38-51.

[9] Kant, for example, was often at pains to point this out. Having claimed that 'judgment is the faculty of subsuming under rules', he adds, "If [general logic] sought to give general instructions how we are to subsume under these rules, that is, to distinguish whether something does or does not come under them, that could only be by means of another rule. This in turn, for the very reason that it is a rule, again demands guidance from judgment. And thus it appears that, though understanding is capable of being instructed, and of being equipped with rules, judgment is a peculiar talent which can be practised only, and cannot be taught." I. Kant, *Critique of Pure Reason*, translated by N. Kemp Smith (Basingstoke: Macmillan, 1929), A 133/B 172, p. 177.

do so would be very costly. But those models may nevertheless explain the extent to which the conclusions we reach by some other method are rational. That, I take it, is their aim.

Chapter Two

Justice, Responsibility, and Fairness

This chapter begins our investigation of the conceptual relationship between justice and responsibility.[1] It outlines in broad terms the possible views of that relationship, and asks what kind of consideration could justify our adopting one of those views rather than another. In particular, we shall consider in this chapter some arguments for the view that justice tracks responsibility. I shall claim that these arguments are not successful.

Whilst our immediate concern is with the relationship between justice and *responsibility*, the underlying issue is whether or not we should try to explain what justice requires by reference to what I shall call *agency-implicating* concepts. A concept is agency-implicating if its correct application depends on understanding the range of possibilities which are within a person's reach, where this idea cannot be explained solely in terms of our predictive theories. I shall explain this in section 1. For the moment, the point is that the concepts of desert and responsibility, and associated ideas, can often be given a non-agency-implicating, or as I shall say *naturalist*, interpretation. One may seek to avoid the problems which agency-implicating concepts bring, by naturalising them

wherever possible. Our underlying concern in this chapter and the next is to understand and evaluate the naturalising strategy in accounts of justice.

1. Four views on Responsibility and Justice

Common sense morality holds that what justice requires depends on what people are responsible for.[2] Suppose that someone is (uncontroversially) badly off. Common sense morality holds that justice may not require, or permit, transfer of resources to that person, if she is responsible for her circumstances. Similarly, it finds a moral difference between wealth for which a person is responsible, and wealth which is simply inherited.

Common sense is often wrong, but the view that what justice requires depends on what persons are responsible for is not obviously false. It does not describe a single account of justice, however, but rather a family of views. These views have two ideas in common:

(1) *responsibility* is conceptually independent of *justice*;

and (2) *justice* is conceptually dependent on *responsibility*.

The first of these ideas may seem uncontroversial (though we'll see in a moment that some deny it), but the second idea is certainly controversial.

---

[1] Throughout this thesis (unless stated otherwise), by 'justice' I mean justice in holdings, as opposed to retributive or commutative justice.

The idea of conceptual dependence is not a psychological notion, and neither is it primarily an epistemological notion. Hence (2) does not claim anything about psychological associations of ideas. The best psychological explanation of our ideas about justice may not mention responsibility. Neither is (2) a claim about how we should justify claims of justice (though it might entail such a claim). It is a claim, instead, about what justice is. It says that justice is partly constituted by appropriate response to persons' responsibility. I shall refer to theories which incorporate this idea as *responsibility-tracking* conceptions of justice.[3]

Notice that (2) does not claim that justice consists wholly of appropriate response to responsibility.[4] Responsibility-tracking conceptions of justice may be simple or pluralist. Simple views claim that justice consists only of appropriate response to responsibility; pluralist views add extra components. Pluralists may claim, for example, that justice consists also of appropriate response to need. We are interested in the respects in which simple and pluralist views are alike, so it will be easier for us to consider, most of the time, simple views only (even though they are less plausible).

Notice also that responsibility-tracking views, whether simple or pluralist, may conceive 'appropriate response to persons' responsibility' in different ways. One way of conceiving it treats responsibility as a disqualifying condition for rectifying real

---

[2] See S. Scheffler, "Responsibility, Reactive Attitudes, and Liberalism in Philosophy and Politics", *Philosophy and Public Affairs* 21 (1992), for discussion of the opposing views taken by common sense morality and recent political philosophy on the relationship between justice and responsibility.

[3] Such theories *aim* to specify institutional arrangements (or principles), such that, if they were realised (or adhered to), goods would be distributed in a responsibility-tracking way. They may fail in their aspiration and still be 'responsibility-tracking' in the way I'm using the phrase.

[4] Compare: the density of an object depends on its mass. (i) This is neither a psychological nor an epistemological dependence, but a conceptual one. (ii) The density of an object depends also on its volume. This is quite consistent with the claim that density depends on mass.

distributions of holdings in accordance with some specified *pattern* of holdings.[5] On this view, we seek to establish and maintain a certain pattern of holdings in the relevant population. One such pattern (an implausible but simple one) is: *strict equality of holdings at all times*. One might combine adoption of that pattern with the view that responsibility is a disqualifying condition for rectification. There would then be two stages in assessing the justice of a distribution: first, one identifies real divergences from the specified pattern; second, one discriminates between those divergences (positive or negative) which are the responsibility of the persons they attach to, and those which are not, treating only the latter as candidates for rectification.

A subtly different kind of view builds reference to responsibility into the specification of the pattern itself. The ideally just pattern of holdings is then something like: *distribution according to responsibility*. Compare this view with a two-stage view in which the pattern specified is one of strict equality. The difference between them lies in the justification of equalities between persons.[6] The two-stage view operates with a presumption in favour of equality, departures from which are justified only if they reflect responsibility. The direct incorporation of responsibility-tracking into the pattern itself has the effect of demanding justification in terms of responsibility for equalities as well as inequalities.[7] So if I've got the same as you, but you're responsible (somehow) for all

---

[5] For the distinction between patterned and entitlement theories, see R. Nozick, *Anarchy, State, and Utopia* (Oxford: Basil Blackwell, 1974), pp. 150-160. Responsibility-tracking patterned theories are examples of what Nozick calls 'historical patterned' theories. They require information not just about the current time, but about how things came about.

[6] Of course, in general the difference between the views is not the justification of equalities, but rather the justification of holdings which fit the pattern used in the two-stage view.

[7] Susan Hurley may be assuming that responsibility-tracking conceptions of justice must incorporate the responsibility-tracking condition directly in the specified ideal pattern, when she argues that the responsibility-tracking aim, plus global scepticism about responsibility, does not generate an egalitarian position. See S. L. Hurley, "Justice without Constitutive Luck", in A. Phillips Griffiths (ed.), *Ethics* (Cambridge: Cambridge University Press, 1993), p. 185, 192. Even if she doesn't restrict

of your holdings, whilst I'm responsible for only part of mine, I do better under the two-stage view than the direct view.

A third conception of appropriate response to persons' responsibility dispenses with any patterning requirement. One might think that *any* distribution which is generated by just transfer of justly held goods is bound to reflect persons' responsibility in an appropriate way. So long as there is some just starting point, allowing individual choices (constrained only by the principles of just transfer) to have their effect will inevitably preserve justice in ways which reflect responsibility, since changes in the original distribution will always be the result of choice.[8] Choices do not have to be corrected, so long as they are constrained by the correct principles and made against the correct background. Whatever distributional outcomes result, they will, in such circumstances, reflect responsibility.[9]

Whether appropriate response to responsibility is conceived in terms of a two-stage view, a direct pattern view, or an historical entitlement view, responsibility-tracking

---

responsibility-tracking conceptions in this way, it would be an understandable restriction. That's because two-stage views treat responsibility as irrelevant to justice so long as the distribution under examination fits the pattern which is applied at the first stage. Within that region, justice does not track responsibility. On the other hand, though, even views which incorporate responsibility directly in the pattern may imply that the difference between some just and some unjust distribution has nothing to do with responsibility, so long as those views are pluralist in the way I've described. Hence I favour the more ecumenical policy.

[8] Hillel Steiner has claimed that: " . . . historical entitlement principles of transfer and rectification are not really independent prescriptive rules at all . . . Such principles necessarily govern any arrangements in which persons take exclusive responsibility for their own actions, in the sense that general compliance with them is a necessary condition of the consequences (valuable or disvaluable) of anyone's actions not enforceably accruing to anyone else." H. Steiner, *An Essay on Rights* (Oxford: Blackwell, 1994), p. 226. Thus, according to Steiner's view, arrangements *must* comply with historical-entitlement principles to track responsibility properly. Two-stage or direct patterned views thus track responsibility only insofar as they are practically equivalent to an historical-entitlement view. (See note 9 below.) See also H. Steiner, "Choice and Circumstance", unpublished manuscript.

[9] Some historical-entitlement theories may be equivalent in practice to some patterned theories. For example, if Steiner is right about satisfaction of historical-entitlement principles being a necessary condition of tracking responsibility, the patterned theory which specifies *distribution according to responsibility* would, presumably, be practically equivalent to an historical-entitlement theory (unless

conceptions of justice seek to explain what justice requires in terms of persons' responsibility. Hence, it seems, such conceptions must incorporate (1), the idea that responsibility is conceptually independent of justice.[10] Only then could responsibility perform its explanatory role with respect to justice.

Not all views of responsibility are consistent with (1). Scanlon's view, for example, is that choice has moral significance only if it is made in reasonably just circumstances.[11] It seems to be a consequence of this view that the concept of responsibility is explained by – and not as usually thought, explanatory of – justice. Whether or not someone is responsible for something depends on what they choose to do in (reasonably) just circumstances – or, more directly, which costs it is fair or just to let them bear. A straightforward logical circle seems to be involved in seeking to explain what justice requires in terms of responsibility, if responsibility is understood in this way.

I shall refer to interpretations of responsibility along the lines which Scanlon suggests, as *naturalist* interpretations of responsibility. They are 'naturalist' in the sense that they offer an interpretation of 'responsibility' which apparently does not implicate our understanding of agency. A concept implicates our understanding of agency if it cannot be explained without reference to counterfactual judgements about *what persons could do*. According to the common-sense view, responsibility is inextricably tied-up with our understanding of persons as agents. In contrast, the naturalist aims to interpret 'responsibility' in such a way that we can tell what someone is responsible for simply by

---

the extra conditions of responsibility-tracking intervene). Compare R. Nozick, *Anarchy, State, and Utopia* (Oxford: Basil Blackwell, 1974), pp. 156-157.

[10] In fact, they could hold that responsibility and justice are conceptually interdependent. I'll discuss this view briefly in a few paragraphs' time, and in Chapter Three.

[11] T. M. Scanlon, "The Significance of Choice", in S. Darwall (ed.), *Equal Freedom. Selected Tanner Lectures on Human Values* (Ann Arbor: University of Michigan Press, 1995), pp. 76-78.

finding out about their *actual behaviour*. If one knows which institutional arrangements are just, and one knows how people have actually behaved, and institutions respond to persons' actual behaviour, not to what they could have done, then one may know which costs it is fair to let persons bear (or which benefits to enjoy) without speculating about what persons could have done.

Note that one may be a naturalist about any concept which involves the idea that choice is morally significant, since it is really that idea that naturalism is oriented towards. Thus one may be a naturalist about desert, for example. The appeal of naturalism is that our understanding of agency is subject to well-known philosophical difficulties, as well as a good deal of less esoteric controversy. People disagree about the extent of persons' responsibility or desert, and this disagreement is underwritten by the problems of free will which have exercised philosophers so much. Naturalist interpretations of these concepts, in contrast, aim not to be agency-implicating. They try to limit themselves to actual human behaviour and rule-governed arrangements.[12]

In order to see whether someone deserves something on a naturalist understanding of desert, for example, one asks whether that person behaved in ways which give rise to legitimate expectations.[13] The explanatory role usually taken by our understanding of agency is taken instead by our understanding of just arrangements. Hence naturalists deny (1), the idea that responsibility is conceptually independent of justice. Since this view is

---

[12] Someone could hold that there are two relevant senses of 'responsible': an agency-implicating sense (R_1), and a non-agency-implicating sense (R_2) according to which someone is responsible for something if and only if it is just to let her bear the costs (enjoy the benefits) of it. R_2 is conceptually dependent on justice, as the naturalist claims. But if this person were to claim also that justice is conceptually dependent on R_1 (so that R_2 depends on J depends on R_1), they would not be a naturalist in my sense. I am grateful to Andrew Williams for discussion of this point.

[13] For the idea of legitimate expectations, or what Scanlon calls "institutional desert", see J. Rawls, *A Theory of Justice* (Oxford: Oxford University Press, 1972), section 48; Scanlon, "The Significance of Choice", pp. 73-76.

essential to straightforward responsibility-tracking conceptions of justice, naturalism is at odds with these views.[14]

Naturalists hope to avoid the obscurity and theoretical problems associated with standard views of agency. I shall assume that standard views of agency share the following feature in common: someone's being an agent consists in her having certain possibilities within her reach, certain things she could do or be, where this range of possibilities cannot be explained simply in terms of our best predictive theories. I make this assumption on the basis of the following argument. Perfect predictive theories would give us knowledge of the future actual world. If someone in that world is an agent, according to standard views, it will be the case that she could have done things other than those which she actually does. Hence perfect predictive theories cannot tell us the whole range of things which persons could have done. Hence *our best* predictive theories cannot do so either.

This argument does not assume a particular stance on issues of determinism or of compatibility between determinism and freedom. These are metaphysical issues, whereas my argument was epistemological. It concerned what we could know using a certain kind of theory, and does not rest on any particular assumptions about the causation of human action, or the compatibility between determinist causation of human action and agency, free will, or morally significant choice. I claim only that, according to our standard understanding of agency, a claim that a person could have done X at t is not always refuted by the claim that in fact she did not do X at t. Our best predictive theories can

[14] See Chapter Three below for further discussion of Scanlon's view, in connection with desert. It turns out that whether a logical circle is involved in explaining justice in terms of responsibility (or desert) depends on how we interpret 'reasonably just circumstances'.

hope only to give us knowledge of the things that people (will) actually do, and so cannot exhaust our understanding of agency, in this standard sense.[15]

Our grasp of the possibilities within a person's reach is difficult to explain, but its existence seems to be a presupposition of the idea that our understanding of agency could inform our understanding of justice. For it is this grasp which makes our understanding of agency something other than a record of actual behaviour – whether that behaviour is in the past, the present, or, in the case of predictive theories, the future. So it is not surprising that naturalising otherwise agency-implicating concepts leaves them incapable of explaining what justice requires. Naturalism is committed to finding other grounds to explain what justice requires, and thence to explain the (naturalised) concepts of responsibility and desert.

Responsibility-tracking conceptions of justice, some elements of which may be found in common sense morality, are characterised by their espousal of both (1) and (2). Naturalist views are characterised, in contrast, by their denial of both (1) and (2). Two remaining views of the relation between justice and responsibility are the *mutual independence* view, and the *mutual interdependence* view. Let me explain these in turn.

The mutual independence view claims that justice does not depend on responsibility, and *vice versa*. The two concepts have no essential relation to each other. One such view is utilitarianism, in its standard forms. On standard utilitarian accounts of justice, what justice requires depends only on what would increase the aggregate or

[15] I've said that the standard view of agency does not rely on any particular assumptions about determinism or compatibility. The flip-side of that point is that the standard view is not, as it stands, a well-articulated view. It simply reports, or so I believe, one feature of our ordinary understanding of agency. It is this feature which causes trouble for naturalism, and which naturalists thus reject.

average utility within some population.[16] Persons' responsibility may enter into this calculation, of course, but only indirectly, insofar as it contributes to expected utility. Similarly, utilitarianism is not standardly committed to any particular view of responsibility. One could combine Scanlonian views about responsibility with a utilitarian account of what justice requires, but that is an addition to utilitarianism; it is neither a presupposition nor an implication of it. Utilitarians typically deny (2) and are agnostic about (1), at least *qua* utilitarians.

Mutual interdependence views are less familiar. They try to combine assertion of (2) with denial of (1), which is an unusual combination of views. Responsibility-tracking conceptions explain (2) by reference to (1): it is the fact that we have an independent grasp of what persons are responsible for, that explains how our judgements of responsibility can give content to the concept of justice. Without (1), it is difficult to see how (2) could be true. This makes mutual interdependence views unstable, if not downright contradictory. I'll return to them in Chapter Three, when I discuss Scanlon's views more fully.

For the moment, I want to focus our attention on the dispute between responsibility-tracking conceptions of justice, and naturalist views. It is sometimes difficult to tell which kind of view a particular writer endorses. He or she may tell us only that, under just arrangements, there is a relation between each person's responsibility and her holdings. That's ambiguous between a responsibility-tracking view, according to which a person's responsibility may be determined independently, and used to explain the justness of her holdings, and a naturalist view, according to which things are exactly

opposite. Yet the difference between these is crucially important. We do not understand conceptions of justice which assert a conceptual connection between justice and responsibility, unless we understand the direction of explanation which is asserted between the two concepts.

It is not easy to say in advance which kind of view is more likely to be correct. Agency-implicating concepts, such as the idea of responsibility which responsibility-tracking views rely upon, have well-known problems. These problems may make a satisfactory responsibility-tracking conception of justice impossible. On the other hand, naturalism faces difficulties of its own. It must show that the concept of justice has enough independent content to serve as an explanation of responsibility, desert, or other similar concepts. In the next section we'll consider how this dispute might be resolved.

2. Arguments for and against the Responsibility-Tracking View

How might someone argue for or against the responsibility-tracking view? The issue on which it takes a stand is a particularly abstract one. We are used to arguments for and against substantive theories of justice, but these come too late on the scene to address our issue. Substantive proposals are likely to have taken a stand on the relationship between justice and responsibility before the explicit argument begins. Criticisms of those positive proposals, on the other hand, often are addressed to the proposals' internal coherence – so unless an incoherence is identified and traced solely to a stance on the relationship between justice and responsibility, these too are unlikely to include explicit argument for

[16] I shall discuss only act-utilitarian views, for simplicity. I believe that other versions of utilitarianism share the mutual independence view of the relationship between justice and responsibility. (For

or against the responsibility-tracking view. Yet the stance one takes on this issue is likely to have a considerable influence on the shape of one's thinking about justice.

Our problem is not exactly epistemological; it is to see what would count as an argument in this area. It will not be solved, for example, by arguing the merits of foundationalist as against coherence accounts of reasoning about justice.[17] It's a problem instead of seeing what kinds of considerations could sway us at such an abstract level, of knowing where to look for reasons to guide our judgement one way rather than the other. Without such reasons our stance will look unnervingly like mere choice, or 'plumping' for one view rather than another.

One putative argument may be quickly dismissed. Some people might find it simply intuitively compelling that the responsibility-tracking view is correct (the converse pre-theoretical intuition seems less likely). 'Justice simply *is* a matter of treating people as responsible for some things and not others', they may say – by inspection of the concept of justice, as it were. Obviously this won't do as an argument. It lacks articulation, amounting merely to assertion of a conviction on the issue.

A better argument for the responsibility-tracking view proceeds from an account of the morally relevant characteristics of persons. Justice, it might be claimed, should treat people in accordance with their highest faculties, morally speaking; and it might be claimed too that the highest moral faculty of humans is their capacity for choice, for determining their own actions. One might conclude as a result that just institutions ought to respond to persons as choosers, and that this requires them not to interfere with those

---

discussion of rule-consequentialism in a different context, see Chapter Six.)

[17] On coherentism, see J. Dancy, *Introduction to Contemporary Epistemology* (Oxford: Blackwell, 1985), chapters 8 and 9.

aspects of their circumstances, whatever they are, for which they are responsible in virtue of their choice.

This argument, broadly Kantian in its portrait of the highest human faculty as that of self-determination, improves on the argument from direct inspection of the concept of justice, since it is at least minimally articulate. On the other hand it suffers, perhaps, from its controversial premise about what the highest human faculty is, morally speaking. On the face of it, it is not clear why the faculty of self-determination should have pre-eminence amongst other morally relevant human capacities, such as the capacity for pleasure, or the capacity for good moral judgement. One might doubt, on those grounds, whether the broadly Kantian argument shows that justice tracks responsibility.[18]

More damaging to this argument from the importance of the capacity of self-determination, however, is the fact that a naturalist conception of the relationship of justice and responsibility could equally well be supported by it. One need only add the naturalist premise, that choice has moral significance only if it is made in relatively just circumstances, to conclude from the broadly Kantian line of thought that people ought to be held responsible for just those things which are assigned to them in virtue of their choices in just institutional arrangements. So the Kantian argument fails to promote the responsibility-tracking view over its main rival. At most it establishes a connection between responsibility and justice; it does not establish a direction of dependence between them.

---

[18] Rawls includes in his 'political conception of persons' ideas about their capacity to form and pursue plans which are, arguably, akin to the Kantian idea canvassed in this paragraph. But his conclusion is not that justice should track responsibility – it is that, " any workable political conception of justice . . . must count human life and the fulfilment of basic human needs as in general good . . .". J. Rawls, " The Priority of Right and Ideas of the Good", *Philosophy and Public Affairs* 17 (1988), p. 254. That is, he emphasises the moral importance of having the means to pursue one's plans (primary goods), rather than the moral importance of being able to form plans.

An alternative argument for the responsibility-tracking view points to the instrumental value of leaving unchanged those aspects of persons' circumstances for which they are responsible. This broadly Millian argument claims that people's lives generally go better if they are left to bear the costs, and enjoy the benefits, of their choices. We should reject this argument as well. No doubt there is considerable instrumental value in leaving people to bear the costs or enjoy the benefits of their choices. But it is equally clear that there is, in very many cases, considerable instrumental *dis*value in adopting this policy. The results of choices made in rotten circumstances are likely to be rotten, and it is little consolation to be told that at least one owns them. It is often unjust to let the outcomes of choices stand. Hence this Millian argument, too, could just as easily work in favour of a naturalist view, according to which the outcomes of choices should be left to stand just in case the circumstances in which they were made were just.

Neither the Millian nor the Kantian arguments get us very far. Perhaps we would do better to look for arguments *against* the responsibility-tracking view. If successful, they would take us some distance towards naturalism (though one would have still to dismiss the rival claims of mutual independence and mutual interdependence).

One such argument claims that we should not seek to explain justice as tracking responsibility, because of the philosophical difficulties and general disagreement which afflict use of the concept of responsibility. These difficulties have theoretical as well as polemical implications. They may prevent us from working out the details of what justice requires, if our judgements of responsibility are not sufficiently determinate. They may

also undermine the polemical strengths of our preferred theory of justice, insofar as they make its justification depend on controversial judgements of responsibility.[19]

What kind of force does this argument have? It appeals to the theoretical and polemical consequences of tying justice to responsibility, and so it is an *instrumentalist* argument. Instrumentalist considerations, we often think, do not provide reasons for belief. They provide us with reasons for acting one way rather than another, but they do not provide us directly with reasons for believing one thing rather than another.[20] Hence, we may think, the philosophical difficulties and controversy in application of the concept of responsibility are at most reasons not to employ the concept of responsibility in trying to persuade someone of what justice requires; they are not reasons to think that what justice requires does not depend on who is responsible for what. If judgements of responsibility are indeterminate or controversial, a proponent of the responsibility-tracking view might say, that's just too bad for our theory of justice.[21]

The instrumentalist considerations are not shown to be irrelevant by these reflections. Admittedly they cannot be the final test of a theory. It may be simply too bad for our understanding of justice, that justice depends on responsibility. But often our direct concern is not with the final test of a theory: it is with the likelihood that a general theoretical approach will succeed. In such cases it is too early to tell whether some

---

[19] Susan Hurley makes these points against responsibility-tracking ('luck-neutralising') conceptions of justice, in her "Justice without Constitutive Luck", pp. 186-187.

[20] One has to be careful here. Suppose that believing in God makes me happy. That's not a reason (for me or anyone else) to believe *that God exists*, but it is a reason to believe something – namely, that I have a certain psychological complexion. Reasons for action may at the same time be reasons for belief, but they are generally not reasons for believing that the good state of affairs, which provides the reason for action, exists.

[21] The contrast between instrumental reasons and proper reasons for belief loses sharpness if one is a coherentist. A coherentist who takes it that *justice* is not indeterminate, and who thinks that *responsibility* is indeterminate, may, on that basis, conclude that justice does not track responsibility. But if this argument escapes the charge of instrumentalism, it lacks force. Coherentists must

particular theory (a responsibility-tracking theory, say) will turn out to be justified, all things considered. We want instead to anticipate the likely success of going down different avenues.

If we adopt this theory-building stance, instrumentalist considerations have some force. If we do not have access to the all-things-considered assessments of various well-developed theories, it is quite legitimate to take note of the theoretical problems associated with different theory-building strategies. We ask what the theoretical consequences of taking some route would be, and may make a decision about which route to follow on that basis. We may, for example, decide to look for a naturalist theory of justice because we find the theoretical consequences of trying to explain justice in terms of responsibility to be unappealing. Instrumentalist considerations are in order so long as we do not treat them as final reasons for belief, but only reasons for theory-building in one way rather than another.

We've considered some arguments in favour of the responsibility-tracking view, and one argument against it, and in each case we've found the arguments to be inconclusive. In the rest of this chapter, I want to examine a different argument in favour of the responsibility-tracking view, which seems to me to be stronger than those we've considered so far. I'll begin, in the next section, by explaining what it is.

---

discriminate between intuitions, and the intuition that 'justice is not indeterminate' looks to me no more antecedently firm than the conviction that 'justice tracks responsibility'.

3. The debate about what egalitarians should try to equalise

Since Amartya Sen's classic paper of 1978, there has been a lot of discussion amongst liberal egalitarians of what it is that egalitarians should try to equalise.[22] Proposals for the *equalisandum* include the following: welfare; opportunity for welfare; primary goods; resources; capability to function; and access to advantage.[23] Clearly, which of these *equalisanda* one adopts has great significance for the egalitarianism which results. The practical implications of, say, welfare egalitarianism may in many cases be directly opposed to those of, say, resource egalitarianism. (Suppose that Jill has more resources but less welfare than Jack, and that we can help only one of them. Welfare egalitarianism favours helping her, whilst resource egalitarianism favours helping him.) Hence it is not surprising that this debate has attracted a good deal of attention.

Much of the debate has consisted of discussion of the merits of compensating people with different kinds of disadvantage. As we'll see, a whole cast of characters is involved. Louis, for example, has expensive tastes: he can't derive as much welfare from a given bundle of resources as others can, because anything other than ancient claret and

---

[22] A. Sen, "Equality of What?", in S. M. McMurrin (ed.), *Liberty, Equality, and Law* (Salt Lake City: University of Utah Press, 1987).

[23] It is difficult to find a modern exponent of an equality of welfare view, though it seems to be a favourite way in which to begin a critical discussion. For discussion of these proposals, see: R. Dworkin, "What is Equality? Part 1: Equality of Welfare", *Philosophy and Public Affairs* 10 (1981); A. Sen, "The Standard of Living: Lecture 1, Concepts and Critiques", in G. Hawthorn (ed.) *The Standard of Living* (Cambridge: Cambridge University Press, 1987); G. A. Cohen, "On the Currency of Egalitarian Justice", *Ethics* 99 (1989), pp. 906-944. On opportunity for welfare, see R. J. Arneson, "Equality and Equal Opportunity for Welfare", *Philosophical Studies* 56 (1989). Rawls's discussion of primary goods is in *A Theory of Justice*, section 15 and *passim*, and "Social unity and primary goods", in A. Sen and B. Williams (eds), *Utilitarianism and beyond* (Cambridge: Cambridge University Press, 1982). On equality of resources, see R. Dworkin, "What is Equality? Part 2: Equality of Resources", *Philosophy and Public Affairs* 10 (1981). For Sen's idea of capability to function, see A. Sen, "Capability and Well-Being", in M. Nussbaum and A. Sen (eds), *The Quality of Life* (Oxford: Clarendon Press, 1993). For Cohen's idea of access to advantage, see his "On the Currency of

plovers' eggs offends his palate. Not only that, but Louis deliberately cultivated those tastes. One question is whether egalitarians should favour giving Louis extra resources. Suppose our considered judgement is that they should not. That might be, in the first instance, a point against welfare equality views, supposing that we are confident that our considered judgement is correct, and sure that such views recommend giving Louis extra.[24]

But that does not exhaust the significance of the example. We may go on to ask what the relevant difference is between someone like Louis, who has expensive tastes, and someone who requires special facilities to move around, say. Expensive tastes and handicaps are alike in one respect: they hinder people from deriving welfare from resources. If our considered judgements vary across the cases (favouring extra resources for the otherwise immobile person, but not for Louis), we should like an explanation of the difference between the cases.

Several explanations could be forthcoming in this case. One could point to the greater moral importance of mobility as compared to Epicurean pleasure, perhaps explaining the greater importance of the former in terms of basic human needs. One might, however, claim that Louis ought not to be compensated because his expensive tastes are his responsibility, whereas handicaps typically are not their bearers' responsibility. After all, the example was specified to include the idea that Louis

---

Egalitarian Justice", *Ethics* 99 (1989), and G. A. Cohen, "Equality of What? On Welfare, Goods, and Capabilities", in M. Nussbaum and A. Sen (eds), *The Quality of Life*.

[24] Dworkin introduces Louis at p. 229 of "What is Equality? Part 1: Equality of Welfare".

deliberately cultivated his expensive tastes. One might doubt whether all expensive tastes are their bearer's responsibility, but Louis' case is not like that.[25]

The argument for the responsibility-tracking view which I want to consider is as follows. Our judgements across a range of individual cases are best explained by our judgements about the responsibility of the individuals concerned for their disadvantages. That is, judgements of responsibility do not seem capable merely of explaining our considered judgements in individual cases. They seem capable of explaining them across a number of different cases, and in doing so, of explaining why we have different intuitions in these different cases. We intuitively seek to track responsibility. The fact that a large number of considered judgements in different cases are explained by our judgements of responsibility, supports the claim that justice tracks responsibility.

This argument is most naturally interpreted as relying on a coherence account of political reasoning, along the lines of Rawls's idea of reflective equilibrium.[26] Coherence accounts (Rawls's included) are distinguished by their view that the coherence of a set of claims itself provides justification of those claims. This is to be contrasted with the view that logical relations between claims can only transmit whatever justification is gained from other sources, or, in the case of contradiction, show that one of a pair of beliefs must be false.[27] The argument for the responsibility-tracking view which I've just sketched

---

[25] Cohen argues that whether or not we should compensate those with expensive tastes depends on whether or not they "can reasonably be held responsible for them". As it stands, this is ambiguous between a responsibility-tracking interpretation and a naturalist interpretation – but I think it is clear that Cohen intends the first of these. The quoted phrase follows his introduction of the case of *Paul*, who is not responsible for his expensive taste for photography. In that discussion, Cohen stresses the importance of the fact that Paul's taste is involuntary. See G. A. Cohen, "On the Currency of Egalitarian Justice", p. 923, and section 5 below.

[26] For Rawls's idea of reflective equilibrium, see J. Rawls, *A Theory of Justice*, pp. 19-22.

[27] Coherence must involve more than mere consistency, if coherentism is to be plausible. The various beliefs must not only not contradict each other, but also support each other. The question is what the notion of support amounts to. One could cash it out in terms of explanation, perhaps, but it's not

seems to rely on the thought that the fact that our considered judgements, and our judgements of responsibility, hang together as a coherent set of beliefs, itself provides justification of both our judgements about responsibility, and our judgements about who should be compensated. And the explanatory relations between these two sets of justified beliefs are then held to justify the responsibility-tracking view.

However, this is not the only way the argument can be construed. Coherentism is a controversial doctrine, but the argument for the responsibility-tracking view need not rely on it.[28] The argument must claim two things: that our considered judgements about compensation in individual cases, or a sufficient number of them, are justified; and that these judgements are explained by our judgements of responsibility. The coherentist version of the argument attempts to solve both of these problems at once – claiming that our judgements of responsibility explain our judgements in individual cases, hang together with them in a coherent set, and thus that both kinds of judgement are justified. But we can imagine taking the steps separately.

Barring scepticism, we should find it possible that our considered judgements about compensation are justified according to our preferred account of justification of normative political beliefs, whatever that is. What about the second step? Not just any explanation of the considered judgements by our judgements of responsibility will do. It might be that we get the right answer by intuitively tracking responsibility in our considered judgements, but for the wrong reasons. We could compare the case with that

---

obvious that the concept of *explanation* is any clearer than that of *justified belief*. For an attempt to spell-out the conditions of coherence, see N. Rescher, *The Coherence Theory of Truth* (Oxford: Clarendon Press, 1973).

[28] For recent doubts about coherentism, see J. Griffin, *Value Judgement* (Oxford: Clarendon Press, 1996), Chapter 1; and J. Raz, *Ethics in the Public Domain* (Oxford: Clarendon Press, 1994), Chapter 13.

of Newtonian explanations of motion. As with the argument for the responsibility-tracking view, here we have reasons to think that our judgements (calculations of motion) are correct, reasons which are independent of the theory concerned (Newton's theory of motion). But the mere fact that these judgements are explained by Newton's theory does not suffice to show that theory to be correct, nor to establish Newtonian calculations as the basis of our calculations of motion.

What could show that? We might accept the conclusion if we had good reasons to think that the theory concerned (Newtonian physics, or the responsibility-tracking view) was the *best available explanation* of our correct judgements. That is, if the theory concerned is the best way we can see of bringing the various correct judgements under a common explanation, we might take that to be justification of the theory, and justification for regarding the theory as a correct portrayal of the basis of such judgements. I shall suppose that this is so.

In fact, something like this argument can be found in the writings of contributors to the debate about the proper *equalisandum* of egalitarianism. For example, John Roemer has written a number of papers exploring the relationship between justice and responsibility, finding a common "egalitarian ethic" in the writings of Dworkin, Arneson, and Cohen. This ethic holds that society should indemnify persons against disadvantages which are not their responsibility, but not against those which are their responsibility. Roemer then tries to show how an egalitarian planner could make judgements of responsibility in practice.[29] Arneson, on the other hand, suggests that many of our

[29] See especially J. E. Roemer, "A Pragmatic Theory of Responsibility for the Egalitarian Planner", *Philosophy and Public Affairs* 22 (1993), pp. 146-150; see also J. E. Roemer, "Equality of Talent", *Economics and Philosophy* 1 (1985), pp. 151-188; J. E. Roemer, "Equality of Resources implies

considered judgements in individual cases reflect the following norm: "Other things equal, it is bad if some people are worse off than others through no voluntary choice or fault of their own."[30]

The argument for the responsibility-tracking view which I have described is clearest in Cohen's review of the debate, however.[31] Cohen engages in a critique of Dworkin, amongst others, claiming that Dworkin's resourcist conception of equality is attractive insofar as it reflects a responsibility-tracking ideal, but that it only imperfectly reflects that ideal. Hence, according to Cohen, Dworkin's conception should be modified, better to reflect the responsibility-tracking ideal as revealed by our considered judgements across a range of cases.[32]

I'll examine in detail Cohen's criticism of Dworkin's views, and possible Dworkinian rejoinders, in section 5. Our interest is in whether Cohen's arguments fulfil the requirements of the argument for the responsibility-tracking view which we've just specified. We want to know whether the judgements in individual cases which Cohen affirms are sound ones, and whether the responsibility-tracking interpretation of them is the best available explanation of them. These are tough standards, designed to evaluate the *argument* for the responsibility-tracking view. That view, and Cohen's egalitarian convictions, may be correct even if the argument we're considering fails. Before we evaluate that argument, however, it will be helpful to discuss in a more general way what

---

Equality of Welfare", *The Quarterly Journal of Economics* (1986), pp. 751-784; J. E. Roemer, "Egalitarianism, Responsibility, and Information", *Economics and Philosophy* 3 (1987), pp. 215-244.

[30] R. J. Arneson, "Equality and Equal Opportunity for Welfare", *Philosophical Studies* 56 (1989), p. 85. See also the remark on p. 82, about common sense and current practices not favouring resource egalitarianism.

[31] G. A. Cohen, "On the Currency of Egalitarian Justice", *Ethics* 99 (1989).

[32] *Ibid.*, section IV.

the debate about the appropriate *equalisandum* is really about. I shall argue in the next section that it has sometimes confused two separate issues.

4. Two separate issues

It is not obvious why the debate about what egalitarians should seek to equalise should be thought to be a specifically egalitarian concern. A first pass at formulating the issue which, one might expect, that debate should be addressing, is this: which aspects of persons' circumstances should an account of justice treat as advantages or disadvantages? That question is pressed on us by worries about basing theories of justice on theories of the good life. We want to know what to treat as an advantage, insofar as we're interested in justice, given that we are reticent to have justice depend on a well-developed conception of the good. These worries are not distinctively egalitarian.[33]

It's true that the debate makes sense only if we assume that what justice requires in part depends on the correct distribution, by some agency, of a stock of goods to which people do not already have binding entitlements. That might be thought to be an egalitarian assumption, if we were to associate egalitarianism with the view that people do not have pattern-independent entitlements to goods, and anti-egalitarianism with denial of that view.[34] But that association is unwarranted, since both historical entitlement egalitarianism, and patterned non-egalitarianism, are established conceptions of justice.[35]

---

[33] Note also that they may be shared by perfectionists, insofar as they suppose there are a number of incommensurable values.

[34] Nozick claims that one of the things which makes Rawlsian egalitarianism plausible, is the false assumption that holdings come into the world like manna from heaven. See R. Nozick, *Anarchy, State, and Utopia*, p. 198.

[35] An example of the former is the account of justice developed in H. Steiner, *An Essay on Rights* (Oxford: Blackwell, 1994). The version of the latter kind of view receiving most current discussion is

Without the association, it is difficult to see why concern with the definition of a conception of advantage, essential to any patterned conception of distributive justice, should be restricted to those who want to *equalise* advantages and disadvantages amongst persons.

There might be any number of sociological explanations for the issue about advantage having been treated, for the most part, as if it were a distinctively egalitarian issue. But there is one explanation having to do with a confusion internal to the debate. The confusion consists of conflating the question about personal advantage with a broader question, which might be formulated as follows: what would be equalised under a just distribution of goods? The latter question asks for more than a conception of personal advantage appropriate to thinking about justice. It asks, in effect, for a complete conception of distributive justice, on the supposition that *something* would be equalised amongst persons were such a conception realised. That *is* a distinctively egalitarian supposition. In order to answer the question, we need to know not just what counts as an advantage with regard to justice, but also what distributions of advantages and disadvantages are required or permitted by justice.

The confusion of these two questions can be detected in some discussions of the significance of our considered judgements in particular cases for our views about personal advantage. Arrow's objection to the use of an index of primary goods as a metric of advantage (and, possibly, Rawls's response to that objection) is a case in point.[36] Arrow complains that the index of primary goods is an unsatisfactory measure of

---

Prioritarianism, for which see D. Parfit, "Equality or Priority?", *The Lindley Lecture*, University of Kansas (1995). But non-egalitarian, non-prioritarian patterned views are quite conceivable.

[36] See J. Rawls, "Social Unity and Primary Goods", in A. Sen and B. Williams (eds.), *Utilitarianism and beyond* (Cambridge: Cambridge University Press, 1982), pp. 167-170.

personal advantage, since it fails to take into account the relative costs of satisfying different preferences. We'll come back to this in Chapter Three. The present point is that Arrow appeals to a judgement about which distributions of advantages and disadvantages justice requires or permits, and not directly a judgement about what counts as an advantage. Thus, to think that without further argument it poses a threat to the use of primary goods as a metric of advantage, is to confuse the two issues I have distinguished.[37]

We must approach Cohen's criticism of Dworkin bearing this distinction in mind. Both Cohen and Dworkin spend much of the time discussing the second issue – what would be equalised under a just distribution of goods – though both also discuss the first. Offhand, it seems likely that issues of responsibility will be more pertinent to the second issue than to the first. The first issue, about the metric of advantage, seems more likely to turn on considerations of liberal neutrality, on one hand, and the availability of information, on the other.[38] In any case, in the rest of this chapter we will examine the argument for the responsibility-tracking view keeping this distinction in mind.

---

[37] This confusion does not afflict all contributions to the dispute, but it is a persistent tendency. Arneson clearly distinguishes the issues, in R. J. Arneson, "Equality", in R. E. Goodin and P. Pettit (eds.), *A Companion to Contemporary Political Philosophy* (Oxford: Basil Blackwell, 1993), p. 494. Cohen himself makes the distinction on pp. 920-921 of "Currency".

[38] Part of Dworkin's criticism of welfare egalitarianism has to do with alleged conceptual difficulties in measuring equality of welfare in terms of anything other than a theory of fair shares such as his resource egalitarianism contains. See Dworkin, "What is Equality? Part 1: Equality of Welfare", sections IV and V. Griffin, meanwhile, notes that governments typically do not have information about individual utility functions – so that utilitarians would favour a principle of equal distribution of "all-purpose means" in the first instance. See J. Griffin, *Well-Being* (Oxford: Oxford University Press, 1986), p. 299. Rawls's later discussions of primary goods makes clear that his idea of using an index of primary goods as a measure of advantage, is tailored to a very specific concern, namely the concern to develop an account of justice which is 'political' in Rawls's special sense. (For that reason, we should perhaps be careful about presenting the Rawlsian account of primary goods as a resourcist stance in the debate about the appropriate *equalisandum*.) See J. Rawls, "Social Unity and Primary Goods", pp. 160-161.

5. Fairness and Responsibility

Several cases are of interest to us. They each combine, in different ways, considerations of voluntariness with different reasons for which persons may require more or less than the average amount of resources to reach a given level of welfare, or to be capable of certain 'functionings'.[39] We've already mentioned Louis, for example, who has an average amount of resources, but requires more than average to reach a given level of welfare, because he has voluntarily cultivated an expensive taste. Dworkin and Cohen both agree that Louis should not receive extra resources to redress his welfare deficiency, though for different reasons. According to Dworkin, Louis should not receive more because he already has his *fair share* of resources, according to the envy test.[40] According to Cohen, Louis should not receive extra because he is *responsible* for his welfare deficiency, since he voluntarily cultivated the expensive taste which generates it.[41]

Dworkin's resource egalitarianism is an example of a naturalist conception of justice. To the extent that it finds a connection between the concepts of justice and responsibility, it holds that the theory of fair shares explains what justice requires, which in turn explains which costs it is legitimate to leave people to bear (hence, which costs persons are to be held responsible for, in the naturalised sense). Dworkin favours explaining the idea of fair shares in terms of an hypothetical auction of those resources up

---

[39] In a series of publications, Sen has developed the idea of capability to function as a measure of personal advantage. See, as a selection: A. Sen, "Equality of What?", in S.M. McMurrin (ed.), *Liberty, Equality, and Law* (Salt Lake City: University of Utah Press, 1987); A. Sen, "Capability and Well-Being", in M. Nussbaum and A. Sen (eds.), *The Quality of Life* (Oxford: Clarendon Press, 1993); A. Sen, *Inequality Reexamined* (Oxford: Clarendon Press, 1992). It should be noted that Sen thinks that the relevance of his capability to function approach is not restricted to discussions of equality: see "Capability and Well-Being", pp. 49-51.

[40] Dworkin, "What is Equality? Part 1: Equality of Welfare", pp. 228-240.

[41] Cohen, "Currency", pp. 922-924.

for distribution, such that individuals determine the division of the various lots, and bid in accordance with their tastes and ambitions using equal purchasing power (clamshells). The results of such an auction would satisfy the envy test, since, "no one will envy another's set of purchases because, by hypothesis, he could have purchased that bundle with his clamshells instead of his own bundle."[42]

Cohen, by contrast, has a two-stage approach to determining what equality in distribution requires. First, we determine the pattern of advantages and disadvantages in the relevant population, where the concept of advantage is a pluralist one, containing elements of welfare, resources, and capability to function. Second, we consider as candidates for redress those disadvantages (or advantages) which are not their bearers' responsibility, in the agency-implicating sense that they did not result from choice.[43] Cohen's account of justice explains what justice requires, then, in terms of a non-naturalist concept of responsibility.[44]

Our task is to decide whether Cohen's two-stage theory provides a better explanation of the correct judgements about whether to compensate or not in these various cases, than does Dworkin's envy-test-based view. The case of Louis is not helpful

---

[42] Dworkin, "What is Equality? Part 2: Equality of Resources", *Philosophy and Public Affairs* 10 (1981), pp. 283-290. The quotation is from p. 287.

[43] The 'two stages' referred to here do not correlate with the stages in one kind of responsibility-tracking view I discussed in section 1. In fact it is possible to interpret Cohen's view either as a two-stage view of the kind described there, or as what I called a 'direct' view.

For the 'two-stage' interpretation, read 'advantages or disadvantages' as 'advantages or disadvantages relative to other persons'. Then Cohen's view is that *in*equalities are candidates for rectification if they are not their bearers' responsibility – which privileges equalities, in the fashion of two-stage views.

For the 'direct' interpretation, read 'advantages or disadvantages' in a way which does not compare different persons' lots, but rather the lot of each person under different arrangements. Then Cohen's view is that advantage-producing or disadvantage-producing (in this non-interpersonally comparative sense) holdings are ripe for redistribution if they are not their bearers' responsibility. This does not privilege equalities, which also must reflect responsibility to be justified. This is, in fact, a direct responsibility-tracking view, which builds-in responsibility to the specified ideal pattern. For Cohen, that pattern is *equal access to advantage*. See Cohen, "Currency", p. 916.

in that respect since, as we've seen, Dworkin's and Cohen's views both account for the judgement, which I take to be correct, that we should not compensate Louis for his expensive taste. We can't yet discriminate between the different reasons they give for not compensating without begging the question one way or the other.

Consider next, then, the case of Paul.[45] Paul has an involuntarily acquired expensive taste. He prefers photography to fishing, but the cost of photography in his society is higher than the cost of fishing and other averagely expensive pursuits. Cohen thinks we should compensate Paul. His involuntary expensive taste is like a handicap, since it makes him unable to reach the average level of welfare for reasons of simple bad luck on his part, and so constitutes a disadvantage for which he is not responsible.[46] Dworkin's resource egalitarianism treats disadvantages due to lack of resources as candidates for compensation, but does not treat disadvantages due to expensive tastes as candidates for compensation, even if they are tastes which their bearer is just stuck with. Cohen here is putting pressure on Dworkin's distinction between handicaps on one hand, and preferences or tastes, on the other. Both make persons inefficient converters of resources into welfare.

Dworkin does treat some tastes like handicaps. He distinguishes between those tastes or ambitions which their bearer identifies with, and those which she repudiates.[47] He has two main arguments for distinguishing cases in this way. First is the argument that only in the cases of non-identification can we say what it is for persons to be unequal in resources because of expensive tastes, without collapsing into welfarism. He writes:

---

[44] Cohen, "Currency", pp. 920-924.
[45] *Ibid.*, pp. 925-926.
[46] *Ibid.*, p. 925.
[47] Dworkin, "What is Equality? Part 2: Equality of Resources", pp. 301-304.

> We cannot say that the person whose tastes are expensive. for whatever reason, therefore has fewer resources at his command. For we cannot state (without falling back on some version of equality of welfare) what equality in the distribution of tastes and preferences would be. Why is there less equality of resources when someone has an eccentric taste that makes goods cheaper for others, than when he shares a popular taste and so makes goods more expensive for them?[48]

The thought here seems to be that there is only one criterion for equality in tastes, which is the prospect of satisfying them given a certain distribution of resources. That's a welfarist criterion, and as such cannot be admitted as a modification to resourcism, but is rather a contradiction of it. A repudiated taste, on the other hand, is a special case, allowing of treatment using a hypothetical insurance market along the lines of the method Dworkin proposes for dealing with inequalities in personal resources such as talents and handicaps, because "we can imagine people who have such a craving not having it, without thereby imagining them to have a different conception of what they want from life than what in fact they do want."[49]

The second argument is that (i) the distinction between personality and circumstances, which Cohen supposes cases of involuntary expensive tastes to pose a problem for, treats involuntary (and voluntary) non-repudiated tastes unambiguously as part of personality, and involuntary (and voluntary) repudiated tastes unambiguously as

[48] Dworkin, "Equality of Resources", p. 302.
[49] *Ibid.*, p. 303.

part of circumstances, on the basis of a first-person view of these matters; and (ii) the first-person view is the correct view of these matters.[50] I cannot view my non-repudiated tastes, whether voluntary or involuntary, coherently as impediments to pursuit of my goals or plans, such as handicaps are, since they help to define what my goals or plans *are*.

The first part of this second argument seems straightforward enough. But why is the first-person view the correct one for this issue? Why isn't the issue whether some third party would count my non-repudiated expensive taste as an impediment to my having a successful life? Dworkin suggests some apparently pragmatic reasons – officials would get it wrong, "no one could apply for compensation with a straight face"[51] – and one principled reason. The principled reason is that the distinction between personality and circumstances should be made within the correct ethical view, which implies that "no one could identify the resources he ought to have to compensate for mistaken beliefs he could not think mistaken".[52] Dworkin's statement here is compressed in the extreme, and seems to presuppose that the first-person view is correct in stating what purports to be a reason for thinking that it is.

I find both Dworkin's first and second arguments, especially the second argument's second part, difficult to assess. There certainly is a difference between inefficient resource utilisation which is the result of a handicap, and inefficiency due to

---

[50] See R. Dworkin, "Foundations of Liberal Equality", in S. Darwall (ed.), *Equal Freedom. Selected Tanner Lectures on Human Values* (Ann Arbor: University of Michigan, 1995), pp. 293-297.

[51] *Ibid.*, p. 297. Cf. T. M. Scanlon, "Equality of Resources and Equality of Welfare: A Forced Marriage?", *Ethics* 97 (1986), pp. 116-117, on compensation for expensive religious convictions. I say that Dworkin's initial reasons are *apparently* pragmatic, because it's not clear to me whether the point about not being able to apply for compensation with a straight face, might not be given some principled interpretation. It looks similar to a charge which Cohen himself makes against rich talented Rawlsians invoking an argument from incentives to justify to the poor their high salaries. I'll discuss Cohen's views on this fully in Chapter Four.

expensive tastes, which does not have to do with personal responsibility for that inefficiency. Cohen claims, contrary to this, that Dworkin must eventually invoke considerations of personal responsibility in order to explain his (Dworkin's) stance on expensive tastes.[53] It is at least clear that this claim of Cohen's is false. Dworkin can explain his stance by reference to the incoherence, from the first person point of view, of treating an non-repudiated taste as an obstacle to one's plans and goals. What is unclear is whether the explanation stops there, or whether the first person stance can be shown to be the appropriate one in these matters.

Suppose Dworkin is right, and the repudiation criterion establishes a morally significant difference between handicaps and non-repudiated expensive tastes. The example of Paul is then a challenge to Dworkin's resource egalitarianism only if we judge that it is right to compensate for *non-repudiated* involuntary expensive tastes. These are cases in which someone has a taste which is expensive to satisfy, which they are not responsible for, in an agency-implicating sense, but which they do not view as an obstacle to the pursuit of their plans, but rather as partly definitive of their plans. Paul just happens to like photography, but he identifies with that taste, judging a life in the darkroom to be *ceteris paribus* better than a life outside it. Does justice require that he be given extra impersonal resources, to compensate for the expense of his taste? My intuitions are divided here. They provide no firm answer to the question, such as may help us decide whether the best explanation of our considered judgements shows that justice tracks responsibility. They invite further reflection on the question of whether it is involuntariness or lack of identification which triggers egalitarian concern – but that's

---

[52] Dworkin, "Foundations of Liberal Equality", p. 297.
[53] Cohen, "Currency", pp. 925-927.

just the question we're trying to answer. So if the repudiation criterion is allowed, the case of Paul does not advance the argument for the responsibility-tracking view which we're considering.

Suppose we fail to find the first person perspective commanding, and as a result we reject the repudiation criterion. Then Paul's case is more straightforward. For without that criterion, treating involuntary expensive tastes as morally in a wholly distinct category from handicaps looks vulnerable to the charge that handicaps are being conceived in a restricted way. They should not be conceived as necessarily corporeal, having to do with functions of the body only, but as possibly mental. Once that point is made, we require a new reason for not allowing preference functions possibly to be handicaps. If such a reason is not forthcoming, we should conclude that some cases like Paul's, at least, are worthy of compensation. This is to affirm Cohen's judgement in these cases, and so to offer some support for the argument from the best explanation of our considered judgements.

So far we've found some arguable support for the responsibility-tracking view, in cases of involuntary expensive tastes. Consider now a final type of case.[54] Jude begins with cheap tastes, then cultivates a single expensive taste (for bullfighting, say) which, when added to his other tastes, still leaves his preferences as a whole cheaper than average to satisfy. What should he receive?[55]

Both Dworkin and Cohen discuss the case of Jude only very briefly. Dworkin discusses it in criticism of welfare egalitarianism, and Cohen discusses it in criticism of

[54] No doubt cases could be multiplied indefinitely, but the one I am about to discuss is, amongst the ones I know of, the most relevant to the assessment of the argument for the responsibility-tracking view which we're considering.

[55] See Dworkin, "Equality of Welfare", pp. 239-240; Cohen, "Currency", p. 925.

both opportunity for welfare egalitarianism and Dworkin's resource egalitarianism. Dworkin claims, for his part, that Jude, having cultivated his single expensive taste, should receive more than he had under welfare egalitarianism with only cheap tastes. That's supposed to pose a puzzle for our account of equality, since, it was assumed, Louis' cultivation of a new expensive taste is not ground for compensation, whereas Jude's is. Both new tastes are deliberately cultivated. Hence, Dworkin concludes, it is the *amount* of resources each character has before cultivating a new taste which is decisive. Louis already had his fair share (defined according to the envy test) and so could not justifiably ask for more, whereas Jude had less than his fair share under the equal welfare scheme, and so can justifiably ask for more. Welfare egalitarianism is wrong – and, just as important for Dworkin, what explains our considered judgements is the theory of fair shares.

Cohen has a different view about what Jude should receive, and draws a different moral. He thinks that Jude should receive more than he would under welfare egalitarianism, but less than he would (less than equal resources) under Dworkin's scheme. This judgement is a straightforward consequence of Cohen's pluralistic conception of advantage, which gives some (unspecified) weight to both resources and welfare.

Cohen's view is in another way ambiguous, however. Consider what Cohen should say about what *Initial Jude* should receive as compared with *Later Jude*. Initial Jude has a set of involuntary cheap tastes, and as such should, on Cohen's view, receive somewhat less than average impersonal resources (those with involuntary expensive tastes receive somewhat more). Later he cultivates a single expensive taste – a change

which, on Cohen's view, does not merit any compensation, since it is something which Jude is responsible for. So Cohen's views commit him to treating Initial Jude and Later Jude in the same way. But his discussion encourages, or at least permits, the opposite interpretation:

> A believer in equality of opportunity for welfare has to keep Jude poor, since he did not have to become a bullfight-lover . . . A believer in Dworkin-style equality of resources ignores Jude's tastes, and their history, and finds no reason . . . to grant him less income than anyone else. I reject both views. Pace equality of opportunity for welfare, I see no manifest injustice in Jude's *getting the funds* he needs to travel to Spain. He then *still* has fewer resources than others, and only the same welfare, so equality of access to advantage cannot say, on that basis, that he is overpaid. But, pace equality of resources, it seems not unreasonable to expect Jude to accept some deduction from the normal resource stipend because of his fortunate high ability to get welfare out of resources.[56]

This passage seems to suggest that Cohen thinks Jude should get more once he has acquired his expensive taste for bullfighting – that's what talk of Jude's "getting the funds . . . [after which he] still has fewer resources than others" conveys to me. On this interpretation, Later Jude should get more than Initial Jude. But that's inconsistent with Cohen's official views.

---

[56] Cohen, "Currency", p. 925, emphasis added.

Let us set aside the exegetical point and suppose that Cohen treats Initial Jude and Later Jude in the same way. Does this case support the responsibility-tracking view? Once again, I find my intuitions genuinely divided, between assent to Cohen's judgement that Jude should get less than the average stipend due to his fortunate constitution, and assent to Dworkin's judgement that Jude's luck in this respect is not relevant to what justice requires – so that Jude should get an equal amount of impersonal resources. Dworkinians will claim that Cohen is unduly harsh on people with involuntary cheap tastes (and that he is unduly indulgent to those with expensive tastes). Cohen, on the other hand, claims that Dworkinians are unduly harsh on those with involuntary expensive tastes, and unduly indulgent to those with cheap tastes. Once again, this uncertainty about the case for compensation invites further reflection about whether it is voluntariness which triggers egalitarian concern in cases like these. But, once again, that's the question we set out to answer, so the case of Jude does not help in that respect.

The conclusion of this section is an unexciting one. The argument from the best explanation of our considered judgements across a range of cases is not decisive. Cohen's responsibility-tracking account of justice succeeds in explaining those of our considered judgements which seem sure. But it is not the only possible, nor clearly the best, explanation of these judgements. Dworkin's naturalist approach seems also to account for them, by appealing to the concept of fair shares and the idea of repudiation of tastes instead of appealing to any overtly agency-implicating concepts. And the cases which seem most likely to divide Cohen and Dworkin – those of Paul and Jude – are cases in which our judgements are insufficiently sure for the argument to the best explanation to

get going. In the end, that argument fails because it does not have enough data to work with.

6. Conclusion

The naturalist strategy escapes unscathed from the arguments we've canvassed in this chapter. We noted theory-building reasons in favour of that strategy, which have to do with the philosophical and other problems surrounding agency-implicating concepts, such as responsibility in its ordinary sense. These reasons have a certain force, though they should not be treated as capable of justifying naturalism in the final instance.

Beginning in section 2, we turned our attention to arguments in favour of the responsibility-tracking view. In total, we rejected four arguments in favour of that view. The 'argument' by inspection of the concept of justice is hopelessly inarticulate, not adding up to a real argument at all. The broadly Kantian and Millian arguments, on the other hand, succeed at most in establishing a conceptual relationship between justice and responsibility, failing to establish a direction of dependence between the two concepts. The argument from the best explanation of our considered judgements in a range of cases looked more promising, but ultimately foundered for two reasons. In those cases where our intuitions seem sure, Dworkin's naturalist approach explains our judgements just as well as Cohen's responsibility-tracking approach. The cases which promised to be decisive turned out to suffer from uncertainty about our intuitions, and demanded further enquiry into precisely the question we hoped they would answer.

I believe that the naturalist strategy can't succeed. But we must look further for arguments to show that it cannot. In Chapter Three, I'll examine Rawls's naturalism about desert (and responsibility), in the hope that it will provide some clues about the viability of naturalism.

Chapter Three

Rawls and the Naturalising Strategy

Chapter Two examined the relationship between justice and responsibility. There I noted that many people find it intuitively plausible that justice tracks responsibility. I noted also that it is possible to reverse this order of ideas, and explain responsibility in terms of justice or fairness. According to this *naturalisation* of responsibility, what someone is responsible for depends on which costs it is just to let her bear. The naturalising strategy is interesting because it promises to protect theories of justice from the problems associated with our understanding of agency.

In Chapter Two we found no strong argument for the dependence of justice on an agency-implicating concept of responsibility, and so no reason to think that the naturalising strategy couldn't succeed. In this chapter we'll examine the relationship between justice and *desert* in Rawls's theory of justice. That theory is naturalist at many points, though it seems also to rely on non-naturalised judgements of desert and responsibility. In examining it, our aim is to test the naturalising strategy in a slightly different context, and to improve our understanding of the problems and possibilities which it involves. I shall argue that Rawls's theory of justice is incompletely naturalist,

but we will find, in this chapter, no reason to think that its defects in this regard could not be remedied.

1. The Concept of Desert and the Naturalising Strategy in Rawls

At least in its normal interpretations, the concept of desert, like the concept of responsibility, implicates our understanding of agency. If someone deserves something, in the standard sense, the reason is usually that she did something or other, when she *could have done* something else. Common sense morality holds that what justice requires depends on what people deserve in this agency-implicating sense.[1] For example, it distinguishes between those who achieve a lot through working hard, and those to whom success 'comes easily', and the distinction seems to rest on the idea that different people have different counterfactual possibilities within their reach. Since *desert* is commonly thought to be conceptually linked with *justice*, and since it is agency-implicating in its normal interpretation, the naturalising strategy may seem attractive with respect to desert as it did with respect to responsibility.

Rawls rejects the view that justice depends on desert. And according to one leading interpretation of his remarks on desert, Rawls not only rejects this view, but seeks also to explain what people deserve in terms of what justice requires.[2] If so, then his remarks on desert are an important example of the naturalising strategy at work. Our aim

[1] David Miller has claimed that desert can provide *pro tanto* reasons of justice. See his *Social Justice* (Oxford: Clarendon Press, 1976), pp. 114-117.

[2] J. Rawls, *A Theory of Justice* (Oxford: Oxford University Press, 1972), sections 12, 13, 17, and 48. I shall refer to this book simply as *TJ*. Scanlon interprets Rawls's remarks on desert as exemplifying what I've called the naturalising strategy. His interpretation emphasises Rawls's distinction between moral desert and legitimate expectations, and treats the idea of legitimate expectations as a revisionist account of the true bases of desert. For Rawls's distinction, see *TJ*, section 48. I'll discuss Scanlon's interpretation fully in section 4 below.

in examining them is to see whether they provide evidence of the viability of the naturalising strategy. His remarks are not perfectly clear, and they support several interpretations. In considering these various interpretations, we shall unearth some additional complexity in possible naturalising views.

There is one argument which threatens to cut short our enquiry, by showing that what justice requires could not depend on what people deserve, regardless of Rawls's remarks on desert. This argument does not depend on worries about the agency-implicating features of the concept of desert. It claims instead that responding to someone's desert out of duty, tends to undermine the practice of responding appropriately to desert. Certain reasons are just extraneous to certain ways of responding to persons. The fact that my admiring someone would make him happy, for example, is extraneous to the practice of admiring persons appropriately. Similarly, duties are extraneous to the practice of responding appropriately to someone's desert. Hence, claims of desert cannot give rise to claims of justice, which entail duties.[3]

It is surely right to think that reasons may be extraneous to practices, and hence practically incompatible with them. It may be true also that duties are extraneous to the practice of responding appropriately to someone's desert. It may nevertheless be the case that desert claims can give rise to claims of justice, entailing duties. Suppose that we are interested in justifying a certain person *A*'s ownership of a holding *P*. The argument we've been considering seems to assume that, if claims of desert were to be capable of justifying holdings, it would have to be the case that we could derive a claim of the form,

(1) *A* has ownership of *P*

from claims of the form,

(2) *B*, *C*, *D*, . . . have duties to treat *A* in certain ways, for reasons of *A*'s desert.

The argument is then that claims of form (2) are false, because of the practical incompatibility of duties on one hand and appropriate response to desert on the other.

Maybe, though, the relationship between desert, ownership, and duties, is like this:

(3) *A* has ownership of *P*, for reasons of *A*'s desert,

therefore

(4) *B*, *C*, *D*, . . . have duties to treat *A* in certain ways.

That is, the duties of others to treat a person in certain ways (not stealing his property, for example), may be explained by the fact that his desert gives him ownership of a holding – rather than the ownership of the holding being explained by the duties of others.

If so, then claims of justice may be explained by claims of desert, whilst it is still the case that the duties which the claims of justice entail are not duties to respond appropriately to desert. The duties may be duties not to steal a person's property, and so on. Desert may establish a relation of ownership between a person and a holding, from

[3] This argument is made by James Griffin. See J. Griffin, *Well-Being* (Oxford: Clarendon Press,

which relations (duties) between others are derived, rather than it being the case that desert establishes relations (duties) between persons first, from which we may derive ownership. Hence the argument which threatened to cut short our enquiry is inconclusive.

Of course, the issues which arise with respect to desert are not exactly the same as those which arise with respect to responsibility. The idea that certain possibilities are, and others aren't, within a person's reach, is essential to the concept of responsibility as I've been treating it. That idea is very important for desert too, as I've said, but in some cases it may not be essential to it. The bases or grounds of desert may extend wider than the bases of responsibility, to include such potentially non-agency-implicating facts as someone's having run one hundred metres faster than everyone else. In other words, the explanation of someone's deserving something may be that she has certain qualities – not that certain possibilities are or were within her reach.[4] Nevertheless, it turns out that Rawls's complaints about desert hinge on its agency-implicating features. So in most cases it is justified to treat his remarks on desert as extending to the concept of responsibility.

However, there is one respect in which Rawls's views on desert are a special case for those interested in the naturalising strategy. Rawls's *contractualism* makes more significant the fact that agency-implicating concepts are typically the subject of persistent disagreement in their application. I noted in Chapter Two that there are instrumentalist reasons against explaining justice in terms of responsibility. One such instrumentalist reason has to do with the fact that people don't agree in their judgements of

---

1986), pp. 257-264.

[4] Joel Feinberg emphasises the heterogeneity of bases of desert, in "Justice and Personal Desert", in J. Feinberg, *Doing and Deserving. Essays in the Theory of Responsibility* (Princeton, NJ: Princeton University Press, 1970).

responsibility. It is polemically unwise to base one's account on judgements about which there is persistent disagreement.

That consideration has some force insofar as we take a theory-building stance towards these issues. We adopt that stance when it is too soon to say whether some general theoretical approach will be successful. But we should not mistake this or any other instrumentalist reason for a reason to believe that justice does not track responsibility (it is, rather, a reason against trying to construct a theory of justice as responsibility-tracking). Justice may track responsibility, for all we know – and if it does, so much the worse for our theories of justice.

But the parallel point about controversy over desert has more significance for Rawlsians. Their contractualism commits them to finding a conception of justice which holders of any reasonable comprehensive conception of the good may accept.[5] An acceptable conception of justice, on this view, will be expressed in terms of "certain fundamental intuitive ideas viewed as latent in the public political culture of a democratic society".[6] To the extent that there is persistent disagreement in judgements of desert in our society, we may suppose that explanations of justice in terms of desert could not be given a Rawlsian justification. Reliance on judgements of desert is not just polemically unwise, for Rawlsians, it is at odds with the official view of the justification of conceptions of justice. For Rawlsians, reasons of internal coherence, not just instrumental reasons, attach to the avoidance of reliance on judgements of desert.

---

[5] Rawls distinguishes between 'comprehensive' and 'political' conceptions of the good. A conception is comprehensive if it contains ideas about non-political virtues or the value of non-political activities. Political conceptions, by contrast, describe only political virtues and values, and do not depend for their justification on any particular comprehensive doctrine. See J. Rawls, *Political Liberalism* (New York: Columbia University Press, 1993), Lectures I and II; J. Rawls, "The Priority of Right and Ideas of the Good", *Philosophy and Public Affairs* 17 (1988), pp. 252-253.

[6] Rawls, "The Priority of Right and Ideas of the Good", p. 252.

The fundamental naturalising method, as we saw in Chapter Two, is to explain otherwise troubling concepts in terms of justice or fairness. One can't dismiss the naturalising strategy in the case of desert, any more than one can dismiss it in the case of responsibility. It is unsatisfactory to complain simply that the concept of fairness or the concept of justice is 'too thin' to support an explanation of desert or responsibility, as some may be tempted to do. Conceptual thinness is not something we can see by inspection. Instead, we'll look in detail at the way in which Rawls seems to use the fundamental naturalising method in his remarks on desert. Our aim is to discover two things. First, does Rawls successfully carry through the naturalising strategy in his account of justice? And second, if he does not, does that tell us anything about the viability *in general* of that strategy?

2. Rawls's two doctrines, and two problems of interpretation

Rawls makes two broad claims about desert in *TJ*. The first is that reflection on desert favours one particular interpretation of his second principle of justice. In its initial formulation, the second principle claims that inequalities must be attached to positions equally open to all, and must be to everyone's advantage, in order to be justified.[7] Two parts of this formulation need further spelling-out: the idea of equal openness of positions, and the idea of everyone's advantage. Rawls treats these as independent ideas, and considers two interpretations of each. Hence he considers four possible

[7] *TJ*, p. 60.

interpretations of the second principle as a whole. These are represented in the following table.[8]

	'Everyone's advantage'	
'Equally open'	Pareto Principle	Difference Principle
Equality as careers open to talents	System of Natural Liberty	Natural Aristocracy
Equality as equality of fair opportunity	Liberal Equality	Democratic Equality

Table 1: Four interpretations of Rawls's second principle of justice

Rawls thinks that reflecting on desert leads to either or both of the following progressions of ideas. First, we find the judgements of desert implicit within the system of natural liberty to be incoherent,[9] and seek to rectify the situation by moving either to liberal equality or to natural aristocracy. Further reflection shows, however, that either of these modifications involves treating morally equivalent facts non-equivalently. Both the

[8] The table is adapted from the one on p. 65 of *TJ*. Note that Rawls refers to the Pareto principle as 'the principle of efficiency'.

[9] On an alternative interpretation, Rawls's claim is that the judgements are *false*, not incoherent. I shall pass over the exegetical issues, and simply adopt the interpretation which I suggest in the text. Our objective is to examine naturalism about desert in its most plausible form, and, I think, attributing to Rawls the charge of incoherence best serves this aim. That's because, if Rawls's charge is that the

system of liberal equality and the system of natural aristocracy make sense only if we treat the luck that persons have in their natural talents, and the luck they have in their social background, non-equivalently with respect to desert. Reflection shows that these kinds of luck are morally equivalent. Hence, " . . . however we move away from the system of natural liberty, we cannot be satisfied short of the democratic conception."[10] Of the alternatives to natural liberty, democratic equality is the only conception which treats the two kinds of luck equivalently, and so it is favoured by the reflections on desert.[11]

Rawls's second broad claim about desert is that it is crucial not to mistake it for a different idea. We may gain 'legitimate expectations', which are entitlements, by acting in ways which are encouraged by some reasonably just arrangements or institutions. Such expectations ought to be honoured, as a matter of justice,

> But what [persons] are entitled to is not proportional to nor dependent upon their intrinsic worth. The principles of justice that regulate the basic structure . . . do not mention moral desert, and there is no tendency for distributive shares to correspond to it.[12]

---

judgements are false, he is himself committed (in an obvious way) to substantive judgements of desert in its ordinary, agency-implicating sense.

[10] *TJ*, p. 75.

[11] Rawls insists that his remarks on desert are not to be considered an *argument* for democratic equality, but are only meant to 'prepare the way' for it, since they do not employ the device of the original position. (*TJ*, p. 75, see also p. 66.) I shall explain in the text below the extent to which I think the reflections on desert do add up to an argument for democratic equality. If they do not, however, that can't be because they don't proceed from the original position. Throughout *TJ* Rawls appeals to the idea of fit between the principles of justice and our considered judgements, and underwrites this kind of appeal with his discussion of reflective equilibrium. (Note that Brian Barry reckons the argument from desert for democratic equality to be stronger than the official argument from the original position. See B. Barry, *Theories of Justice* [Hemel Hempstead: Harvester-Wheatsheaf, 1989], pp. 213-225, and see section 3 below.)

[12] *TJ*, p. 311, see also p. 103.

If in a just society I play and win a lottery, I am entitled to the winnings whether or not I morally deserve them. My entitlement does not depend on my moral desert, or moral responsibility, or any other agency-implicating concept. It depends only on my having actually behaved in certain ways against a certain background of rules and arrangements. The entitlements generated by the rules are, for that reason, in no danger of being disrupted by variable moral desert amongst those who comply with the rules. For the same reason, a person's entitlement on grounds of legitimate expectations implies nothing about her moral desert. Rawls's second claim is that we should not confuse legitimate expectations with moral desert, and that, whilst justice honours the former, it has no tendency to correspond to the latter.[13]

The idea of legitimate expectations can, but need not, be used in an attempt to explain the true grounds of desert claims. According to such a view, when we speak of someone 'deserving' something, our claim is to be understood as meaning that she has a legitimate expectation to something. That is to explain desert in terms of justice, since the idea of legitimate expectations makes essential reference to the concept of justice. (Legitimate expectations are generated by behaviour in accordance with reasonably just arrangements.) That would be a way of naturalising desert, and it is this view which Scanlon attributes to Rawls, as we'll see. On the other hand, we could say that the idea of legitimate expectations is simply distinct from, and not a revisionist account of, the idea of desert. I'll return to this point in section 4.

In brief outline, Rawls's remarks on desert consist of these two doctrines. Now let me outline the two main problems in interpreting those remarks. The first is to see exactly how reflection on desert favours democratic equality as a conception of justice. One can

[13] Rawls draws a sharp contrast in this regard between retributive and distributive justice – in the

see how it could be thought to undermine its three rivals on the list that Rawls considers. Each of those conceptions seeks to track desert, so reflection on desert looks like the kind of thing that might upset them. But democratic equality itself seems not to contain any desert-tracking principle, and so it is difficult to see how reflection on desert could be relevant to its justification – except in the weak sense that all its rivals on a particular list are rejected. The question is whether democratic equality is favoured only in this weak sense.

The second interpretative problem is to decide whether natural liberty, liberal equality and natural aristocracy all fail to track desert because tracking desert is impossible in principle, or because it is the wrong thing for a principle of justice to do, or instead for reasons which are specific to those particular conceptions. Each of these possibilities has been canvassed in the literature. In characterising the precise way in which Rawls's remarks manifest the naturalising strategy, it will be important to see which of them is right.

In the next section I'll discuss two influential but unsatisfactory interpretations of Rawls's remarks. It is important to dismiss these interpretations, as they could mislead us later on. Then in section 4 I'll explain two interpretations of Rawls's remarks which seem to me to present the strongest versions of his argument.

3. Two unsatisfactory interpretations

Each of the three rivals to democratic equality, in Rawls's view, tries and fails to make distributive shares reflect moral desert. All three conceptions incorporate some desert-

---

former, he allows, claims of moral desert may be important. See *TJ*, pp. 314-315. In my view the

tracking principle, but all three treat matters of luck as real bases of desert. The system of natural liberty treats matters of luck in "natural talents . . . social circumstances and such chance contingencies as accident and good fortune" all as bases of desert, and wrongly allows distributive shares to be influenced by them.[14] Natural aristocracy and liberal equality each restrict the factors which are allowed to influence distributive shares, in some way, but in both cases the restriction is arbitrary, drawing a false distinction between factors which are all really matters of luck. One does not deserve one's parents or social class, but neither does one deserve one's talents. Even very central aspects of one's character, such as the ability to make an effort, depend on one's natural and social luck.[15] Hence all three conceptions of justice are vitiated by incoherent views about desert. But why exactly does this favour democratic equality?[16]

The problem is that, as we've seen, Rawls claims that justice has 'no tendency' to track desert. If, as this suggests, democratic equality contains no desert-tracking principle, then remarks on desert seem simply irrelevant to its justification. No claim about the real grounds of desert – whether, for example, they include the ability to make an effort – would seem to bear on the justification of a conception of justice which purports to be independent of considerations of desert.

In effect, Brian Barry has suggested one way in which this problem (the first of our interpretative problems) may be solved. His suggestion is that democratic equality *does* contain a desert-tracking principle, the so-called principle of redress.[17] This principle

---

contrast is too stark, but I shall not argue that point here.

[14] *TJ*, p. 72. See below, section 5, for discussion of the extent to which Rawls relies on substantive judgements of desert in making these claims.

[15] *TJ*, p. 74.

[16] I shall not challenge Rawls's view that, for example, a person does not deserve her natural talents. Our interest is not directly with the truth of Rawls's views, but with what they can tell us about the viability of the naturalising strategy.

[17] See B. Barry, *Theories of Justice* (Hemel Hempstead: Harvester-Wheatsheaf, 1989), pp. 213-215.

requires undeserved inequalities to be compensated. If we add to the principle of redress the thought that inequalities of character, ability, and social circumstance are undeserved (as Rawls seems to claim), a strong default preference for equality in distributive shares is established. Democratic equality appears not to be a desert-tracking conception only because it adds to these considerations, some overwhelming considerations about the possible good *effects* of inequalities. The desert-tracking principle of redress is submerged but present in democratic equality. Its presence explains why the reflections on desert favour that conception.[18]

Barry's interpretation has the advantage of showing clearly how the reflections on desert could be said to favour democratic equality. It is true, also, that Rawls says that the difference principle, which is part of the conception of democratic equality, "gives some weight to the considerations singled out by the principle of redress".[19] But there is strong textual evidence against Barry's view. Elsewhere in *TJ*, Rawls states explicitly that his account of distributive justice does not contain a desert-tracking principle, even as "a prima facie principle".[20] What he means when he says that the difference principle gives weight to the considerations singled out by the principle of redress, I suggest, is only that, in placing a check on inequalities, it does part of what the principle of redress requires, though for different reasons.

If Barry's interpretation is wrong, we can settle the first of our interpretative problems. If as I suggest democratic equality does not contain any desert-tracking principle, even as a submerged *prima facie* principle, then it must be the case that it is

---

[18] Susan Hurley points out that correcting undeserved *in*equalities need not amount to tracking desert, since equalities too may be undeserved. See S. L. Hurley, "Justice without Constitutive Luck", in A. Phillips Griffiths (ed.), *Ethics* (Cambridge: Cambridge University Press, 1993), p. 185. For further comment on Barry's reconstructed Rawlsian argument, see G. A. Cohen, "The Pareto Argument for Inequality", *Social Philosophy and Policy* 12 (1995).

[19] *TJ*, p. 100.

favoured by the remarks on desert only in the weak way I indicated earlier. It is favoured insofar as all its rivals on the list Rawls considers fail in their desert-tracking aspirations. But that is only a weak recommendation, since it is obvious, and Rawls accepts, that the list of four conceptions is not exhaustive.

Now consider one of the interpretations which Nozick places on Rawls's remarks.[21] He attributes to Rawls a view about desert which would make it impossible in principle for persons to deserve things. According to this view, which incorporates what has been called the *regression principle*, someone deserves something only if there is some ground *X* of her desert, such that she deserves all the grounds of *X* too.[22] That would explain why the three rivals to democratic equality fail in their aims, of course. They can't track desert, because there's no such thing as desert to track. The regression principle, if it is amongst the necessary conditions of desert, makes desert impossible all by itself – regardless of the truth about determinism, for example. It leads directly to an infinite regress, assuming only that there cannot be a ground which does not itself have grounds.[23]

Attribution to Rawls of the regression principle may be encouraged by some of his remarks about desert, but it is at best an uncharitable interpretation. It may be encouraged, for example, by the following famous passage:

---

[20] *TJ*, p. 310.

[21] Nozick considers a number of ways of construing Rawls's remarks on desert, of which I shall discuss only one. My aim is to criticise not Nozick, but a certain way of understanding Rawls.

[22] See R. Nozick, *Anarchy, State, and Utopia*, p. 225: "It needn't be that the foundations underlying desert are themselves deserved, *all the way down*" (italics in the original). For formulation and discussion of the 'regression principle', see S. L. Hurley, "Justice without Constitutive Luck", pp. 180-185. George Sher also attributes the regression principle to Rawls: see G. Sher, *Desert* (Princeton, NJ: Princeton University Press, 1987), Chapter 2.

[23] This argument is made by Galen Strawson, who accepts the regression principle, and accepts that it makes desert 'in the strong, moral sense' impossible in principle. See G. Strawson, "Consciousness, Free Will, and the Unimportance of Determinism", *Inquiry* 32 (1989), esp. pp. 10-11.

> It seems to be one of the fixed points of our considered judgements that no one deserves his place in the distribution of native endowments, any more than one deserves one's initial starting place in society. The assertion that a man deserves the superior character that enables him to make the effort to cultivate his abilities is equally problematic; for his character depends in large part upon fortunate family and social circumstances for which he can claim no credit.[24]

This passage directs our attention to the *causes* of those things which common sense takes as giving rise to claims of desert: cultivated ability, effort, and character are presented as results or consequences of underlying social or natural luck. It may seem that implicit appeal is being made to the regression principle as incorporated in the conditions of desert, in reaching the conclusion that these common appeals to desert do not survive reflection. For otherwise the mere fact that the putative grounds can be causally explained, might not show that they are not real grounds.

This interpretation is unsatisfactory for two reasons. First, conceptions of desert which incorporate the regression principle are put in doubt by the very fact that they immediately yield the conclusion that desert is impossible. Our normal intuitions that people may deserve things – and not just in the sense that they legitimately expect them – are inconsistent with the regression principle. Those intuitions are not immune to revision, of course; but neither do they count for nothing. The regression principle may seem axiomatic to some. But, perhaps for that reason, there seems to be no real argument

[24] Rawls, *TJ*, p. 104.

in favour of it. It reflects merely a conceptual prejudice, that that is the kind of thing desert *must* be.[25]

We should not accept the regression principle even if we happen to think that the conditions of desert are impossible to satisfy. If it is impossible for persons to deserve things, that should be explained by a substantive thesis, whether metaphysical or physiological, psychological or sociological. It should not fall out directly from what is, in effect, a conceptual stipulation.

Attributing the regression principle to Rawls is, for that reason, uncharitable. There are also textual reasons not to do so. In the passage reproduced two paragraphs back, Rawls claims that seeking to base desert claims on effort is "problematic", because ability to make an effort "depends in large part" on matters of luck. Someone who endorsed the regression principle, we might expect, would couch the rebuttal of appeals to effort in rather stronger terms. Elsewhere, Rawls indicates that he thinks it is possible for persons to be deserving, in the moral sense, when he says that "there seems to be no way to discount for . . . greater good fortune. The idea of rewarding desert is *impracticable*."[26]

Neither Barry's nor Nozick's interpretation is satisfactory. Barry shows how the reflections on desert could be thought to favour democratic equality, but in doing so he wrongly attributes to Rawls a desert-tracking principle. Nozick's interpretation addresses the second of our interpretative problems, claiming that Rawls thinks no-one could ever deserve anything, and hence no desert-tracking principle could help explain what justice requires. But there is no reason to attribute the regression principle, which underlies this

[25] Cf. S. L. Hurley, "Justice without Constitutive Luck", p. 184.
[26] Rawls, *TJ*, p. 312, emphasis added. Cohen cites this passage in criticising Nozick's interpretation of Rawls. See G. A. Cohen, "On the Currency of Egalitarian Justice", *Ethics* 99 (1989), pp. 914-916.

view, to Rawls. In the next section I'll explain the two interpretations of Rawls's remarks which construe them most plausibly.

4. Two persuasive interpretations

The interpretations we have considered so far place most emphasis on the first of Rawls's two doctrines, which is his criticism of the three desert-tracking rivals to democratic equality. Barry sought to reconstruct this criticism, showing, as Rawls failed to do, exactly how it might favour democratic equality in any but the weak sense I've explained. Nozick sought to explain the view about desert which he took to lie behind Rawls's criticism of desert-tracking principles – with the aim of criticising that underlying view and rebutting the criticism. T. M. Scanlon's interpretation of Rawls's remarks, in contrast, places the greatest emphasis on Rawls's second doctrine about desert. That doctrine is about the importance of distinguishing the idea of moral desert from the idea of legitimate expectations. In Scanlon's view, that distinction explains Rawls's rejection of the three rivals to democratic equality. Hence the second doctrine explains the first.[27]

We should recognise Scanlon's version of Rawls's argument as a clear example of the naturalising strategy. I noted earlier (section 2) that Rawls's idea of legitimate expectations may, but need not, be employed in an attempt to naturalise the concept of desert. Such an attempt says something like this: the true grounds of desert are determined by the rules governing reasonably just arrangements. Someone deserves

[27] See T. M. Scanlon, "The Significance of Choice", in S. Darwall (ed.), *Equal Freedom. Selected Tanner Lectures on Human Values* (Ann Arbor: University of Michigan Press, 1995), pp. 73-78.

something, on this view, in those cases where she has entitlements generated by legitimate expectations.[28]

The idea of legitimate expectations may be used, however, not as a revisionist account of the grounds of desert, but as a concept which is parallel to desert and, perhaps, somewhat like it. Persons may have an entitlement-generating legitimate expectation of something, but deserve something quite different. As Feinberg notes, this is the way we often think of the results of sporting competitions: the winner did not deserve to win, but in winning fairly had a legitimate expectation of the prize.[29] If the idea of legitimate expectations is treated in this way as parallel to desert, it involves no claim about the true grounds of desert, and nor does it involve scepticism about desert (though it may, of course, be combined with an independently-motivated scepticism about desert).

It matters a great deal in which of these two ways the idea of legitimate expectations is understood. If it is understood as a revisionist account of the true grounds of desert claims, then we can see why, as Scanlon suggests, it could explain Rawls's criticism of desert-tracking principles. For, if what someone deserves depends on what they do in just arrangements, a logical circle seems to be involved in explaining justice in terms of desert. Like Nozick's interpretation, Scanlon's seems to show why all desert-tracking conceptions of justice must be wrong in principle – without, however, attributing the regression principle to Rawls. What's more, the revisionist employment of the idea of legitimate expectations with respect to desert seems bound to extend also to the concept of responsibility. For in both cases the underlying idea is that the moral significance of choice depends on some reasonably just background circumstances.

---

[28] Utilitarians could try to explain 'desert' in terms of the utility of entitlement-bestowing institutions. Sidgwick, apparently, sought to do this. See D. Miller, *Social Justice*, p. 95.

If, on the other hand, the idea of legitimate expectations is understood only as a parallel, somewhat desert-like concept, then *desert*-tracking conceptions of justice do not involve a logical circle. Legitimate expectations depend logically on justice, which depends logically, on this view, on desert.

There is an argument in favour of the non-revisionist understanding of legitimate expectations. For the revisionist interpretation seems to revise our common understanding of desert too radically. Normally we don't think of someone's desert as itself an entitlement, though it may give rise to an entitlement.[30] We think of their desert as a more primitive fact about them, depending only on their character, behaviour, or other attributes. As such, they may deserve things, on the common view, even in unjust circumstances, so long as they have the relevant qualities. To claim that the only true grounds of desert claims are entitlements generated by legitimate expectations, is to claim that many of our desert claims are false, and to change our understanding of those which remain – perhaps in too radical a fashion.

It is an interesting high-level issue whether there are any limits on philosophical revisionism. The argument I've just outlined supposes that there may be, so that we may cast doubt on an argument by pointing out how radically it revises our pre-theoretical beliefs. To some philosophers that smacks of unjustified theoretical conservatism, reminiscent of the cruder versions of ordinary language philosophy. According to those philosophers we should not place *a priori* limits on what philosophical argument can show. Although I cannot argue for it here, my view is that there are some limits on the extent to which philosophical argument can justifiably revise our pre-theoretical beliefs,

---

[29] He concludes that: "A person can be entitled to a reward he does not deserve, or deserving of a reward he has not qualified for." See J. Feinberg, "Justice and Personal Desert", in his *Doing and Deserving*. The quotation is from p. 72.

[30] Compare D. Miller, *Social Justice*, pp. 84-87, 90-92.

but that we don't know where those limits are. So argumentative appeals to them should always be tentative. If we adopt that policy, the worry about radical revision does not appear to be a decisive reason to reject the revisionist view.

Suppose, though, that we *are* somewhat concerned by the radical revisions which seem to follow if we seek to explain the true grounds of desert in terms of legitimate expectations. There is a way of retaining the revisionist view whilst softening its implications. Recall that the key idea is that persons gain entitlements by acting in accordance with the rules of reasonably just arrangements. It is very important how we understand 'reasonably' here. Presumably persons can gain entitlements by acting in accordance with the rules of *nearly* just arrangements. I do not lose the entitlement to my winnings if I play and win the lottery in a slightly unjust state. But exactly how just must the background arrangements be, for the idea of legitimate expectations, or Scanlon's broader idea of the moral significance of choice, to get off the ground?

I'll call this the *threshold* issue. The idea of naturalising agency-implicating concepts is that we can explain our judgements involving those concepts by reference to two things: (a) actual behaviour, and (b) a set of background arrangements, which must meet a certain threshold of justice. Reliance on the concept of justice is what promises to allow us to dispense with the idea of possibilities being within persons' reach, where what is meant by that can't be explained in terms of our best predictive theories (it is that idea which naturalism is hostile to). So the threshold issue arises in general for the naturalising strategy.

There is a trade-off between holding the threshold high, and being able to explain our real-world judgements involving agency-implicating concepts. For the real world is only very imperfectly just, and so naturalism with a high threshold can't get going for

very many of our judgements of desert, responsibility, and so on. If we think that some of our judgements of responsibility or desert, made in the context of significantly unjust arrangements, are nevertheless correct, and we're committed to the naturalising strategy, that's a reason for thinking that the threshold must be low. Only then could we explain how, for example, persons in Nazi Germany could be responsible for their actions.

If the threshold is not very high, then even the revisionist use of the idea of legitimate expectations may avoid the logical circle apparent in desert-tracking principles of justice. Suppose that we may specify justice imperfectly without any reference to desert, and that the threshold of justice (for the idea of legitimate expectations) is low enough for regimes satisfying the imperfect specification to meet it. Then justice has enough independent content for a naturalist explanation of desert to be possible. But if judgements of desert draw not only on the idea of legitimate expectations, but also on some independent considerations, desert may have sufficient independent content to help refine our understanding of justice, beyond the threshold. According to such a view, the two concepts are *mutually interdependent*, each with some content which is independent of the other, and each helping to determine the other.[31] It may be, for example, that our understanding of agency would leave the concept of desert indeterminate if it were not explained, in part, in terms of the idea of legitimate expectations. But our grip on agency may nonetheless contribute to our understanding of justice.

So Scanlon's emphasis on the second of Rawls's doctrines does not quite rule-out all desert-tracking conceptions of justice. First, it must be the case that the idea of legitimate expectations is intended as a revisionist account of the true grounds of desert.

[31] I referred to mutual interdependence views of the relationship between justice and *responsibility*, in Chapter Two (section 1). There I noted that it seems puzzling to claim that justice could be dependent on responsibility, *and* that responsibility is not independent of justice. But really the puzzle is a

Second, the threshold must be reasonably high, or the idea of mutual interdependence between justice and desert must be wrong for some reason we have not considered.

(The distinction between the mutual interdependence view, and the view which I've been discussing up to now, is an important one. The former is naturalist only in a limited way, since it does not deny the dependence of justice on agency-implicating concepts. *Thoroughgoing naturalism*, by contrast, claims that justice *is* independent of agency-implicating concepts such as desert and responsibility. In Chapter Four I shall discuss some claims which cut against naturalism in its thoroughgoing, but not its limited, form.)

Now consider a final interpretation of Rawls's remarks on desert. Nozick and Scanlon gave a common answer to the second of our interpretative problems. Both explained Rawls's criticism of the three desert-tracking rivals to democratic equality, as based on a quite general argument against desert-tracking principles. For Nozick, Rawls's view was that desert-tracking principles must be wrong, because there's no desert to track. For Rawls according to Scanlon, desert-tracking principles fall into logical incoherence because what people deserves depends on what justice requires (although we've just lodged qualifications to this view).[32]

---

manifestation of the general unfamiliarity of mutual interdependence as a relation between theoretical terms. It is not, I believe, the result of any incoherence in mutual interdependence views.

[32] A prominent interpretation which I haven't mentioned is common to Gauthier and Sandel. They present Rawls as claiming that desert-tracking principles are wrong because they treat as individual assets what are really common assets, namely, individual talents and abilities. I think this gets things the wrong way around. Rawls says that talents and abilities may be treated as 'in effect' common assets, but this is explained by, not explanatory of, his criticism of desert-tracking principles. See *TJ*, p. 101; D. Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986), pp. 220-221; M. J. Sandel, *Liberalism and the Limits of Justice* (Cambridge: Cambridge University Press, 1982), pp. 73-74.

G. A. Cohen offers a quite different stance on the second interpretative issue.[33] According to him, Rawls's thought is that the three desert-tracking rivals to democratic equality fail for reasons which are specific to *them*. They fail because *they* don't track desert, not because there's no desert to track, or because who deserves what depends on what justice requires. According to this view, Rawls does not object to the desert-tracking aim in general, and he accepts that persons can deserve things in a way which is independent of their legitimate expectations, but he thinks the three rivals to democratic equality don't track the right things. They track the distinction between luck in natural abilities and luck in social background, in different ways, and this distinction is morally arbitrary – at least insofar as it doesn't reflect what people really deserve.

Cohen portrays Rawls as worried by the practical difficulties in getting institutions to track desert. Once again, the attributed criticism of desert-tracking conceptions extends naturally to responsibility-tracking conceptions. If desert is difficult to track, then so is (non-naturalised) responsibility. That's because the difficulty in tracking desert is largely a result of its agency-implicating features. In order to know what someone deserves, in the non-naturalised sense, we usually have to know which possibilities are or were within their reach. That's the difficult part – not listing the various ways in which persons may deserve things (as compensation, reward, and so on) – and that difficulty is present too when we judge what persons are responsible for. For that reason, one would expect a Rawlsian of the type Cohen describes to want to naturalise responsibility as well as desert.

Cohen raises the issue of whether Rawls thinks that the practical difficulties in tracking desert are decisive against all attempts to do so. Rawls may think that it is just

[33] See G. A. Cohen, "On the Currency of Egalitarian Justice", *Ethics* 99 (1989), pp. 914-915. See also

the three rivals to democratic equality which fail to track desert. It's plausible that they do, perhaps – but then democratic equality is recommended only very weakly by this thought, since not only is the list of four conceptions which Rawls considers not exhaustive, it is not exhaustive even of possible desert-tracking conceptions. The reflections on desert would then have very limited critical scope. On the other hand, Rawls may think that the practical difficulties in tracking desert extend to all feasible institutional mechanisms, so that no desert-tracking conception would escape the difficulties. That would do more to recommend democratic equality, but it would also require further argument. As Cohen puts it, "The practical difficulty of telling how much of it [effort] merits reward hardly justifies rewarding it at a rate of 0 percent, as opposed to a rate somewhere between 0 percent and 100 percent, for example, through a taxation scheme whose shape and justification escapes, because of its deference to effort, the writ of the difference principle."[34]

The further argument might be forthcoming. It could be that the most accurate feasible mechanism for tracking desert would involve significant inaccuracies. It could be, too, that the injustice involved in responding to desert using this mechanism would outweigh the injustice involved in not responding to desert at all (supposing as we are that, if we leave the practical difficulties aside, the desert-tracking aim is correct). But we lack any such detailed evaluation of the feasibility of desert-tracking mechanisms, and it is unwarranted simply to assume that none could promote justice on balance.[35] Rawls's

---

R. Young, "Egalitarianism and Personal Desert", *Ethics* 102 (1992), p. 323.

[34] Cohen, "On the Currency of Egalitarian Justice", p. 915.

[35] Roemer's 'pragmatic' proposal for designing responsibility-tracking institutions is relevant here, though the sense of responsibility he considers is not the ordinary one, but one culled ('pragmatically') from the political culture of the society in which the institutions are to operate. See J. E. Roemer, "A Pragmatic Theory of Responsibility for the Egalitarian Planner", *Philosophy and Public Affairs* 22 (1993).

criticism of desert-tracking principles, in its Cohenesque interpretation, is vulnerable to criticism in this regard.

Cohen's interpretation is attractive because, like Scanlon's, it avoids attributing to Rawls an implausible view about the conditions for desert. The regression principle is controversial, and it is probably false. Moreover, like Scanlon's interpretation, Cohen's is favoured by textual evidence as well as charity. In his discussion of effort, Rawls writes,

> The effort a person is willing to make is influenced by his natural abilities and skills and the alternatives open to him. The better endowed are more likely, other things equal, to strive conscientiously, and there seems to be no way to discount for their greater good fortune. The idea of rewarding desert is impracticable.[36]

It may be, of course, that Rawls's thoughts about desert are confused, embracing at one minute scepticism based on the regression principle, at another moment Scanlonian revisionism, and Cohenesque worries about practicalities at still other times. That's an interesting issue, but we can leave it aside. Our primary aim is to explain the strongest case against desert-tracking principles which a Rawlsian could mount.

Scanlon's and Cohen's interpretations of Rawls provide a map of the most promising possibilities for pursuing the naturalising strategy in the Rawlsian framework. Scanlon's raises two issues: that of the revisionist versus the parallel use of the idea of legitimate expectations, and that of the threshold of justice required for naturalism to get off the ground. The naturalising strategy in its full-fledged form requires the revisionist path, and if we are worried by the subsequent difficulties in explaining judgements of

desert and responsibility in significantly unjust circumstances, a low threshold seems most plausible. Rawls's argument on Cohen's interpretation, on the other hand, is not by itself an example of the naturalising strategy. But it is compatible with certain elements of that strategy. One may find desert-tracking impracticable, and at the same time use the idea of legitimate expectations to describe a ground of entitlements which is somewhat similar, but parallel, to desert. That does not strictly naturalise desert, but it may be useful in allaying fears that democratic equality does not recognise, for example, the entitlements which people have, in certain circumstances, to the things they make.

5. Rawls's use of judgements of desert and responsibility

Rawls is critical of conceptions of justice which seek to track desert, and his criticisms apply equally to appeals to responsibility. It is true that, according to either Scanlon's or Cohen's interpretations, Rawls's remarks do not quite rule-out desert-tracking conceptions. But he does insist that the conception of justice he favours has no tendency to track moral desert. So it is interesting to note that Rawls seems to make significant use of judgements of desert and responsibility in his theory of justice.

First, his claim that differences in (i) productivity, or (ii) talents, or (iii) effort, are not satisfactory grounds for desert, is itself a claim about what people deserve, albeit one needing interpretation. One could interpret it Nozick's way, as following from a general scepticism about desert which is a consequence of adopting the regression principle. One might then demur at saying that it is a claim about what people deserve. It would be a claim about the failure of all humans and human activities to satisfy the (impossible to

[36] Rawls, *TJ*, p. 312. See G. A. Cohen, "On the Currency of Egalitarian Justice", pp. 914-916. I cited

satisfy) conditions of desert. People wouldn't be so much *undeserving* of their natural talents, on such a view, as wholly *non*-deserving creatures. They are not deserving of anything in any cases at all. If that were Rawls's view, it might be better to say that he rejects all claims of desert, than that he relies on the substantive judgement of desert, that people do not deserve their natural talents.

We can ignore that scruple, however, since we should not interpret Rawls's claim in Nozick's way. If we treat it instead in Scanlon's way, it means something like this:

> (S1) when the threshold of justice *is not* reached, differences in (i) productivity, or (ii) talents, or (iii) effort, are not satisfactory grounds for desert,

and

> (S2) when the threshold of justice *is* reached, differences in (i) productivity, or (ii) talents, or (iii) effort, are not satisfactory grounds for desert.

Now, if one holds Scanlon's view, it follows from the doctrine about the conditions for the moral significance of choice, that (S1) is correct. But (S2) is still controversial: it is a judgement about what people deserve that one may reject, but one which Rawls relies upon. It is by no means obvious that, given a reasonably just institutional background, differences in one or more of these factors could not be satisfactory grounds for desert.

If, on the other hand, we treats Rawls's view in Cohen's way, it means:

---

this passage in section 3, in discussing Nozick's interpretation.

(C) differences in (i) productivity, or (ii) talents, or (iii) effort, do not map in any simple way onto the real grounds for desert, though those grounds may well include differences in any of these factors.

(C) says: what persons deserve may well depend on their productivity, or talents, or the effort they make, but the weight or significance of these factors relative to each other varies from case to case. Hence we could not track desert using an institutional mechanism which tracks one or more of these factors according to a standard function. Once again, that may be true, but it is not obviously and uncontroversially so.

So whether we follow Cohen or Scanlon, Rawls's criticism of desert-tracking principles rests on a controversial judgement about what persons deserve. In addition, Rawls relies on controversial judgements of responsibility in his discussion of primary goods.[37] Rawls responds to an objection from Kenneth Arrow, that his idea of an index of primary goods is a poor measure of personal advantage for the purposes of claims of justice, since it does not take into account variations in preferences between persons. Arrow's idea is that since the same bundle of primary goods can produce much less preference-satisfaction in one person than another, because they have different preferences, the idea of primary goods is unsuited to its purpose. Rawls replies that:

As moral persons citizens have some part in forming and cultivating their final ends and preferences. It is not by itself an objection to the use of primary goods that it does not accommodate those with expensive tastes. One must argue in addition that it is unreasonable, if not unjust, to hold such persons responsible for their

> preferences and to require them to make out as best they can. But to argue this seems to presuppose that citizens' preferences are beyond their control as propensities or cravings which simply happen. Citizens seem to be regarded as passive carriers of desires. The use of primary goods, however, relies on a capacity to assume responsibility for our ends.[38]

This seems straightforwardly to invoke a controversial judgement about what persons are responsible for. Arrow's objection fails, Rawls seems to be saying, because persons are responsible for their tastes, and hence it is fair to let them bear the costs of their tastes.[39]

In interpreting this passage, however, it is important to bear in mind the distinction between (a) being responsible in the sense that one is the prime agent with respect to some circumstance, and (b) being responsible in the sense that, whether or not one is the prime agent, one is a legitimate bearer of costs with respect to some circumstance. Cases of strict liability illustrate the second of these senses of responsibility. If I am strictly liable for something, I am responsible for it whether or not my agency determines what happens to it: the costs are legitimately imposed on me in virtue of my liability, not my agency. Rawls might be claiming that primary goods provide the appropriate metric of advantage, not because persons are usually agents with respect to their preferences (bringing those preferences about), but because persons are the legitimate bearers of costs for their expensive tastes.

---

[37] See J. Rawls, "Social Unity and Primary Goods", in A. Sen and B. Williams (eds.), *Utilitarianism and beyond* (Cambridge: Cambridge University Press, 1982).

[38] Rawls, "Social Unity and Primary Goods", pp. 168-169.

[39] Cohen argues that justice sometimes requires compensation for expensive tastes, and that whether it does depends on the person's moral responsibility for those tastes. See Cohen, "On the Currency of Egalitarian Justice", p. 923. But see also note 43 below.

This reinterpretation of the concept of responsibility is an application of the naturalising strategy. As with the revisionist account of desert in terms of legitimate expectations, a concept which seems to implicate our understanding of persons as agents is rendered less troubling by subordinating it to an independent account of what is fair, or just. We find out who is responsible, in the sense of (b), by asking what distribution of burdens would be fair or just. Such a view, we should expect, would be attractive to Rawls, at least as portrayed by Scanlon. And some of Rawls's phrases in the passage I've quoted may support this view. (He asks what it is reasonable or just to hold persons responsible for, and later in the same article writes of a "social division of responsibility".[40])

If this is Rawls's view, however, he has not answered Arrow's challenge but merely thrown down a different one of his own. Faced with the question of why we shouldn't be worried about expensive tastes, Rawls asks why it is unreasonable to let people bear the costs of their own expensive tastes. He can claim that the sense of 'responsibility' he invokes in making this response is derived from our idea of fairness, not our idea of moral responsibility – but, in doing so, he foregoes the putative explanation of *why* it is fair to let people bear the costs of their expensive tastes. If they were morally responsible for their tastes, we'd have an explanation; if the claim is just that it's reasonable because it is fair to let them bear the costs, we simply have two ways of making the same assertion.

Rawlsians might find other ways of explaining why it is fair to let persons bear the costs of their expensive tastes. They could appeal to Dworkin's criterion for compensation of expensive tastes. As we saw in Chapter Two, Dworkin asks not whether

[40] *Ibid.*, p. 170.

a person is morally responsible for her tastes, but whether she would repudiate them.[41] He treats tastes which the bearer regrets, but whose non-satisfaction would nevertheless be painful, as handicaps. And he thinks it is just to compensate for such tastes, though it is not just to compensate for expensive tastes which are not regretted.[42]

The repudiation criterion may avoid implicating our understanding of person as agents, and at the same time avoid a vacuous explanation of the requirements of fairness in terms of fairness itself. It does not rely on the same kind of understanding of counterfactual possibilities as does our usual understanding of agency – though it does appeal to the hypothetical 'would be repudiated'. (But perhaps this hypothetical can be explained as a special kind of prediction?) Insofar as it does avoid implicating our understanding of agency, it is well-suited to explaining the treatment of expensive tastes, for those who wish to pursue the naturalising strategy.

6. Conclusion

There seems to be a tension in Rawls's views on desert and responsibility. He is critical of desert-tracking principles of justice, and I've claimed that these criticisms extend naturally to responsibility-tracking principles. On the other hand, he relies on the controversial judgement that differences in productivity, talent, and effort are not proper

[41] R. Dworkin, "What is Equality? Part 2: Equality of Resources", *Philosophy and Public Affairs* 10 (1981), pp. 302-303.

[42] It is interesting that Cohen, who is one of the most prominent champions of the idea that justice tracks responsibility, adds to his conception of justice a Dworkin-like condition. In response to some of Scanlon's arguments, he makes the following amendment: "Instead of saying, 'compensate for disadvantages which are not traceable to the subject's choice,' say, 'compensate for disadvantages which are not traceable to the subject's choice and which the subject would choose not to suffer from.'" G. A. Cohen, "On the Currency of Egalitarian Justice", p. 937.

grounds of desert, and he seems to rely also on a controversial judgement about responsibility, that persons are responsible for their ends or tastes.

There is no outright contradiction in this, because according to the best interpretations, Rawls's criticisms of desert-tracking (and, by extension, responsibility-tracking) principles do not apply equally to all such principles. And as we've just seen, Rawlsians could appeal to Dworkin's repudiation criterion, or something like it, to explain why it is fair to let persons bear the costs of their expensive tastes. Nevertheless, Rawls's reliance on controversial judgements of responsibility and desert is in tension with his contractualism. As I argued in section 1, it is difficult to see how any conception of justice which relies on (non-naturalised) judgements of desert or responsibility, could be given a contractualist justification of the kind Rawls seeks.

It is tempting to treat this tension in Rawls's views as evidence that the naturalising strategy cannot be applied across the board. Perhaps there is some essential dependence of justice on responsibility or desert, which resists naturalisation. Perhaps we just can't explain what justice requires without relying on non-naturalised judgements of desert or responsibility. But it would be premature to draw this conclusion.

With regard to the doctrine of responsibility for ends, Rawls is perhaps just insufficiently scrupulous in his naturalism – failing to see the importance of stating clearly whether the doctrine of responsibility is to be understood in a naturalised or a non-naturalised sense, and if the former, what independent considerations explain why it is fair to let persons bear the costs of their expensive tastes. I've suggested that a more thoroughgoing naturalist might appeal at that point to something like Dworkin's criterion of repudiation.

On the other hand, Rawls's judgement that differences in productivity, talent, or effort are not proper grounds of desert can for some purposes be dispensed with. We were interested in seeing exactly how it was supposed to count in favour of democratic equality, because of our general interest in getting Rawls's remarks about desert straight. The conclusion of that discussion, however, was that the remarks on desert do very little work in justifying democratic equality. Rawls could, in one sense, just dispense with them, and rely instead solely on the argument from the original position, as presented in Chapter 3 of *TJ*. Together, these revisions seem sufficient to complete Rawls's naturalism.

In another sense, however, the remarks on desert cannot be dispensed with. That is because a thoroughgoing naturalist about desert and responsibility must do two things. She must show that an account of justice can be developed which does not rely on any non-naturalised judgements of desert or responsibility. But ultimately she must also provide us with an non-instrumentalist justification for naturalising in the first place. We want to know whether naturalism can be carried through, but we want to know also whether naturalism should be carried through.

The instrumentalist reasons for naturalising agency concepts are strong ones. It is difficult to explain what justice requires in terms of responsibility or desert, because these concepts are the subject of persistent dispute in their application, and they are philosophically troublesome. But the instrumentalist reasons are in order only so long as we adopt the theory-building stance, in which we try to anticipate the success of broad strategies or general approaches. Ultimately, we want to know why desert and responsibility aren't properly understood independently of what justice or fairness requires.

In the last two chapters we've been looking for arguments against the naturalising strategy, and so placing the burden of proof on those who reject it. But ultimately there is a sense in which the burden of proof is on the naturalist, to show why common sense gets the relation between justice and fairness on one side, and desert and responsibility on the other, the wrong way around. Something like the judgement that differences in productivity, talent or effort are not proper grounds for (agency-implicating) desert is needed to give us a non-instrumentalist reason for naturalising all other judgements of desert in an account of justice. Without it, we have an internally consistent naturalised theory, but no argument against rival, non-naturalised theories.

So we have reached a mixed conclusion. We have found no reason to think that the naturalising strategy can't be carried through. Rawls's doctrine of responsibility for ends can be interpreted naturalistically, and can be motivated by an independent consideration such as Dworkin's repudiation criterion. And, in one sense, the judgement about desert can be dispensed with. On the other hand, some non-naturalised judgements of desert or responsibility seem required to motivate naturalism in a non-instrumentalist, non-question-begging way.

In these early chapters we have explored the naturalising strategy, looking for clues as to its viability, first in the debate about what egalitarians should equalise, and second, in this chapter, in Rawls's stance on judgements of desert and responsibility. We have yet to find good reasons to think the strategy can't be carried through, though equally we've found no strong non-instrumentalist motivation for it. In the next chapter, we'll consider a more direct challenge to the naturalising strategy. I shall claim that Cohen's arguments about incentives raise a general problem which cannot be solved

without relying on agency-implicating judgements about *what persons could do*. Hence no satisfactory account of justice can avoid relying on such judgements.

Chapter Four

Incentives, Agency, and Benefit

In this chapter and the next I argue against thoroughgoing versions of the naturalising strategy.[1] In its purest form, the naturalising strategy operates on the assumption that political theory could be made independent of all agency-implicating judgements, but I shall argue that this assumption is false. It is false, at least, for any political theory which makes use of the concept of *benefit*, or the concept of *promoting the good.* These concepts cannot be understood, I shall claim, without invoking agency-implicating judgements.

Impure forms of naturalism may seek to make political theory independent of some agency-implicating judgements, but not all.[2] The argument that follows does not cut against them directly; though, insofar as impure forms of naturalism share the underlying rationale of pure naturalism, they may be undermined indirectly by what follows. The desire to avoid agency-implicating judgements simply because they are agency-implicating is shown to be wrong-headed, I believe, by the conclusion that we cannot do without relying on some such judgements. The relevant dispute is about *which* ones we should rely on, not about whether we can do without relying on them altogether.

---

[1] The conclusion is consistent with limited naturalism, where that is understood to be the view that justice is conceptually *interdependent* with agency-implicating concepts (possibly including desert and responsibility). Thoroughgoing naturalism, by contrast, claims that justice is conceptually independent of agency-implicating concepts, but agency-implicating concepts such as responsibility and desert are conceptually dependent on justice.

[2] So I have narrowed the target in two ways: I am arguing against *pure* forms of *thoroughgoing* naturalism. It may be objected that no-one espouses such a view. My response is that the possibility of pure, thoroughgoing naturalism is often relied upon in political theory, as justification for avoiding, or at least postponing, discussion of agency-implicating judgements. So in that sense the view I am arguing against has an important role in the way political arguments are conducted. (And I think that Scanlon, arguably, espouses pure thoroughgoing naturalism as his official doctrine: see my discussion

Our concern until now has been with the conceptual relationship between justice on one hand, and desert and responsibility on the other. 'Naturalism', as I've been using the term, is the view that justice is not conceptually dependent on desert or responsibility. However, the focus of discussion in this chapter will be neither desert nor responsibility, but the idea of *agency-implicating concepts and judgements*. A concept is agency-implicating, if it cannot be explained without reference to counterfactual judgements about *what persons could do*; and a judgement is agency-implicating if it makes essential use of such a concept. Agency-implicating judgements about the future, I shall assume, pick-out a broader range of possibilities than do our best predictions about what will actually happen.[3] That's because, on our usual understanding of agency, what persons actually do (including what they will do) is a subset of what they could do.

As I suggested in Chapter Two, it is agency-implicating concepts and judgements which naturalism is really hostile to. For it is those which bring the philosophical and other problems which the naturalist wants to avoid. But it is nevertheless true that I shall be arguing that justice depends on agency-implicating judgements, and not, strictly speaking, that it depends on desert or responsibility. For it could be that desert and responsibility are agency-implicating, and that justice depends on agency-implicating concepts, without it being true that justice depends on desert or responsibility. So although our first interest in the naturalising strategy was via our interest in desert and responsibility, we now approach it at the deeper, if less well-specified level, at which it is the view that justice is independent of agency-implicating concepts.

The argument takes off from G. A. Cohen's criticism of a certain putative justification of social and economic inequalities, the so-called *incentives argument* for

---

of his views in Chapter Three, section 4, where I suggested that his naturalism could be interpreted as either thoroughgoing or limited.)

[3] Something may be unpredicted in two senses: it may not feature in our predictions, or it may be incompatible with them. This latter, stronger sense is the one I mean to invoke. See also Chapter Five below.

inequalities.[4] I find Cohen's critique of the incentives argument compelling, but I shall depart from his discussion in two respects. Cohen thinks that the argument fails because it doesn't respect a certain kind of community. I shall offer a different diagnosis, according to which the argument in question fails for quite general reasons to do with the concept of *benefit*. The trouble with the concept of benefit is that it is inherently comparative, asking us to compare one policy, action, or state of affairs, with some others, but it is often unclear what the relevant range for comparison should be. If I'm right, the upshot is that the issues Cohen discusses extend further than he suggests. Very many political arguments are subject to this problem: they make implicit reference to a range of alternatives, but the reasons for including some possibilities in the range and leaving others out, are unclear.

The nub of the argument against the naturalising strategy is that the extent of the range ought to depend on what persons could do, where 'could' is understood in a counterfactual sense. But it turns out that we must be careful to distinguish two forms of political argument. This is the second departure from Cohen's discussion. In *ideal theory*, the range of relevant alternatives is generally wider than in *deliberation*, in which information about the particular circumstances of the agent in question should be used to rule many possibilities out. And, in fact, it is much more difficult to argue against naturalism if it is applied to deliberative argument. The present chapter deals on the whole with the argument against naturalism applied to ideal theory, whilst the next chapter attempts the trickier task of extending the anti-naturalist conclusion to deliberative arguments. The first task, however, is to explain the argument from incentives.

---

[4] See G. A. Cohen, "Incentives, Inequality, and Community", in S. Darwall (ed.), *Equal Freedom. Selected Tanner Lectures on Human Values* (Ann Arbor: University of Michigan Press, 1995); see also, G. A. Cohen, "The Pareto Argument for Inequality", *Social Philosophy and Policy* 12 (1995), G. A. Cohen, "Where the Action Is: On the Site of Distributive Justice", *Philosophy and Public Affairs* 26 (1997). I shall refer to these articles from now on as "Incentives", "Pareto", and "Site", respectively.

1. Efficiency, Inequality, and Benefit

Most theories of justice treat the fact that someone benefits under a scheme as a reason *pro tanto* in favour of that scheme, as far as justice is concerned.[5] Most theories also treat distributive considerations as having some intrinsic importance.[6] But any theory of justice with these two features will be subject to the following tension. It may turn out, as a matter of fact, that departures from the desired distribution predictably benefit at least one person, without harming any others (on the plausible assumption that there is not a constant sum of advantage). Since all such theories, we've said, think that someone's benefiting is *pro tanto* good for justice, all such theories will be under pressure to allow (or even require) such departures from the stipulated pattern.

The tension is the result of giving intrinsic weight to two facts about advantage: how much advantage persons have, and the distribution of advantage amongst persons. The former kind of fact concerns *efficiency*, and the purest cases in this respect are those where at least one person could do better, without anyone doing worse, under an alternative arrangement. Considerations of efficiency needn't dominate in a theory of justice, but I assume that theories of justice, for the most part, will not simply ignore considerations of efficiency. At any rate, I shall be discussing those which do not.

Such theories require principles for reconciling the two kinds of concern with advantage. Political theorists have recently paid attention to the complex ways in which this may be attempted.[7] Our topic, however, is an argument which is

---

[5] *Pro tanto* reasons may be contrasted with *prima facie* reasons and all-things-considered reasons. *Prima facie* reasons need be reasons only on the face of it: there is no guarantee that they will survive reflection. *Pro tanto* reasons survive reflection, but may address only part of the situation at hand. All-things-considered reasons survive reflection and address all of the aspects we deem relevant. (Recall that, by 'justice', I mean 'justice in holdings'.)

[6] The distinction between distributive and aggregative principles (albeit with respect to 'want-satisfaction' in particular) is made in B. Barry, *Political Argument* (Hemel Hempstead: Harvester Wheatsheaf, 1990), pp. 43-44.

[7] On the complexities of possible principles of reconciliation, see L. Temkin, "Inequality", *Philosophy and Public Affairs* 15:2 (1986). Similar issues are discussed by Parfit, though the point of his discussion is to suggest that the true object of concern for many who thought they were worried about distribution, is not the intrinsic value of certain patterns of distribution at all, but rather the special

compatible with any such theory. This argument exploits the tension between concern with distribution and concern with efficiency.

We can see that considerations of efficiency may conflict with considerations of distribution in theory, but is there any reason for thinking that they are *likely* to do so in practice? The commonest reason for thinking that such conflicts are likely in practice has to do with *incentive effects*. Any distributive pattern is liable to be disturbed by free market transfers, which, we might charitably assume, tend to promote efficiency. But many of these gains, supposing they exist, will be insufficient to justify departure from the distributive pattern, because they benefit only the parties to the transaction, and so may fail to benefit the persons or groups targeted by the principles of reconciliation (between efficiency and distribution).[8] In the case of the labour market, however, it seems likely that gains of the required sort will be available. That's because transfers may have predictable effects on third parties – including the target beneficiaries specified by principles of reconciliation – insofar as they affect productive output. Hence, incentive effects are particularly likely to pose a case for departure from the specified distributive pattern.

Cohen considers an argument of this sort which is stated in terms of Rawls's difference principle, and which purports to give an *egalitarian justification* of *inequalities*. This is not paradoxical, since, as we've observed, even those who give weight to distributive concerns, as egalitarians do, are likely to give some independent weight to considerations of efficiency. Rawls's difference principle reconciles the two concerns with advantage as follows:

> (DP) Economic inequalities are justified when they make the worst off people materially better off.[9]

---

urgency of the claims of the worst-off. See D. Parfit, "Equality or Priority?", *The Lindley Lecture,* University of Kansas (1995). Also relevant is J. Raz, *The Morality of Freedom* (Oxford: Clarendon Press, 1986), Chapter 9.

[8] For example, Rawls's difference principle targets the least advantaged. By no means all efficiency-promoting market transfers would benefit this group. See Rawls, *A Theory of Justice*, p. 97f.

[9] This is the statement of the difference principle which Cohen uses in his construction of the incentives argument. See Cohen, "Incentives", p. 339.

Hence, according to DP, departures from the specified distributive pattern (equality of primary goods) are justified when they lead to efficiency gains for the worst-off members of the population.

Cohen's target argument involves also a factual claim about the effects of different income structures and tax regimes on the levels of well being of badly off people. It is claimed that, at some level, increasing the top rate of income tax would make badly off people worse off, since there would be less incentive for talented, potentially highly productive, people to work hard. This would lead to a decrease in total product, offsetting the distributive gains for the egalitarian of the more progressive tax regime. Given knowledge of reliable economic theories, one can predict the harmful effect, by the egalitarian's own lights, of raising the top rate of income tax.

These two thoughts, about the importance even to egalitarians of considerations of efficiency, and the predictable effects of incentive-generating inequalities on the well being of the badly off, are politically and theoretically challenging for socialist egalitarians. For they enable proponents of the incentives argument to claim that they are enlisting economic truth on the side of a restriction on redistribution, in the name of a plausible interpretation of the value of equality.[10] In practical politics they add up to a dilemma for those with redistributive aims: beyond a certain point, redistribution is either a mark of economic ignorance, or a mark of willingness to harm the badly off in the name of an ideology which claims to help them. As Cohen notes, this kind of challenge was politically effective in the 1980s in seeming to justify Thatcherite economic policy.[11] Judging by New Labour's tax policy, it seems that it retains weight in Britain in the 1990s.

In what follows I shall restrict my discussion to this particular argument which Cohen sets out to criticise. It has sufficient pedigree to justify the restriction. But it is

[10] Cohen centres his discussion of the argument for inequalities from incentive effects on versions which appeal to Rawls's Difference Principle. See Cohen, "Incentives", pp. 333-334.
[11] Cohen, "Incentives", p. 333.

worth emphasising that it is not the incentives argument's incorporation of the difference principle, nor its egalitarianism (such as it is), nor even its particular concern with incentive effects, which generates the problem that Cohen latches onto. Its essential feature is that it attaches weight to considerations of efficiency – or, in other words, it is minimally beneficent. The *minimal principle of beneficence* states that,

(MPB) We have a (*pro tanto*) reason of justice to make persons better off.

MPB is a very uncontroversial principle. It says merely that it is in one way good to make persons better off. Almost every theory of justice includes or entails this principle.[12]

MPB should be contrasted with more complex and stronger principles of beneficence, such as Rawls's difference principle. Such principles are designed to guide action when we face complex circumstances. The simplest circumstances, in which the simplest principles of beneficence are sufficient guide, involve a fixed number of identifiable potential beneficiaries. (The simplest case involves a single identifiable potential beneficiary.) Complications arise when one or more of the following factors is variable with respect to different possible courses of action: the size of the total relevant population; the identity of the total relevant population; the number of potential beneficiaries; the identity of potential beneficiaries; the amount of benefit which could be conferred on each potential beneficiary.[13] MPB is a modest

[12] The principle is so minimal that it might seem that no theory of justice could deny it. It is much less controversial than the Pareto Principle, for example, since it says only that it is in one way good to make persons better off – not that it is always right, all things considered, to make someone better off if, in doing so, no-one is made worse off. It's true that some historical-entitlement theories may deny MPB at ground-level, if they disallow all ground-level principles which do not have the requisite procedural form. But if the underlying theory of rights is a Benefit theory, MPB is likely to be included nevertheless, as part of the justification of the particular set of rights which they endorse.

A residual class of theories which would not include MPB, are those which conceive of reasons of justice as separate in kind from reasons of beneficence. Such theories are likely to say that we have a reason to make people better off, but that the reason is not a reason of justice. It has become customary to include reasons of beneficence amongst those which an adequate theory of justice is expected to comprehend, however.

[13] Different-number and different-identity issues are discussed by Parfit and Temkin. See D. Parfit, "Equality or Priority?" *The Lindley Lecture*, University of Kansas (1995), and L. Temkin, "Inequality *Philosophy and Public Affairs* 15 (1986).

principle – it does not give us any guidance about what to do in cases like these, where we may have to judge the relative importance of the size of benefit to a single person against the number of persons who could be benefited, and so on.

Yet though MPB is a very uncontroversial and uninformative principle, it is sufficient to generate the kind of difficulty which Cohen addresses. That's because it is the notion of benefit itself which leads to the underlying problem, or so I shall argue in section 3 below. Cohen's critique of the argument from incentive effects relies on a point which has, so I claim, much wider application than Cohen himself indicates. The argument from incentive effects is just a special manifestation of the general problem, albeit one with particular saliency for those who give weight both to considerations of efficiency and to distributive issues.

2. Cohen's critique

Cohen thinks that the argument from incentive effects justifies much less inequality than its proponents believe, and possibly none at all.[14] He considers the argument applied to a decision whether or not to raise the top rate of income tax from 40 to 60 percent:

> Economic inequalities are justified when they make the worst off people materially better off . . .
>
> When the top rate of tax is 40 percent, (a) the talented rich produce more than they do when it is 60 percent, and (b) the worst off are, as a result, materially better off . . .
>
> Therefore, the top tax should not be raised from 40 percent to 60 percent.[15]

However, he sets out to challenge this argument whilst granting its normative premise (the difference principle), and *also* granting the assumption that (our best predictions

tell us), if the tax rate were raised, the talented would not work so hard and the worst off would be harmed.[16] How, then, does Cohen hope to challenge the conclusion?

In the first instance he sets out to criticise certain *utterances* of the incentives argument (that is, certain acts), rather than the argument itself.[17] Cohen notes that, "an argument changes its aspect when its presenter is the person, or one of the people, whose choice, or choices, make one or more of the argument's premises true".[18] He gives as an example of this phenomenon the argument which could be made for paying a kidnapper. This argument joins the normative premise that children should be with their parents, and the factual premise that the kidnapper will not return the child unless he receives payment, to reach the conclusion that the kidnapper should be paid. Of course, the normative premise is too simple; but Cohen's point is that the argument in any case looks different, and would be received differently, if it were to be presented by the kidnapper himself – even though the propositional content of the premises might be identical.[19] Kidnappers succeed only in showing themselves to be more vile by presenting this argument themselves, since the factual premise is true only because they make it true. And, he claims, the incentives argument is similar insofar as the talented rich *make it true* that they would not work so hard without incentives – like the kidnappers, they make the factual premise of the (respective) argument true, and our response to their uttering the argument reflects this fact.

Cohen does not conflate the justification of the act of making an argument with the justification of the conclusion of an argument. But he does claim that, in the case of the incentives argument at least, the former affects the latter. His argument is as follows: the talented do indeed make it true that they would not work so hard at the higher tax rate[20]; they cannot justify making this true; if they cannot justify making

---

[14] Cohen, "Site", pp. 5-6; "Incentives", p. 338.
[15] Cohen, "Incentives", p. 339.
[16] Cohen, "Incentives", p. 340.
[17] Cohen, "Incentives", pp. 340-347, esp. p. 345.
[18] Cohen, "Incentives", p. 344.
[19] Cohen, "Incentives", pp. 344-345.
[20] As I shall explain in the text below, the incentives argument is, in its standard versions anyhow, a comparative argument, which recommends one scheme of economic arrangements over another or some others. For simplicity, I shall follow Cohen in speaking of higher and lower tax rates. 'Would not

this true, then the incentives argument cannot be used (by them or anyone else) as a justification of inequalities, without violating a certain desirable kind of community.[21]

Cohen calls the desirable kind of community he has in mind, 'justificatory community'. A society enjoys justificatory community if most of its policies are *comprehensively justified* – and a policy is comprehensively justified if it is recommended by a valid argument with sound premises, *and* any behaviour which is mentioned in the premises of the argument is itself justified.[22] The second part of this conjunctive condition is what distinguishes comprehensive justification from ordinary justification. Cohen's idea seems to be that, in a just society, persons would not only benefit from the behaviour of others, in appropriate ways, but would also be able to justify their behaviour to each other.[23] Mutual justifiability of behaviour is a feature of just societies.

How does the idea of comprehensive justification show the conclusion of the incentives argument to be unjustified? Suppose we grant Cohen's critical premises – that the talented make the minor premise of the incentives argument true, and could not justify doing so. We want still to avoid saying that political argument has an eccentric logic, such that a valid political argument with true premises does not necessarily have a true conclusion. There are three ways in which we can construe Cohen's critique of the incentives argument, which do not involve attributing this view to him in respect of that argument. These ways of construing his critique produce three separable arguments, though Cohen himself does not clearly distinguish them.

The simplest argument does not, in fact, rely on the ideas of comprehensive justification or justificatory community. It simply claims that the normative premise of the incentives argument is *ambiguous*. The difference principle says that inequalities are justified only if they benefit the worst-off members of society. Cohen

---

work so hard at the higher tax rate' is likewise shorthand for 'would work at a rate which has the consequence that the worst-off are harmed, relative to some relevant alternative scheme'.

[21] Cohen, "Incentives", p. 354f.

[22] This is my formulation. For Cohen's, see "Incentives", pp. 347-353.

[23] Cohen, "Incentives", p. 350.

focuses our attention on what 'benefit' means here, exactly. He cashes it out as follows: an inequality benefits the worst-off only if it is *necessary* to enhance their position.[24] Now, there are two relevant senses of 'necessary' here, according to Cohen: necessary taking as given the motivations of the talented, and necessary "*tout court*, that is, independently of human will".[25] The difference principle trades, for its plausibility, on the second, unconditional sense of necessity; but the premise about the future behaviour of the talented, and consequent effects on the worst-off, relies on the former sense, which takes the motivations of the talented as parametric.[26] What's wrong with the argument from incentives, then, is that it equivocates in the sense of 'necessary' which is used, or implied, in the two premises.

Note that this argument relies on the assumption that the premise about the effects of raising tax rates would be false if 'necessary' were to be given the strict interpretation. (Without that assumption, there need be no equivocation in the incentives argument.) This assumption is justified only if it is true that the talented could work just as hard at the higher tax rate. Cohen thinks that this is true, and hence that the assumption is justified, at least for possible tax rate increases on the political agenda.[27] But the issue is not straightforward. It may be that some motivations, and so some patterns of work behaviour, just can't be produced without material incentives.[28]

The second argument does rely on the ideas of comprehensive justification and justificatory community, and hence it is more complicated than the first. It says that the incentives argument does not show that justice permits inequalities, because the inequalities which it justifies depend for their justification on behaviour which

---

[24] Cohen, "Site", p. 6: "The inequality consequent on differential material incentives is said to be justified within the terms of the difference principle, for, so it is said, that inequality benefits the worst off people: the inequality is *necessary* for them to be positioned as well as they are, however paltry their position may nevertheless be", emphasis added.

[25] Cohen, "Site", p. 8.

[26] Cohen, "Site", pp. 7-10.

[27] Cohen, "Incentives", pp. 355-362.

[28] See Cohen, "Incentives", pp. 359-361, where Cohen discusses a point which was made by Samuel Scheffler when the lecture was first delivered. Scheffler's point is that not all motivations can be summoned at will; some may require material incentives. Cohen concedes that the issue requires further attention, since, as he puts it, "what people are able to do depends on the reasons they have for doing it: with different reasons, the adrenalin flows to different extents" (p. 360).

could not be justified to the worst-off. What makes it true that the incentives scheme will benefit the worst-off, is the fact that they would do better under it than under a more equal division of benefits, given the motivations and the resulting behaviour of the talented. But, Cohen supposes, the talented could not justify their motivations or behaviour to the worst-off. Hence the incentives argument may justify inequalities generated by incentive schemes, but it does not show them to be comprehensively justified.[29]

This second critical argument, like the first, rests on the assumption that the talented could work just as hard at the higher tax rate. For presumably they could justify their behaviour to the worst-off if they were unable to work just as hard without the incentives. It relies also on the assumption that comprehensive justification, while not a rival logic, is somehow a requirement of an account of justice. I'll explain later the sense in which I take this to be true. But for the moment, note that the second critical argument can be presented in two ways: either as amounting to a denial of the first premise of the incentives argument, since that premise does not include the conditions for comprehensive justification; or as amounting to a second charge of equivocation, between the sense of justification employed in the first premise (the standard sense), and the sense presented in the conclusion (comprehensive justification).

Finally, Cohen criticises Rawlsian advocates of the incentives argument on grounds of internal inconsistency. Rawls, he notes, insists that persons in a just society both accept and act on the principles of justice.[30] Now Cohen supposes, as we've seen, that (in almost all cases) the talented could work just as hard without incentives as with them. The difference principle licences only those inequalities which benefit the worst off; hence it does not licence those inequalities which are necessary just because of the attitudes of the talented (by the first critical argument); hence, the talented cannot both accept the difference principle and hold out for the

[29] Cohen writes: " . . . if I am right, . . . the incentive argument can justify inequality only in a society where interpersonal relations lack a communal character, in the specified sense." Cohen, "Incentives", p. 353.

incentives, on pain of incoherence. Cohen concludes that, if we exclude those rare cases in which the talented could not work just as hard for less, "the difference principle can justify inequality only in a society where not everyone accepts that very principle. It therefore cannot justify inequality in the appropriate Rawlsian way."[31] Either the Rawlsian view of justification is wrong, or the incentives argument does not justify (in the Rawlsian way) the inequalities which it purports to. Note, once again, that this third argument too depends on the assumption that, in almost every case, the talented could work just as hard without material incentives.

3. What Cohen should have said

The third of Cohen's arguments has force against Rawlsians, I think – though it could be challenged.[32] But since it has this specialised application, I propose to leave it aside and concentrate instead on his other two arguments. The first of these suffers, I think, from an unhelpful way of diagnosing what is wrong with the incentives argument. I shall claim, in this section, that a better diagnosis brings out Cohen's underlying point – or at least, the point that Cohen should have made. Then in the next section I'll use this suggested diagnosis to explain the argument which invokes the idea of comprehensive justification.

Recall that the first argument operated by putting pressure on the idea of inequalities benefiting the worst-off. Cohen claimed that the appropriate way to evaluate claims that an inequality benefits the worst off, is to ask whether it is

---

[30] Cohen, "Site", pp. 8-9, 16-17.

[31] Cohen, "Site", p. 9. Cohen is sensitive to the objection that the difference principle is supposed to apply only to the organisation of the basic structure of society, not to everyday conduct within it, so that the incoherence he alleges cannot arise. The article "Site" is intended, in fact, as a response to that objection. Cohen's view is that Rawls cannot sustain a normatively significant distinction between everyday behaviour and choices, on one hand, and the basic structure of society, on the other, given that the normative significance of the idea of basic structure is supposed to depend on its denoting the factors which have the profoundest effects on the distribution of benefits and burdens in society. Cohen points out that non-(legally) coercive institutions such as the family, and behaviour within them, have very profound effects on the distribution of benefits and burdens.

[32] Most obviously, Rawlsians could object to Cohen's assumption that, in almost all cases, the talented could work just as hard without incentives. Or they could make the objection outlined in the preceding footnote, and go on to reject Cohen's response to that objection on the grounds that the basic structure does not, as Cohen alleges, expand on examination to include everyday behaviour.

necessary to enhance the position of the worst-off. He then distinguished two senses of 'necessary': necessary given the motivations of the talented, and necessary *tout court*. Now I think that Cohen is right that the idea of an inequality's benefiting the worst-off is unclear, and right also that the incentives argument trades on this unclarity. But the idea is not best clarified by reference to the concept of necessity.

The problem with the idea of benefit, or making people better-off, and hence with the difference principle and the incentives argument, lies in the implied comparison with an alternative arrangement or state of affairs. The idea of benefit is essentially comparative. It does not make sense to say of a single state of affairs, or economic arrangement, considered in isolation, that it benefits the worst-off (or anyone else). Someone is benefited by an arrangement insofar as they do better under that arrangement than they would under some *relevant alternative arrangement.*

Now Cohen's suggestion that we should interpret 'benefiting the worst-off' in terms of necessity, comes to this: the range of relevant alternatives comprises all of those which are possible. For to say that *X* is necessary (*tout court*) for *Y* is to say that there is no possible world in which *Y* obtains and *X* does not. Therefore, an incentive scheme is necessary to benefit the worst off, only if there is no possible world in which the worst-off do at least as well without that scheme as they do in some possible world with it. Cohen's suggestion that we cash-out 'benefit' in terms of necessity asks us to consider every possible world as a relevant possibility.[33]

One can see the motivation for doing this, but it casts the net too wide. The motivation is a desire to wring, from the essentially comparative idea of 'benefit', an unconditional sense in which incentive schemes may be justified. Because the difference principle, like every other principle which incorporates the idea of benefit, inherits the essentially comparative nature of that concept, the justifications which it typically gives are conditional on there being only a certain range of possible arrangements. Cohen wants to eliminate the conditionality, by expanding the range of relevant alternatives to include every possible world. But not every possibility *is* a

relevant possibility. It may be, for example, that the possible world in which the talented are totally self-abnegating, working very hard out of moral concern for the worst-off, with no material incentives, is not a relevant possibility. Cohen thinks that Rawls is right to suppose that in a just society persons would be motivated by the principles of justice.[34] But surely citizens of a just order could have other legitimate motivations, too.[35]

Whether or not the self-abnegating scenario is a relevant possibility is surely a *moral* question, which raises issues of the demandingness of morality. It is not merely a physical or metaphysical question, as Cohen seems to think, having to do simply with what the talented could do.[36] Those physical and metaphysical issues are of course pertinent to the central issue, but they do not exhaust it. (One way in which a possibility can be irrelevant, is that the talented could not act in the way described, but that is not the only way.) One should not assume in advance that it is possible to remove the conditionality of justifications employing the concept of benefit. Whether it is possible or not depends on how the moral considerations play out.

The idea of benefit which the difference principle, and hence the incentives argument, employs, involves the idea that there is a range of relevant alternative arrangements. I'm suggesting that the range is *morally* constrained, by considerations of the fairness of requiring persons to behave in certain ways and to have certain possible motivations. Cohen might reply to this that any relevant considerations of fairness or demandingness are already taken care of in his way of doing things. For the talented could not be made worse-off, all things considered by the lights of the difference principle, than anyone is legitimately required to be by that same principle. Thus, the various strains of requiring them to have certain work patterns are taken into

---

[33] There are different senses of 'necessity *tout court*', including physical necessity and logical necessity. I assume that Cohen means physical necessity.

[34] Cohen, "Site", p. 10.

[35] At one point, Cohen allows (for the sake of argument?) that the talented rich could claim a morally justifiable partiality towards their own interests, but goes on to say: "a modest right of self-interest seems insufficient to justify the range of inequality, the extremes of wealth and poverty, that actually obtain in the society under discussion [Britain]." Cohen, "Incentives", pp. 370-371. The present point is that such considerations bear on the range of relevant alternatives. Cohen effectively concedes that in this passage – even though he thinks that they do not help the incentives argument much.

account, even if, following Cohen, we take the range of relevant alternatives to be the set of all possible worlds. Any legitimate concern with the demandingness of work patterns would modify our assessment of the talented's level of advantage.

I agree that Cohen's approach has this feature. But I think that it is unhelpful to present the issue as, in the first instance, a purely physical or metaphysical one having to do with the full range of possible worlds. Presenting it in this way conceals the relevant moral considerations. It is much better to bring the moral issues to the surface, so that we can see that the range of relevant alternatives is morally constrained – and so we can see, also, how the various moral considerations are factored-in to the assessment of advantage. Cohen's second critical argument is better in this respect, since the idea of comprehensive justification makes plain that what is at issue is whether persons can justify their behaviour to each other, and not simply whether they behave in ways which are beneficial with respect to some other, narrowly conceived, alternative. Thus it acts as a kind of filter of possible arrangements, excluding those in which persons cannot justify their behaviour to each other as inappropriate benchmarks for evaluating the justice of an arrangement. In section 5 I'll return to the question of whether the idea of comprehensive justification is the right kind of constraint on possible alternatives.

So Cohen should have said that the incentives argument trades on the unclarity in the idea of benefit contained in the difference principle, because it tacitly supposes that the range of relevant alternative arrangements is constrained by facts about the actual motivations of the talented, when it is not. Some possible arrangements in which the talented do not have the motivations they actually do in our world, are relevant to the claim that any particular incentive scheme benefits the worst-off. We can express this thought in its most general form, like this:

> (C) The states of affairs relevant to moral and political argument may include some which are radically different from the actual state of affairs.

---

[36] The bearing on justice of issues of fair concern with one's own interest, is discussed in T. Nagel,

Reflection on (C) brings to light the issue, of course, of what determines the range of relevant states of affairs, or arrangements, if it is neither strongly constrained by the actual world nor, as Cohen suggests, radically unconstrained. I've already suggested that it is constrained by moral considerations, and I'll try to explain exactly what I mean by this in the rest of this chapter and the next. But we do not need to know what, in general, determines the range of alternatives which is relevant to a moral or political argument, in order to appreciate the force of Cohen's critique of the incentives argument. The incentives argument fails to consider alternatives in which the talented have motivations different from those which they actually have, and that is uncontroversially too narrow a focus.

It is worth emphasising the theoretical and practical importance of judgements about the relevant range of possibilities. In most if not all cases, political actors have insufficient power to secure their aims perfectly, and so must choose between a range of imperfectly effective courses of action. Since, therefore, courses of action will rarely be recommended on the grounds that they perfectly secure the actor's objectives, most arguments in favour of taking one course rather than another must be arguments which, explicitly or not, point to the greater effectiveness of one course *compared to* some other(s). Almost all arguments recommending a course of action must have the form of recommending one course as the most effective course in some range. That's because few if any political actors have the power to secure their objectives perfectly.[37] But the upshot is that it is crucial, in assessing almost all arguments recommending some course of action, to ask what the relevant range of alternative courses was taken to be, and why. A change in the list of relevant possibilities which is considered, may have the effect of reversing the recommendation of an argument about whether to pursue a particular course, *without changing any assumptions* about the circumstances or objectives of the actor.

Equality and Partiality, Chapter 2 especially.

We're all familiar with the practical importance of the range of alternatives amongst which we must choose in the case of menus, for example. In the case of menus, it's easy in principle to see how the range of alternatives gets determined. There is an historical story we could tell about the evolution of menus. We might be able to do the same for the evolution of policies on the political agenda in a certain state, though the causes may be more obscure. But it is far from obvious how the range of political possibilities *should* be determined. Given the admitted practical importance of considering one range rather than another, it is of great theoretical importance to try to understand the considerations which should guide us in these judgements.

4. Ideal theory contrasted with deliberation

In order to understand the considerations which we should use to judge the range of relevant possibilities, we must distinguish two modes or kinds of normative political argument. This distinction is a very important one, both for general understanding of the character of political argument, and for the argument against naturalism. For, as I indicated at the start of this chapter, the argument against naturalism is much easier (and quite different) if applied to one of the modes rather than the other.

One mode of political argument is that of *ideal theory*.[38] Ideal theory attempts to specify states of affairs or institutions which are ideal in some respect. An ideal theory of justice attempts to specify ideally just states of affairs or institutions; an ideal theory of democracy would attempt to specify ideally democratic states of affairs or (more likely) institutions.

---

[37] Rational political actors *typically* secure their aims perfectly, we might suppose, if we take that to mean that the expected cost of further activity cancels the expected benefit, at the margin. But few if any secure their aims perfectly, if we exclude from the calculation the costs of political action itself.

[38] As far as I know Rawls introduced the term 'ideal theory', in *A Theory of Justice*, pp. 8f and 245f. He identifies ideal theory with strict compliance theory, in which one is to evaluate principles of justice by imagining a society in which everyone follows those principles. I have no objection to this identification, but merely note that it is a substantive and controversial thesis. Some utilitarians, for example, would want to keep open the possibility that a perfectly just society is one where only a few people consciously endorse and act upon utilitarian principles. See S. Scheffler, *The Rejection of Consequentialism* (Oxford: Clarendon Press, 1982), pp. 44-52.

Ideal theory deals with the justifiability of states of affairs or institutions, and so its most straightforward application is in the *evaluation* of states of affairs or institutions, including actual ones. If one has a specification of ideally just institutions, one can use it to say whether an actually existing institution is ideally just, or whether a proposed institution would be. This exercise could be driven by merely contemplative interest; one may simply want to know which institutions are just, with no interest in remedying injustice. But ideal theory also lends itself to an interest in political action, and to an interest in giving reasons for acting one way rather than another. If one can show that a present state of affairs is unjust, and one can show that some action would probably bring about a state of affairs which realises the specified ideal of justice more closely, then one can provide reasons of justice for performing that action.

Ideal theory lends itself to argument about which actions should be performed, but its conclusions are not that some action or other should be performed. Its conclusions are that some state of affairs or institution, hypothetical or actual, is or is not just, or democratic, or is so to a certain degree, or except in certain respects. Ideal theory evaluates, testing the justifiability of a state of affairs or institution. By contrast, *deliberation*, which is the second mode of normative political argument, seeks to recommend one course of action rather than another. Deliberative argument is concerned with the particular circumstances of particular actors, since some action is the thing to do not just in virtue of one's aims or ideals, but also in virtue of one's circumstances. The conclusion of a piece of deliberation is that some actor should do one thing rather than another, in some situation.

The two modes of political argument are obviously related, though it is not obvious what their relationship is. The main issue is whether what the ideal state of affairs is, depends on which actions are justified. I take it that, for those who accept the very idea of a specification of ideal states of affairs or institutions, the opposite

dependence is uncontroversial.[39] But there is a good deal of room for disagreement about whether the permanent unattainability, by permissible actions, of some ideal, would serve to discredit the ideal. Is world peace part of correct ideal theory? Which actions are justifiable depends, as I said earlier, not just on what is ideal, but also on the actor's circumstances. If it is allowed that what is ideal depends on which actions are justifiable, then wretched actual circumstances could undermine what would otherwise be an ideal.

I leave this interesting issue aside. Consider the uncontroversial relation of dependence, that of justifiable actions on ideals. Deliberative political argument must somehow bring ideals to bear on the circumstances of the agent whose actions are to be guided by the argument (hereafter referred to as 'the actor'), in order to recommend one course of action rather than another. I shall adopt the following model of such argument.[40]

Deliberative argument aims to recommend some course of action as more effective than others. Hence, the actor to which it is addressed must face a range of *options*. Options are things which are practically relevant and which, the deliberator supposes, the actor could do. Thus my options vary according to the task in hand, not just according to my physical powers. Deliberation consists in part, then, of the generation of a list of options. But it must also evaluate those options for their *expected effectiveness*, and finally recommend an option on the basis of these evaluations. The evaluation of expected effectiveness is in terms of the actor's aims[41], and is made on the basis of assumptions about the future behaviour, conditional on

---

[39] An example of someone who seems not to accept the idea of ideal theory, at least in moral contexts, is Jonathan Dancy. His particularism shuns appeal to general characterisations of ideals to be aimed at by actors. According to him; moral thinking (note: not moral *argument*) involves correct description of the morally relevant features of particular situations, and not much else. It does not involve subsuming particular cases, or aspects of particular cases, under general principles, or appealing to values (or, I suppose, ideals) which provide reasons for action in the same way across a range of cases. See J. Dancy, *Moral Reasons* (Oxford: Blackwell, 1993), especially pp. 111-119. I find Dancy's views about the nature of moral thinking challenging, but I cannot address them here.

[40] Recall what I said in Chapter One, that the claims I shall make about the rational structure of political argument, are not to be interpreted as claims that political actors who do not consciously follow the pattern described are somehow faulty in their decision-making, just for that reason. Nor do they entail such claims.

[41] The actor's aims might be the aims which she takes herself to have, or those which the deliberator takes her to have, or those which right reason assigns to her. The model of deliberation presented here is compatible with each of these views.

the actor's pursuing the option in question, of the actor's environment, including the behaviour of other persons.[42] In the simplest cases, the option with the greatest expected effectiveness is recommended, though one might specify more complex selection rules (perhaps reflecting side-constraints or agent-centred prerogatives).

Consider an example of a piece of deliberation described in terms of this model. A certain actor wants to learn French, say. One of the things she must do is become clear about her aims. She may, for example, want to learn conversational French rather than written French; she may want to study in the evening or at weekends, rather than during the day; she may want to pay as little as possible. A logically separable exercise – though it is by no means always distinct in practice – is finding out one's options (practically relevant things one could do). Our actor may discover that there are two options relevant to her: a course at the local FE college every Wednesday afternoon, and a course of private tuition available any weekday evening. In order to select one of these options, she should evaluate their expected effectiveness in promoting her aims. Finally, she should pick an option according to these evaluations and according to her decision rule.[43]

Having distinguished between ideal theory and deliberation, we can restate the issue Cohen raises with greater clarity. In the context of ideal theory, as I've said, the issue concerns the range of possible arrangements which should be treated as relevant to the evaluation of some institution or state of affairs. The issue arises (at least) in every case in which the evaluation employs the concept of benefit. In the context of deliberation, on the other hand, the issue concerns the range of possible future behaviour of the actor's environment, which should be treated as relevant to

---

[42] Reliance on the idea of expected effectiveness does not make this model of deliberation inherently consequentialist. As I shall explain in chapters Five and Six, it is possible to evaluate expected effectiveness not only in terms of the predicted consequences of acts, but also in terms of counterfactual consequences.

[43] The decision rule may not be the simple one, which says that she should choose the option with greatest expected effectiveness. It may be, for example, that our actor considers herself to be subject to a moral duty not to choose private services when publicly-funded services are available. In that case, we might say that the private route is expected to be most effective in promoting her aims, all things considered, but that her decision rule includes a constraint forbidding selection of that option. My aim is for this model of deliberation to be inclusive. Those who think that any constraints worthy of respect can be interpreted in terms of effectiveness, all things considered, can adopt the model and specify the appropriately simple decision rule: choose the option with greatest expected effectiveness.

evaluating the expected effectiveness of the actor's various options. The issue arises in every case in which the actor's aims and powers are such that she cannot perfectly secure her aims with any particular option.

The two issues are certainly connected, but they are not the same. The greatest difference between ideal theory and deliberation is that the latter, but not the former, is addressed to particular agents in particular circumstances, and is to be judged according to how sensitive it is to those particularities. For that reason, we can indulge our imaginations in ideal theory with counterfactual assumptions which would look absurd or reckless in deliberation. In the following two sections, I'll explore the contrast between the two modes of argument in this respect in greater detail.

5. What determines the range of relevant alternatives in ideal theory?

The contrast between ideal theory and deliberation enables us to be much clearer about the considerations which determine the range of alternatives which should be treated as relevant to a political argument. First consider ideal theory. Since ideal theory is by its nature not addressed to the deliberative problems of particular agents, any claim that some possibility is irrelevant to ideal theory must have general force. Ideal theory can perhaps accommodate some relativity to particular circumstances: what justice requires, as a matter of ideal theory, may be different in modern Britain than in medieval Britain, for example. But even if justice admits this level of relativity, it cannot be very specific in ruling some possible arrangements irrelevant in some circumstances but not in others. We would balk at a theory of justice which declared that an unconditional basic income is a relevant possibility in late twentieth century Belgium, but not in late twentieth century Britain, for example.[44]

It is difficult, but not impossible, to establish general irrelevance. We feel the constraint on what is counted as relevant to ideal theory when we consider some

possibility to be utopian. Some possibilities really are utopian, in the sense that we should not organise our activities around the goal of realising them. It may be interesting for some theoretical reason to know about such possibilities, but they do not have practical value as ends or aims. We may be inspired by them to act, but we should not hold out for them.

It is not true of ideals in general that they are utopian. Some of the characteristics of the idea of an ideal state of affairs or institutional arrangement may be inferred from the assumptions which are appropriate to thinking about it, such as these: 'this arrangement may be very different from the way things currently are'; 'we may not be able to bring this arrangement about, in the foreseeable future'; 'people would behave differently under these circumstances'. Assumptions such as these give some content to the idea of an ideal state of affairs. They describe, in the haziest fashion, some of the features of ideal states. They correlate with the idea that we do not currently live in an ideal state. But they do not imply that we could never live in such a state, let alone that we could never realise even portions of the ideal. Ideals are not *generally* practically irrelevant.[45]

It is not clear how we should judge which proposals of ideal theory are utopian. Both ends of a spectrum of stances seem to be mistaken. Facts about human nature or motivation, for example, should not be allowed directly and straightforwardly to defeat a proposal, so long as we allow for the transformation of human nature or character to be part of the aim of political action. On the other hand, it seems equally wrong to suppose that such facts have *no* bearing on the relevance of a proposal – for, following that policy, we would be liable to describe what justice (or some other value) requires for creatures very different from us.[46]

---

[44] The idea of an unconditional basic income is discussed and defended in P. Van Parijs, *Real Freedom for All. What (if anything) Can Justify Capitalism?* (Oxford: Clarendon Press, 1995).

[45] Radicals think that often the net is not cast wide enough, in considering which possibilities are politically relevant. This thought seems to be implicit in the idea that relations between the sexes are properly described in terms of gender, rather than the biological categories of sex. It is implied that gender characteristics are more open to change than are sexual characteristics. Hence, the implication continues, more possible relations between the sexes are politically relevant. Many disputes in social science involve issues of this kind.

[46] Thomas Nagel discusses 'the problem of utopianism' in Chapter 3 of his book, *Equality and Partiality* (New York and Oxford: Oxford University Press, 1991).

One idea about the difference between proposals which are relevant, and those which are irrelevant or utopian, is that the former but not the latter are likely to be *stable arrangements*.[47] A proposed set of ideal institutions is not irrelevant just because the proper functioning of those institutions would require persons to have motivations which they do not currently have. That test would hold political ideals hostage to actual human badness. Instead the proposal is irrelevant if it would require motivations which we can't imagine humans coming to have. For then the proposal would be unstable in the sense that deviations from the ideal, supposing it were instituted in the first place, would tend to acquire momentum instead of being counteracted by stabilising forces. Judging whether a proposal is utopian is then an exercise of imagination which aims to discern whether the proposed arrangements would cause persons living under them to have appropriate motivations.[48] Judgements like this aren't necessarily arbitrary – we may cite evidence from history and the social sciences in favour of them. But they are undoubtedly difficult to make.

In this light, Cohen's underlying claim (C) is difficult to dispute. C says,

> (C) The states of affairs which are relevant to moral and political argument may include some which are radically different from the actual state of affairs.

C does not say *which* radically different states of affairs are relevant, only that some are. That bare claim is surely true.

Cohen's further claims are more controversial. The idea of comprehensive justification, and the associated idea of justificatory community, try to tell us more about which arrangements are relevant, since they impose normative constraints on

---

[47] On the idea of stability, see J. Rawls, *A Theory of Justice*, pp. 453-458 (also pp. 175-177 on the 'strains of commitment'); R. Nozick, *Anarchy, State, and Utopia*, pp. 299-300.

[48] Cohen endorses roughly this test, when he concedes that "it needs to be shown that a society of people who believe in equality and act accordingly is reproducible, that it is not fated to collapse under disintegrative strains". Cohen, "Incentives", p. 360. In a similar vein, Brian Barry proposes the following requirement of a theory of justice: it should be able to explain how its account of the motives for behaving justly is related to its account of the principles of justice. See B. Barry, *Justice as Impartiality* (Oxford: Clarendon Press, 1995), pp. 46-51.

relevance. They say that a proposed arrangement is relevant to ideal theory only if it supposes that persons behave in justifiable ways under it. This constraint is similar to the Rawlsian idea that arrangements should be judged, for the purposes of ideal theory, on the assumption of full (or 'strict') compliance.[49]

Cohen's fundamental criticism of the incentives argument is vindicated by our analysis of the considerations which determine the relevance of proposals to ideal theory. The incentives argument relies on the hidden assumption that the motivations of the talented are fixed, as far as justice is concerned. That assumption is false: it lies at one end of the spectrum of views about the bearing of facts of human nature on the relevance of ideals, and both ends of that spectrum are mistaken. We do not need to know precisely what the correct view *is* about the bearing of such facts, to reject any version of the incentives argument which relies on this assumption.

On the other hand, Cohen's claim that no version of the incentives argument could justify much inequality, does presuppose a definite view about the bearing of such facts. Or rather, it presupposes the following:

> Either (a) the facts about human motivation are such that difference principle-superior, equalising regimes are *obviously feasible*, and so relevant for that reason; or (b) the proper view about the bearing of facts about human motivation on relevance, is such that difference principle-superior, equalising regimes are relevant even though they are not obviously feasible.

Susan Hurley has disputed (a). She finds it not obvious that there is a relevant possible equalising regime which is difference-principle superior to a relevant possible inequality-permitting regime. That's because she points out the value which could be lost by people coming to have the views about responsibility which would be required for them to be motivated under the equalising regime.[50] In any case, there is surely room for disagreement, centring on (a) or (b), about which possible incentive

[49] J. Rawls, *A Theory of Justice*, pp. 245-246.

regimes and patterns of work should be judged relevant to ideal theory, and which should be judged irrelevant or utopian. Cohen's confidence that no version of the incentive argument could justify much inequality is open to challenge, even if we allow the idea of justificatory community to act as a filter on relevant possible incentive regimes.

The argument against naturalism as applied to ideal theory is quite simple. Any plausible view as to how the range of relevant alternatives is constrained, in ideal theory, must be consistent with (C). That means that the range is not picked-out, for ideal theory, by our best predictions, since the possibilities selected in such a way would be too strongly constrained by facts about the actual state of the world. But the range must be picked-out by reference to *some* (deep and persistent) facts about humans. That leaves only one candidate type of judgement, so far as I can see: the range of relevant possibilities must be picked-out, in part at least, by judgements about what persons could do.[51] We decide whether a proposal is utopian by forming some judgement about whether humans could live in the ways the proposal describes. If this is correct, then ideal theory, including ideal theory of justice, is inescapably agency-implicating.

6. What determines the range of relevant alternatives in deliberation?

Applied to ideal theory, Cohen's underlying claim (C) is surely true. But it is much more controversial if it is understood as a claim about the range of possibilities which is relevant to deliberation. That's because there is much more information available, in principle, when we're discriminating between relevant and irrelevant possibilities with a particular agent in particular circumstances in mind. Assumptions about the range of possibilities against which a proposal should be judged, which are quite acceptable in ideal theory, are obviously misplaced, or even reckless, in deliberation.

---

[50] S. L. Hurley, "Cohen on Incentives", pp. 56-58.

For example, consider the case of the kidnapper which Cohen discusses. Suppose our task is to decide whether to pay the kidnapper. It may well be reckless to judge the expected effectiveness of paying on the assumption that one relevant possibility is that the kidnapper undergoes a change of heart, and comes to accept principles of justice. In effect, to do so would be to extend the assumption of full compliance to a deliberative context. But whereas the assumption of full compliance, or Cohen's assumption that in just societies people would be able to justify their behaviour to each other, looks reasonable if applied to ideal theory, these assumptions are often clearly inappropriate to deliberation. Kant's view that one should not lie even to a murderer at the door, for example, may be understood as the view that one should apply the assumption of full compliance even to deliberation. But that view has widely been thought to be a regrettable part of his philosophy.[52]

If stricter standards are appropriate for judging a possibility to be relevant to a deliberative problem, what are they? We should distinguish between three objects of such judgements. Call the agent whose actions are the subject of the deliberation, *the actor*. One may judge (i) that a possible action on the part of the actor is irrelevant, or (ii) that a possible action on the part of some other agent is irrelevant, or (iii) that some other possible event or behaviour (of the non-human environment, for example) is irrelevant. Each may be irrelevant for two reasons.

A possible action on the part of the actor may be irrelevant, first, because it is *practically* irrelevant to the task in hand. By appeal to the criterion of practical relevance, we can eliminate many possibilities from consideration. If the actor's task is to deal with the kidnapper, then her digging the garden, say, is almost certainly practically irrelevant. Admittedly, practical relevance is not an altogether clear criterion. We can appeal to conventional patterns of action to explain it, in which case we say that an action is practically irrelevant if it is not a conventional method of

[51] Admittedly, arguments of this form must be used with care. They are prone to the discovery of some new candidate, not previously considered. But I think in this case there really are no other candidates.

promoting the actor's aims. But this is not complete as it stands, since we must allow for the invention of new patterns. Hence, something is practically irrelevant, roughly, if it is neither a conventional means to the actor's ends nor an innovative means to them.[53] This gives us a somewhat unclear, but not entirely arbitrary criterion of relevance.

The other reason why a possible action of the actor's may be irrelevant, is that it is something the actor could not do, in the circumstances. Actions may be possible for an agent, and yet not something that the agent could do in the circumstances at hand. For example, running for a bus may be a possible action for me, in general, and yet not something I could do in the present circumstances, because I have a broken leg. Or, alternatively, the action in question could be possible for an agent even if it is not usually within her powers. She may be capable of learning to do it, or of otherwise acquiring the power. As with deliberation generally, our judgements about what the actor could do must be tailored to the particular circumstances at hand.

Next consider what would make possible actions of agents other than the actor, or *third parties*, irrelevant to a deliberative problem. Again, the actions in question could be practically irrelevant (practically irrelevant, that is, to the actor's deliberative problem), and the notion of practical relevance is as much use here as it is with respect to the actor. If we leave this aside, the remaining ground of irrelevance is the expectation that the third party will not in fact behave in that way. This is what grounds resistance to Kant's view about the murderer at the door: it is reckless to

---

[52] On Kant's view, see C. M. Korsgaard, "The Right to Lie: Kant on Dealing with Evil", *Philosophy and Public Affairs* 15 (1986), pp. 325-349. For further discussion of Kant's views, and comparison with the stance on these issues for which I argue, see Chapter Six below.

[53] In simple cases, Hobbes's account of 'regulated thoughts' provides an adequate model: "In sum, the discourse of the mind, when it is governed by design, is nothing but *seeking*, or the faculty of invention, which the Latins called *sagacitas*, and *solertia*; a hunting out of the causes, of some effect, present or past; or of the effects, of some present or past cause." T. Hobbes, *Leviathan*, edited by M. Oakeshott (Oxford: Basil Blackwell, 1946), p. 15. If our object is to light a stove, we need to imagine causal series which begin with some action of ours and end with the stove's being lit. But in more complex cases, reasoning about means cannot be sharply distinguished from reasoning about ends. David Wiggins gives the example of the aim of having an entertaining evening. Consideration of the various options (such as *going to the cinema*, *going to the restaurant*, and so on), is at once (i) consideration of means and (ii) further reflection on the nature of the end – on *what counts* as an entertaining evening. D. Wiggins, "Deliberation and Practical Reason", in his *Needs, Values, Truth* (Oxford: Basil Blackwell, 1987), p. 225. We may be interested, in other words, not just in the causal relation of some action of the actor's to the desired end, but also in the constitutive relation between the two.

decide whether to lie to the murderer on the assumption that he might not put the information to bad use, if all the evidence indicates that he will put it to bad use.

One main ground for judging some possible behaviour of third parties to be irrelevant to a piece of deliberation, then, is that we do not expect the person in question to behave that way. Hence, we use predictions of the third party's behaviour, conditional on the actor's choice of option, to rule-out possible behaviour of that party as irrelevant. On any understanding of these matters, such conditional predictions are very important in discriminating between relevant and irrelevant possibilities. In Chapter Five, however, I will argue that, even if they are certain, such predictions are not always sufficient grounds for judging a possibility to be irrelevant. Sometimes we should act on the assumption that third parties might behave in ways which, we predict, they will not.

The third category is relatively unproblematic. Again, some possible behaviour of the non-human environment of the actor may be *practically* irrelevant. But there seems no room for dispute that the remaining appropriate grounds for judging it to be irrelevant are simply that we believe it will not in fact behave that way. If all our best predictions tell us that a machine will not behave a certain way, there are no grounds, analogous to those which, I claim, exist in the case of third parties, for deliberating on any alternative assumption. The range of relevant possible behaviour of the non-human environment is straightforwardly picked-out by our best predictions about its behaviour. (In the next chapter, I shall explain this asymmetry between other persons, on one hand, and the rest of the actor's environment, on the other, as having to do with the possibility of *joint action* with other persons.)

Now consider Cohen's fundamental claim (C) in application to a deliberative problem. Instead of asking what justice requires in the mode of ideal theory, we are to deliberate about the behaviour of some taxing authority, which is faced with two options: *no change*, or *raise the tax rate* (to some particular level). Suppose that we are Rawlsians, and that both of the options satisfy Rawls's first principle of justice, so that they are to be judged only as regards the difference principle. Suppose, finally,

that our best predictions tell us that the talented will work less hard at the higher rate of tax, to a degree which is sufficient to make the worst-off less well-off than they currently are. Which option should we recommend?

Cohen's insistence that we should treat as relevant possibilities which are radically different from the actual state of affairs, looks less pertinent in a case such as this. A taxing authority which is trying to apply the difference principle in these circumstances, it appears, should not change the tax rate. Our best predictions of the behaviour of the talented seem to show that the possibility of their working just as hard for less is an irrelevant one. We believe that they won't do so, and our best predictions are usually sufficient grounds for ruling a possibility to be irrelevant. The incentives argument, in its deliberative form, seems to work – even if we agree with Cohen that it doesn't work as an argument of ideal theory.

We may be uncertain in our prediction of the behaviour of the talented. Perhaps we are unsure whether they would work just as hard for less, given the right encouragement and so on. In these cases, the behaviour of the talented is *epistemically open*: our knowledge of their future behaviour is deficient. Hence, it may be reasonable to suppose that they might do a range of things, and so a range of possibilities may be relevant to the deliberative problem. On these grounds, the difference principle may give us some reason to raise the tax rate. But these reasons seem to disappear as our knowledge increases – until, with perfect knowledge, the incentives argument seems to acquire irresistible force.

Note that one way in which our predictions may become more certain, is by the talented having a stronger resolve not to work so hard without incentives. On the face of it, it is a bit odd that their becoming more strongly resolved to hold out for incentives can remove our reason for withholding those incentives on grounds of justice. One could attempt to dispel the feeling of oddness, by saying that the incentives argument, which recommends granting the incentives if our best predictions really do tell us that to remove them would harm the worst-off, provides

in these cases only reasons of prudence, not reasons of justice.[54] But that is a bit strained. The difference between the incentives argument as applied to ideal theory, and the same argument applied in deliberative contexts, is not a matter of principle, or of the character of the reason-giving ideas involved in each. It has to do instead with the aim of the argument in each case. In one case we want to say what kind of world or institutional arrangement would best satisfy the principle, taking as constraints only the most general, immutable facts about human nature and social organisation; and in the other we want to say which action would best satisfy *the same principle*, taking as constraints the circumstances of the actor concerned. So the re-description of the deliberative problem in terms of reasons of prudence is unconvincing. However, in Chapter Five I'll argue that the reasons of justice for removing incentives do not disappear in cases where we are certain that the talented would not work so hard at the higher rate, despite appearances.

There is a common assumption about the relationship between ideal theory and deliberation, which is easy to make but which may be dangerously misleading in practice. It is often said that the point of ideal theory is to sketch where we should aim to head, and that we do best by trying to travel directly from here to there. It's not immediately obvious what that injunction means, but one natural way of understanding it is as claiming that it is always a good thing to realise more rather than less features of the ideal world. But in unjust circumstances, we may do badly if we follow this advice. For example, we may harm the badly off, if we seek to institute a part of the ideal world by denying the talented incentives. Ideal theory should not be understood as the first stage of deliberation, the second stage being conceived as plotting a course from the actor's circumstances to the ideal. It should be conceived instead as a separate exercise, with intrinsic interest. It is directly applicable to the evaluation of actual states of affairs or institutions, but it does not issue directly in

[54] See C. Bertram, "Principles of Distributive Justice, Counterfactuals and History", *Journal of Political Philosophy* 1 (1993), p. 225.

recommendations for remedial action. That's because the possibilities which are relevant to ideal theory may well be irrelevant to deliberative problems.[55]

7. Conclusion

Cohen's critique of the incentives argument consists of a fundamental claim (C), plus further claims. C claims only that the possibilities which are relevant to moral and political argument may include some which are radically different from the actual state of affairs. Cohen's further claims take a stand on *which* other possibilities are relevant to the incentives argument, and these may be disputed without challenging C itself.

C raises the general issue of what makes a possibility relevant to a political argument. This is really a very general issue, applying not just to the incentives argument, nor only to arguments about justice – but to any political or moral argument which implicitly or explicitly makes a moral comparison between alternatives. C is simply the denial of one extreme stance on this issue, which claims that the actual state of the world places a very strong constraint on the range of relevant possibilities. Cohen's further claims often assume the converse view, which is that something is a relevant possibility so long as it is a possibility. I argued that both of these extreme views are wrong.

Whether or not some counterfactual state of affairs, institutional arrangement, action, or event, is a relevant possibility, depends on whether we are engaged in ideal theorising or in deliberation. Ideal theory by its nature is not (whereas deliberation is) concerned with the particular circumstances of particular agents, and so the grounds for judging something to be irrelevant are more slender in ideal theory than in deliberation. As a corollary, C (together with Cohen's other claims, such as the claim that justificatory community is a feature of just societies) is less controversial as a doctrine about ideal theory than it is as a doctrine about deliberation. The importance

[55] The assumption that ideal theory lends itself straightforwardly to deliberation is criticised in M.

of the distinction between ideal theory and deliberation is shrouded, to some extent, by the fact that we often deliberate with very imperfect information. But in cases where we have very good predictions of the behaviour of third parties, C may seem false. Only those possible future actions of third parties which, we predict, they will actually perform, seem relevant to deliberation. The incentives argument may escape Cohen's criticisms in such cases. Cohen himself seems to accept this.[56]

The argument against naturalism as applied to ideal theory is straightforward, I think. What justice requires depends, as a matter of ideal theory, on which patterns of work are relevant possibilities, and this depends on whether the talented are unable, or merely unwilling, to work just as hard for less. It depends on which possibilities are within their reach, in an agency-implicating sense. In the next chapter, I shall attempt to show that naturalism is false even when it is applied to deliberative arguments.

Phillips, "Reflections on the Transition From Ideal to Non-Ideal Theory", *Nous* XIX (1985).

[56] Cohen, "Incentives", pp. 325-326.

Chapter Five

The Rationality of Acquiescence

In Chapter Four I argued that agency-implicating judgements about what persons could do are essential to determine what justice requires. Following from Cohen's discussion of incentives, and in particular from his claim that the incentives argument fails to justify much inequality, I explored the issues raised by judgements about which alternatives are relevant to moral and political argument. I made a distinction between two modes which such argument may take: those of ideal theory and deliberation. Cohen's criticism of the incentives argument is straightforwardly cogent, I claimed, if applied to an ideal-theoretic version of that argument, but is much more controversial if applied to deliberative versions of it. In particular, it seems to fail in cases where there is no dispute about the normative premise of the incentives argument, and our best predictions really do tell us that the talented would not work so hard for less.

In this chapter I seek to explain how Cohen's critique might apply even in these hard cases. The argument turns on the rationality, or otherwise, of a certain attitude towards acquiescent options. I shall distinguish four kinds of worry about acquiescence. Three of these are easy to rationalise, whilst the fourth kind, which I shall call 'pure worries' about acquiescence, are more difficult to account for. But I shall argue that pure worries *are* rational in some circumstances. In order to explain their rationality we must abandon a view about the nature of deliberation which many of us take for granted; we shall have to modify our view of the appropriate way to evaluate courses of action.

As I've indicated, this conclusion is much more controversial than the conclusion of Chapter Four, and the argument for it is more difficult to make. To some extent, we leave our particular concern with justice behind – since the rationality of acquiescence extends further than the realm of justice in holdings. We shall be concerned with fundamental issues about the nature of deliberation. (We shall be concerned with these in the next chapter too.) But the application of these arguments to the topic of justice remains particularly important. For in arguing that pure worries may be rational, I shall be arguing that the naturalising urge to make political argument independent of agency-implicating claims, is misplaced not just in ideal theory, but also in deliberation.

1. Four kinds of worry about acquiescence

The first stage of the argument is the identification of the different kinds of worry about acquiescence. Imagine that some decision must be taken on the legal provision for parental leave from work. Suppose that the present situation is one of moderate statutory provision for maternity leave, with provision for paternity leave left to the discretion of employers. The actor considers a range of options, which include: (i) no change; (ii) increasing maternity leave from moderate to high; (iii) equalising maternity and paternity leave, so that both are moderate and statutory.

Suppose it is true that, according to our best predictions, the fathers and employers will behave unjustly in such a way as to make the ranking of options as follows:

Best option: (i) No change
Next best: (ii) Increasing maternity leave
Worst: (iii) Equalising maternity and paternity leave

Assume however that (i) and (ii) are second best options, which beat (iii) only because it is expected that a sufficient number of the employers and fathers would behave as they should not, were (iii) to be chosen. It is predicted that a sufficient number of employers would break the law, or exploit loopholes in it, to leave women in fact disadvantaged in the labour market; it is predicted also that a sufficient number of fathers would shirk their responsibilities, leaving the shortfall of childcare to be made up by paid extra help, or more unpaid maternal care, or both.

The details of this imagined case may or may not be realistic. The important point, however, is that an otherwise favoured option, here option (i), may be thought to have reasons against it from considerations of acquiescence. Option (i) seems wrong, because the calculation of expected effectiveness comes out in its favour only because a proportion of fathers and employers are expected to behave as they should not. This fact favours that option, but also taints it. In such a case, we may feel some resistance towards choosing the apparently most favoured option, because we do not want to acquiesce in unjust behaviour.

The general notion of resistance to acquiescence may cover a number of more specific concerns. Three may be distinguished which are quite straightforward to explain. The first of these is the idea that the apparently most favoured option looks best only because of insufficient ingenuity in thinking of ways to change the background conditions for the other person's predicted behaviour. In the case of income tax and the talented, one might worry about maintaining incentives on the grounds that there could be ways of changing the attitudes of the talented, so that it would be false that they would not work just as hard without them. It may be, for example, that our best predictions about the behaviour of the talented assume that prevalent views about responsibility will continue to hold in the future, so that people continue to feel themselves responsible for at least part of what flows from their talents. Perhaps the legislating authority should count attempting to change these views as one of its

options.[1] Similarly, one might worry about recommending option (i) (no change) in the parental leave case, for the reason that there might be some chance of influencing the behaviour of the recalcitrant fathers and employers.

This first kind of worry about acquiescence is directed at the specification of options for the actor. The thought is that the evaluation of the list of options is defective because it considers too few options. More options are available than at first sight, because the predictions of the behaviour of other persons are conditional on certain background conditions remaining constant. If the actor can do something to change these background conditions, then she might be able to change the way in which the others are expected to behave. The others' behaviour is open in the sense that our best predictions of their behaviour portray that behaviour as being conditional on factors which may lie within the actor's control. Attention to these factors may lead to an expanded set of options.

The second kind of worry about acquiescence is directed not at the specification of the list of options, but at the way in which each option is evaluated. One may worry about the *institutionalising effects* of choosing an otherwise favoured option, such as the option of retaining present parental leave arrangements. That option is favoured because of the attitudes of a certain proportion of fathers and employers; these attitudes are wrong, and the favoured option will predictably reproduce them. This kind of worry about acquiescence concerns the correct relative weighting of short-, mid-, and long-term consequences of a course of action. One may worry that mid- and long- term consequences of a policy, such as retaining incentives, or retaining the present arrangements for parental leave, have been under-weighted with respect to the short-term consequences, which seem to favour that option. This worry, moreover, may be independent of worries about the uncertainty of distant

[1] The importance of attitudes about responsibility to issues about incentives and inequality is discussed in S. L. Hurley, "Egalitarianism without constitutive luck: Incentives and Responsibility", unpublished manuscript. Nagel discusses related matters, in T. Nagel, *Equality and Partiality* (New York and Oxford: Oxford University Press, 1991), Chapter 10.

outcomes.[2] I might be quite sure that certain long-term outcomes would follow a certain policy, but unsure whether I had given these the right weight against equally sure nearer outcomes.

The third kind of worry about acquiescence is directed towards the aims themselves, or towards the constraints which the actor adopts. One may be unsure whether one has got the right project. In the parental leave case, reflecting on the actual attitudes of the fathers and employers may make one wonder whether the task in hand is simply one of reviewing parental leave arrangements, or really a broader project of reviewing work practices and relations between the sexes. Similarly, reflecting on the talented's resolve not to work just as hard for less, may make one review one's adherence to the difference principle as the only relevant normative commitment.[3] If one comes to think that the task in hand is not simply to arrange inequalities so as to benefit the worst off, but also to encourage people to have good characters, one may formulate a more complex normative premise.

All three of these kinds of worry about acquiescence are perfectly compatible with our usual thoughts about the nature of deliberation. None of them calls for special explanation. The first may lead to the rejection of the conclusion of a practical syllogism by undermining the minor premise. It may do this if the deliberator comes to think that certain background conditions of the erstwhile predicted behaviour will no longer hold, because the actor can do something to change them. When we reflect about the range of things the actor could do, we see that it includes some action which invalidates the predictions about the behaviour of others we were relying on, because it upsets some of the background conditions.

---

[2] Parfit discusses the rationality of different attitudes towards time, in Chapter 8 of *Reasons and Persons* (Oxford: Oxford University Press, 1984).

[3] Cohen entertains the possibility of the electorate doing just this, at p. 305 of "Incentives, Inequality, and Community". He writes, "the poor . . . might believe that the minor premise [of the incentives argument] is true, they might notice that its truth reflects the insistence of the rich on an unusually high standard of life and work, and they might not want to condone that insistence but to resist it, even at the cost of their own material self-interest. They would then reject the major premise of the incentive argument."

The second kind of worry about acquiescence also may undermine the minor premise. It may do this by leading to a more complex evaluation of the effectiveness of the options, which differentiates between the short-, mid-, and long-term effects of each option. This calls for a more complex prediction, and so the minor premise may be rejected because it no longer reflects our best predictions. For example, suppose that worrying about the institutionalising effects of retaining incentives leads the deliberator to attach more weight than she had to long-term consequences of income tax policy. She now needs to consider whether the prediction that the talented would not work so hard without incentives reflects the long term expected consequences of the various different options. It may be that removing incentives would have quite different long-term than short-term consequences, in which case the minor premise may be rejected.

The third kind of worry about acquiescence may undermine the major premise, rather than the minor premise, of a practical syllogism. The major premise may be undermined if we reflect on the unjust behaviour or intentions of third parties, since we may come to adopt a wider project, or to adopt a more complex set of aims for the existing project. Instead of being concerned to arrange parental leave in the best way possible, for example, we may become concerned to do that *and* to promote sexual equality.

There is, however, a fourth category of worry about acquiescence, which is difficult to account for. *Pure worries*, as I shall call them, cannot be explained as directed towards the expected actual consequences of an action, as are the first two kinds of worry, or as directed towards the specification of the task or the aims governing the task, as is the third kind. They are left over when one is satisfied that one has predicted the consequences correctly, and specified the practical problem correctly. One just does not want to go along with the other person's will, and one feels that choosing a course of action which is most favoured just because that person has a certain will, is a way of going along with it. One feels that it is preferable to choose a course of action which would better realise one's

aims only if the other person does not behave as he is expected. In other words, one feels it is right to do the thing which, one predicts, will not secure the best outcome.

Pure worries about acquiescence look irrational, and acting on them looks reckless, since it is likely to lead to an outcome worse than one which was available to the actor. Someone who chooses a course of action in the knowledge that it will lead to a worse outcome than another which is available to her is, we are likely to say, bloody-minded or intransigent, or irrational. The theoretically interesting thing about pure worries, therefore, is that if we could make sense of them, and show them to be rational, we would have to make adjustments to our picture of practical rationality.

2. Could pure worries be rational?

Consider, in a purely speculative mode for the moment, how pure worries about acquiescence could possibly be rational. Such worries are not directed towards the specification of the actor's options; nor towards the possible under-weighting of long-term, institutionalising effects of certain courses of action; nor towards the description of the agent's ends or practical task. Nor can they be explained as based on doubt about the truth of the predictions on which calculations of expected effectiveness are made, since in the cases we've been interested in, we've assumed that the predictions may be entirely certain. What could they be directed towards?

One answer is that they could be directed towards the very *use* of predictions of other persons' actual behaviour as the appropriate method to calculate expected effectiveness. Without as yet looking for a justification of the view that expected effectiveness should not, in these cases, be calculated using our best predictions of actual behaviour, we can see how that view could make theoretical sense of pure worries. To see this, consider how such a doctrine might enable us to reject the conclusion of the deliberative form of the incentives argument:

Economic inequalities should not be removed if they make the worst off people materially better off;

When the top rate of tax is 40 percent, (a) the talented rich produce more than they do when it is 60 percent, and (b) the worst off are, as a result, materially better off;

Therefore, the top tax should not be raised from 40 percent to 60 percent.[4]

Remember that we noted, in Chapter Four, that the deliberative form of this argument (as contrasted with the ideal theoretic form) seems to retain its force despite Cohen's critique, if we suppose that there is no quarrel with the major premise, and we assume that the predictions embedded in the minor premise are correct. Such assumptions describe, it seems, the point at which Cohen's critique of the incentives argument ceases to apply. For, if we make these assumptions, we seem to be granting both premises of a valid argument.

But if we suppose that the expected effectiveness of raising the tax rate (as compared with keeping it at 40 percent) should not be calculated on the basis of our best predictions of the future behaviour of the talented, then we have a way of challenging the conclusion even granting the major premise, and granting that the predictions embedded in the minor premise are correct.[5] We can describe the challenge in two ways. We can say that part (b) of the minor premise does not follow from part (a), and may indeed be false, since what will actually happen, according to our best predictions, is not what is at issue in calculating expected effectiveness. Or alternatively we could say that part (b) of the minor premise follows from part (a), but

[4] See Cohen, "Incentives", p. 339. My statement of the incentives argument is a modification of the one which Cohen gives at this point. As Cohen states it, it is arguably invalid, since the major premise refers to the justifiability of an inequality – that is, a state of affairs, or perhaps an institutional arrangement – whilst the conclusion recommends a certain *action*. This is just one example of the distinction between evaluation of states of affairs or institutions, and deliberation aimed at recommending some action, being blurred.

[5] The method for calculating expected effectiveness which I shall recommend does involve the use of predictions, in a sense. It uses them to estimate what would happen if other people were to behave in ways which, we predict, they will not. So when I refer to 'using our best predictions', this qualification

that the sense of 'benefit' at issue in the minor premise is not the same as the sense of 'benefit' at issue in the major premise, so the argument equivocates. Either way, we can concede that the normative premise is correct, and agree that the talented will behave in the way the minor premise claims, and *still* deny the conclusion, that the tax rate should not be raised.

Whichever way we describe it, the criticism rests on the idea that predicting what will happen is not, at least in some cases, the appropriate method for calculating expected effectiveness. I shall be trying to explain and justify this idea in later sections, so it is worth conceding straightaway a clumsiness in expressing it in the language of 'expected effectiveness'. Such language conveys, on ordinary understanding, an interest in the actual consequences of actions. And, by definition, our best predictions are our best guide to the future actual consequences of actions.[6] So the language of expected effectiveness seems to cut against the idea we want to express, which is that our best predictions are not the appropriate basis, in some instances, for evaluating options.

I think this reflects, in part, the depth at which the assumption that our best predictions are the appropriate basis for calculating the value of different options, is buried in our thoughts. Yet it must be possible to suspend and examine the assumption in question. In any case, I have been unable to think of a suitable alternative expression. So I propose to use 'expected effectiveness' in the following artificial sense: the *expected effectiveness* of an action is the contribution we think it would make to promotion of the good, whatever our conception of the good is. Once we understand the term in this way, we should be able to see the bare intelligibility at least of the idea that our best predictions may not be the appropriate basis in all instances for calculating expected effectiveness.

---

should be borne in mind: it is the use of our best predictions about the *actual future behaviour* of third parties which I am criticising.

[6] 'Best' in terms of the reasons we have for believing any suggested indicator of future consequences. It may turn out, quite fortuitously, that some other indicator is more reliable in fact. What is true by definition is that our best predictions will be what we have reason to believe, all things considered, about the future.

If pure worries about acquiescence are sometimes rational, it is because we have too simple a view about the appropriate way to calculate the expected effectiveness of an actor's options (so I shall suppose). I've tried to explain how an unorthodox view about expected effectiveness could, in principle, rationalise pure worries, by showing how such a view could support the formal extension of Cohen's critique of the deliberative incentives argument to the hardest cases, where we suppose no quarrel with the predictions or the normative premises. The implications of an unorthodox view would not, of course, be restricted to the incentives argument in any of its forms. They would extend as far as our reliance, in practical (including political) reasoning, on the idea of *an action's contribution to promotion of the good.* All we've considered so far is the theoretical significance of the idea that the contribution an action can be expected to make to the good, is not appropriately calculated using our best predictions; and there is nothing in that abstract consideration to differentiate the incentives argument from any other piece of deliberation.

To aid discussion, I shall refer to two views about the appropriate basis for calculating expected effectiveness, which I shall call, respectively, the *hard-nosed view*, and the *more-inclusive view*. The hard-nosed view claims that:

(HNV) Insofar as we rely on assumptions about the possible behaviour of other persons to calculate an option's expected effectiveness, we should use our best predictions of their actual future behaviour.

In contrast, the more-inclusive view claims that:

(MIV) Sometimes we should calculate expected effectiveness on the assumption that people will behave in a way which is inconsistent with our best predictions about their actual future behaviour.

The more-inclusive view asserts, and the hard-nosed view denies, that unpredicted behaviour may be relevant to the assessment of expected effectiveness.[7] But what exactly does this mean?

Behaviour may be 'unpredicted' in two ways: it may not feature in our predictions; or its non-occurrence may be featured in our predictions.[8] That is, our predictions may have nothing to say about its occurrence, leaving its occurrence quite compatible with them, or they may claim that it will not occur, in which case its occurrence obviously would be incompatible with them (strictly, incompatible with their truth). More-inclusive reasoning treats both kinds of unpredicted behaviour as potentially relevant to the assessment of expected effectiveness. The former kind's potential relevance is, however, less controversial than the latter kind's. It is uncontroversial, I take it, that when deliberating we should keep an open mind about the possibility that someone will do something which is practically relevant, but whose occurrence or non-occurrence we have not thought about, and made no prediction about.

It is the second type of unpredicted behaviour whose relevance is controversial. The more-inclusive view claims that a piece of behaviour may be relevant even if our best predictions rule it out. Moreover, this claimed relevance has nothing to do with uncertainty in our predictions. We may be absolutely sure that someone will not do *X* at *t*, and yet still count their doing *X* at *t* as relevant to the evaluation of another person's options.

Much depends on what we understand by 'relevant' here. There are some kinds of relevance which are not in dispute between the hard-nosed view and the more-inclusive view. A hard-nosed reasoner may treat the possibility that the talented could work just as hard at the higher rate, as relevant to the assessment of the option of keeping the top tax rate high, in the following sense: the fact that this is possible

[7] The more-inclusive view is inconsistent with the hard-nosed view. (The hard-nosed view is of the form: 'No unpredicted behaviour is relevant . . .', whilst the more-inclusive view is of the form: 'some unpredicted behaviour is relevant . . .'.) Therefore, the argument for the more-inclusive view is at the same time an argument against the hard-nosed view, and thus, against the naturalising strategy.

for them, in some sense, activates our concern with the institutionalising effects of incentives regimes on attitudes to work. If it was impossible for the talented to work so hard at the higher rate, there would be no scope for concern about institutionalising effects; since it is not impossible, we may worry that the long-term harm done by institutionalising the expectation of incentives, outweighs the short-term benefit of increased production redounding to the worst-off. Thus, for purely hard-nosed reasons, the possibility of the talented working just as hard at the higher tax rate may be relevant to assessing the expected effectiveness of the various options.

In order to get at what is controversial, we must distinguish between two components of the evaluation of expected effectiveness. On one hand is the contribution which the agent's aims and other commitments make to the evaluation.[9] On the other, is the contribution made by assumptions about what will or could happen. The first kind of issue has to do with our practical task and goals. Thus our assessment of the expected effectiveness of consulting a General Practitioner, as against consulting an Osteopath, will depend on whether we want a routine diagnosis or investigation of a back problem. The second kind of issue concerns our expectations about others' behaviour. Supposing our task and aims held constant, different assumptions about the behaviour of others will generate different expectations of effectiveness. Such expectations are the product of these two factors.

The claim which is controversial, then, is that it may be appropriate to assume that others would behave in a way which, we predict, they will not, when assessing expected effectiveness. This claim enables us to challenge the conclusion of the deliberative form of the incentives argument even whilst granting its normative premise, and the correctness of the predictions embedded in the minor premise – since, as I've explained, we deny that our best predictions are the appropriate basis for calculating expected effectiveness. In that way, we can see, from a purely abstract

---

[8] I shall not discuss the complications which arise when predictions have a probabilistic form. I do not think these complications change the underlying argument.

[9] Once again, I leave aside tangential issues about what those aims and commitments might include. They may be her actual desires, or the aims she would have were she better informed, or the aims

perspective, that the more-inclusive view may help explain pure worries about acquiescence. But we should like further explanation of the supposed connection between pure worries and the more-inclusive view.

The connection is this. When we are worried about acquiescence in the pure way I've tried to describe, we resent granting significance to another person's will, intention, or disposition. They are set to do something which leaves us with only second-best (or worse) options, and what they are set to do is unjust or immoral, or otherwise unreasonable.[10] The unreasonableness of their will makes them recalcitrant in the face of reasons, and, though in one sense we would do well to take note of the fact that they are set to act in this way, in another sense we do not want to grant significance or weight to it. It's as if taking note of it, when we deliberate, pays it more respect than it deserves, and even increases the injustice of the situation – since we must not only accommodate our *actions* to it, but also our *reasons*.

The more-inclusive view, meanwhile, licences disregard of predictions about other persons' future behaviour in calculating the expected effectiveness of some options. Amongst the class of all of our best predictions about the future behaviour of others, is the set of predictions about their behaviour which we base on beliefs about their intentions, will, or dispositions. So the more-inclusive view licences disregard of information about other persons' wills (and so on), which answers to the intuitive concerns of pure worries about acquiescence.

There is a plausible connection between pure worries about acquiescence and the more-inclusive view, though noting it does not amount to providing an argument for the latter. I shall explain very shortly how I intend to do that. But first it is worth noting a source of possible confusion in the discussion to follow. I shall be discussing different views about the appropriate method for evaluating the expected effectiveness of courses of action. It is easy to confuse this issue – how to evaluate

---

commanded by reason, and so on. And the aims might or might not be supplemented by side-constraints.

[10] Situations of potential acquiescence thus have three features: (i) the actor's options are deficient, in the sense that none of them offers first-best results; (ii) this is because of the intention or disposition of another person; (iii) what the other person is set to do is unreasonable in some way.

courses of action – with a separate, but more widely-discussed issue, which concerns how we should determine the unit of agency when evaluating actions.[11] That's because evaluations of effectiveness based on the more-inclusive view may involve similar counterfactual assumptions about the behaviour of other persons, as do some forms of collective consequentialism. In each case, we suppose that there are things that other persons could do which are relevant to evaluating different options.

Moreover, in motivating the more-inclusive view I shall invoke the possibility of *joint action* (see section 4 below). So it may seem that I'm really discussing unit of agency issues, as usually understood, after all. But discussions of the unit of agency issue operate on the assumption that, no matter how small, *there is some chance of collective action occurring*; in contrast, I use the idea of joint action to account for the actor's reasons not to acquiesce even in cases where there is *absolutely no chance* that joint action will occur. It is this which separates the issues I will discuss, from the issues typically addressed when the unit of agency is discussed. I shall say more about this in discussing the similarities between more-inclusive reasoning and variant forms of consequentialism, in the next chapter.

The hard-nosed view and the more-inclusive view take opposed stances on a very abstract issue. The issue is this: what range of possible behaviour of others is relevant to evaluating the expected effectiveness of an actor's options? The hard-nosed view says that the range is picked-out by our best predictions about their future behaviour, whatever those are. I think that this answer is deeply entrenched in much of our thought, which is why the issue on which it takes a stand is rarely raised. The more-inclusive view is a genuine alternative, however. What's difficult is to see how we could argue in favour of it. What would settle the issue about the range of possible behaviour which is appropriate to evaluating effectiveness?

My approach will be to examine the *significance of others* for us as deliberators – what it is about other persons, that makes them worthy of note for someone deciding how to act. A doctrine about the range of possible actions of others

[11] For discussion of the latter issue, see S. L. Hurley, *Natural Reasons*, pp. 136-148.

which is relevant to evaluating the expected effectiveness of an option, should reflect our views about how other persons can be significant for us as deliberators. One may put this the other way around equally well. Indeed, there is only a narrow gap between thinking about the *significance* of others, and thinking about which possible actions of theirs are *relevant* to assessing effectiveness.

There is some gap, however, so that trying to get our thoughts about each to support our thoughts about the other, is not just a case of trying to pull oneself up by one's bootstraps. In particular, we can ask and argue about how other persons can be significant for deliberators whilst leaving aside thoughts about the range of possibilities which are relevant to assessing effectiveness. In that way, independent light can be shed on the latter issue. I shall claim that other persons can be significant for us as deliberators in two distinct ways, and that the hard-nosed view takes account of only one of these.

I mean to discuss the significance of others for us as deliberators in a rather special sense. Others can be 'significant' for a deliberator in a number of senses. We are used to thinking about this in moral and political philosophy, if at all, in terms of the significance which other persons can have for us as sources of value, originators of obligations, bearers of rights, potential beneficiaries of care, and so on. These things have to do with the aims we might formulate as deliberators, and the constraints on our action we might judge to exist. These are very important matters, but they leave something out. Once we have formulated our aims, and decided the moral or political constraints on how we may pursue them, there is typically still some practical reasoning to be done. We still have to formulate possible courses of action, and choose between them on grounds of their effectiveness in promoting our aims, whatever those are. I described these further processes of practical reasoning in Chapter Three, as the specification of *options*, and the *evaluation of their expected effectiveness*. At this stage in deliberation, we have to think of ways in which we could act to promote our aims, and assess how effective these courses of action might be.

Of course, these two stages of practical reasoning are not likely to be chronologically or logically ordered in this simple way. Often, reflection on what may be done to promote some already formulated aims may throw up some further consideration which leads to reformation of the aims themselves.[12] This does not detract from the present point, however, which is that even the final specification of aims is very unlikely to bring with it a full specification of what needs to be done to realise them. Rather, once the aims have been settled, some further consideration will need to be given as to how to realise them. It is the significance of others at this stage which I mean to discuss.

I shall claim that other persons can be significant at this stage in two ways. The first is the most obvious, and concerns the effect which their behaviour may be expected to have on the success of our actions. What they do is likely to help determine whether I can realise my aims – just as this will depend also on how the non-human parts of my environment will behave. It is, I hope, uncontroversial that other persons have this kind of significance for us. What may be controversial is my claim that there is a second kind of significance, which has to do with the patterns of behaviour which are open to myself and the other persons jointly. I think that the availability of such patterns of behaviour can give us reasons for action which cannot be accounted for by those who hold the hard-nosed view.

3. The actor's vulnerability to others

The most straightforward way in which others can be significant for us as deliberators has to do with the expected effects of their future behaviour on the success of our actions. In a very simple case, the success of my action will depend on what you do.[13] This makes me interested in your future behaviour, at least insofar as I am interested

[12] Wiggins makes this point, at pp. 225, 232-233, and 237 of "Deliberation and Practical Reason", in his *Needs, Values, Truth. Essay in the Philosophy of Value*, Aristotelian Society Series Volume 6 (Oxford: Basil Blackwell, 1987).

[13] Someone else can affect the success of my actions in two familiar ways. He might interfere with what I do, or he might help determine the success of what I do in a different way, by evaluating it.

in the success of my action. I will want to estimate what you will do, so that I can plan my action accordingly.

I call this aspect of the significance of others, the actor's *vulnerability* to them. In using this term, I do not mean to suggest that other persons are usually, let alone always, hostile, or disposed to thwart our actions. Obviously, other persons can help as well as hinder. Rather, the term is meant to evoke a certain sense in which each actor stands apart, in her first-person view, from the rest of the world. Let me try to explain this.

Consider first the appropriate way for me to deal with the fact that what you do will affect the success of my actions. Remember that I've assumed that I've already decided what my aims are, and what moral and political constraints there are on my action. It seems to me that the best thing that I can do is to try to predict what you would do if I chose each option, and then, provided that I think my predictions are fairly reliable, choose the option which would best realise my aims if you do what I predict you will. In this respect, you are just like the rest of my local environment. That too will have some affect on the success of my actions, and the best thing for me to do is to predict how it would behave if I chose each option, and then decide accordingly.

The reason you are just like a clock, or a switch, or a car, is not because I'm treating you as a means rather than an end. At least, that isn't right just as it stands. I've already treated you as an end, in the earlier stage where I decided what my aims are, and considered the things which I may not do in pursuing them. Now I have to work out how best to try to realise those aims within those constraints. In a certain sense I am now treating you as a means, since I am wondering how to manipulate your behaviour to realise my aims. But there already exists a framework of value and constraint which takes into account your being an end for me, a creature to whom I could in principle be called to account.

What's interesting and puzzling here is not that *you* are lumped in with the rest of my environment, including the non-human world, but that *I* am separated from

the rest of the world. I'm interested in how you will act, as regards this first kind of significance, because I'm interested in how you will affect the success of my actions. That means that I'm interested ultimately in what will happen, in the future course of the actual world, for it is in the actual future that my actions will succeed or fail. By definition, my best guide to the future course of the actual consists of my best predictions, so it makes perfect sense for me to try to predict your behaviour, conditional on my choice, as best I can – for exactly the same reason that it makes perfect sense for me to try to predict the behaviour of my non-human environment, again conditional on my choice.

Of course, how well my aims will get realised is likely to depend also on what I do. (At least, I must take this to be the case at the time of acting, in order to think that these aims provide me with reasons for acting in these circumstances.) But I cannot adopt the same perspective towards *my* future behaviour as I do towards your future behaviour and the future behaviour of the various devices and so on which are relevant to my action. In order to choose the right option, I ask myself how you will behave, and how the rest of the environment will behave, for each option in turn. In other words, I make a series of conditional predictions, guided by the best predictive theories available to me.[14] For each option, A, B, C, and so on, I ask questions like this:

How will she behave, if I do A?

How will that thing behave, if I do A?

Note however, that I cannot ask:

How will *I* behave, if I do A?

At least, I cannot do so in cases where the range of behaviour of mine about which I want a prediction, is just my doing A, *or* B, *or* C, and so on. (I *can* ask: how will I behave if I force myself to endure John's company instead of going alone? But I

[14] Here and throughout this chapter I ignore the pragmatics of decision making, which have to do with deliberative costs. I am interested in an idealised picture of deliberation, which abstracts from these issues, because I am interested in the fundamental principles according to which we should discriminate between the possible behaviour of other persons which is relevant to deliberation, and that which is not. I assume that issues about deliberative costs and so on should modify some underlying

cannot ask: will I force myself to endure his company, or will I go alone, if I force myself to endure his company?)

There seems to be a kind of linguistic absurdity in formulating conditional predictions about that range of my future behaviour about which I'm deliberating, even though these are just the things which I do best to use in taking account of the rest of the world – of you, and everything else which will affect the success of my actions – in deciding how to act. That sets each of us apart from the rest of the world in the deliberative stance we have to our own action. This is surely not just a linguistic fact, however, but at least a fact about the most pervasive features of our conceptions of action and of deliberation. There is a kind of conceptual impossibility in treating that range of my future behaviour about which I'm deliberating in just the same way that I treat the behaviour of everything else, insofar as it touches on the success of my actions. All actions are either successful or not *actually*. This may make it appropriate to use our best guide to the future course of the actual – that is, our best predictive theories – in taking account of the influence of the environment, human and non-human, on the future success of whatever we choose to do. But this cannot be our stance towards our own future behaviour, insofar as we are deliberating about it.

This fact about our conceptions of deliberation and action may or may not reflect the metaphysical truth about action, whatever that is. But it is worth noting that the primary rationale for adopting a different stance towards our own action is not incompatible with determinism in any direct or obvious way. It's not that we treat our own behaviour differently from the behaviour of others, when we deliberate, *because* we think that our behaviour is not determined, whereas theirs is (or something like that). Instead it seems to be a presupposition of practical reasoning that we adopt a different, not purely-observational, stance towards our behaviour when we deliberate.

One could imagine someone whose thoughts about his future behaviour were only ever contemplations about how that part of the natural world which consisted of

---

picture, which gives us an idea of how we should decide what to do in cases where further reasoning

his body and mind would behave in the future. Such a person would have a purely observational stance towards his own behaviour, to which prediction would be appropriate. But such a person would not be a deliberator, nor what we think of as an agent. Practical reasoning presupposes a different stance towards our own future behaviour. It could turn out that this presupposition is inconsistent with some metaphysical truth – but it is not immediately clear that it is, and if it should turn out to be, then not just punishment and blame, and the other reactive attitudes, but any practical reasoning, would be shown to be irrational.[15]

My interest in the effects of your future behaviour on the success of my actions, then, puts you in the same category as the rest of my environment, bar me. But that's a result of special reasons (having to do with our conceptions of action and deliberation) removing me, or more precisely that future behaviour of mine about which I'm deliberating, from what is otherwise a natural residual category. It's not that some principle of cynicism puts you in the same category as things. Being interested in the success of my future actions is a species of interest in the future course of the actual world, and our best guide to this, by definition, consists of our best predictive theories. It's just that I can't take the same stance towards my own future behaviour, and at the same time deliberate about it.

The hard-nosed view is built to respond to the significance of others which has to do with an actor's vulnerability to them. For, I've argued, the best way to manage vulnerability is to get hold of the best conditional predictions about the behaviour of those aspects of one's environment which will affect the success of one's actions. And that is just what the hard-nosed view recommends, for that portion of one's environment which consists of other persons.

---

has no cost.

[15] Strawson famously argued that we could not give up our reactive attitudes. See P. F. Strawson, "Freedom and Resentment", in his *Freedom and Resentment and other essays* (London: Methuen and Co., 1974).

The future behaviour of the agent whose actions are being deliberated about features in a different way in deliberation than does the future behaviour of her environment, including other persons. We usually think of agents as individual persons. But could the agent be something other than an individual person? Suppose *we* were to be the agent whose actions are being deliberated about. Then, perhaps, in deciding how to play my part in the action, *your* future behaviour would be, like mine usually is for me, not a suitable subject of prediction. If so, other persons would, in some cases at least, have a significance for deliberators in addition to that which I have been discussing under the heading of vulnerability. They would be significant also as potential participants in joint action.

There is an uncontroversial sense in which we can do things together. We can walk together, or paint the wall together, and so on. These are things which individuals can do by themselves or together. Some things, on the other hand, persons can do only together – these are *essentially joint acts*, such as having a debate. There is no difficulty in talking of us as doing these things. But can we, together, be a single agent? Or do we simply do these things together as a collection of separate agents?

One may deny the possibility of joint agency on the following simple grounds. To be an agent with respect to some behaviour, I must control it, or, at least, typically control the behaviour of the object in question. You satisfy this condition with respect to your own behaviour, since you are in control of it, or, at least, typically in control of your body. But you do not satisfy the condition with respect to my behaviour. We may do things together, in the sense that we co-operate, but we do not act as a single agent, since neither you nor I is in control of the other's behaviour. You simply rely on my controlling my behaviour in ways you find advantageous, and I do the same with respect to you.

I think this argument is inconclusive. It's not clear that my actions cannot be in the control of someone else, in the appropriate way. People who do things together

a lot become very sensitive to cues from each other. Ballroom dancers, say, may react very quickly indeed to each other's movements. The same may be said of musicians who play together, and of a wide variety of other cases. These are cases in which it makes perfect sense to say that each individual is controlling the other's behaviour, in a way which is not much less immediate, if at all, than the clumsy kind of control I often have over my own behaviour.

Perhaps this is a legitimate use of 'control', but yet not the right sense. That may be so, but in the absence of a positive account of the kind of control which I have over my own behaviour, which explains why I am an agent with respect to it, and why one ballroom dancer cannot be an agent with respect to her partner's behaviour, we should not rule-out the possibility of true joint agency. Nor should we restrict consideration of that possibility to those cases, like that of the ballroom dancers, where the causal relations between one person's decision to act, and another's behaviour, are tuned to an unusually high degree. Without a positive account of 'control', we should consider it an open question how highly tuned relations between persons have to be, for them to be capable of true joint agency.

These speculations are meant only to prepare the way for the idea of joint agency, so that we may take it seriously. I do not pretend to have explained how several persons could be a single agent; I hope only to have rebutted a simple argument for thinking that they could not, and thereby to have made the idea that they could, less open to immediate rejection. To that end, it is worth pointing out also that we sometimes attribute responsibility to persons for their part in joint actions, not just for the things they do individually. Consider the case of a getaway driver for a bank robbery. Suppose that the robbers agree amongst themselves not to shoot anyone during the robbery, but that, in the event, someone is shot. I think we are likely to attribute *some* responsibility for this to the getaway driver, who was waiting outside all the while. If so, and if we are right to do so, that can't be because of anything he did by himself. He was, after all, in agreement with the others that the guns would not be used. But he did something *with them*, which had as one of its consequences the

shooting: they robbed the bank, and his behaviour was a part of that joint action. (It would not have been the action that it was, if he had not done what he did, even though none of his individual actions had the shooting as their consequence.)

In cases like this, we are used to thinking of persons as acting together, in a single unit of agency. Doing so does not preclude thinking of the persons who make up the joint agent as, at the same time, individual agents. Two or more people can do things together and by themselves at the same time. (When the others were in the bank, the getaway driver adjusted the vehicle's mirrors. Hence we might say that while he was *taking part in the robbery*, he was also *adjusting the mirrors*.) What we need to see now is whether the possibility of joint action helps us understand how other persons can have significance for deliberators, significance of a kind which illuminates the pure worries about acquiescence.

If we take the possibility of joint action seriously, what are its implications for deliberation? We need to distinguish between the significance we attach to other persons in virtue of our vulnerability to them, and the significance we attach to them in virtue of the possibility of joint action, by us together. As I've just suggested, others are likely to remain individual agents even whilst they join with us as a single unit of agency. Hence, one's vulnerability to them is still salient. But insofar as we focus only on their significance as potential participants in joint action, it seems to me that we should take the following view. Some possible piece of behaviour of theirs, like some possible piece of behaviour of my own, should not be made irrelevant to the evaluation of the expected effectiveness of *my* options, just because of their resolve to behave in a different way.[16]

[16] Someone's resolve to do something is often a good basis on which to predict his behaviour, but it is not the only ground. It is of particular importance for the present argument, however, because it seems to be the incompatibility between *resolve* and deliberation which explains why I cannot take a predictive stance towards my future behaviour, and still be deliberating about it. If the basis of my prediction is, say, that I have a broken leg and so cannot go skating, and not that I am resolved not to go skating, it does not seem incoherent so much as misplaced, to deliberate about whether to go skating. It's just not something I could do (not an option). The incoherence arises when I 'could' do X (in some sense which is difficult to explain – see below, Chapter Seven), but resolve not to. Similarly, a prediction that you will not go dancing which is based on the fact that you have a broken leg, *is* sufficient to rule out the possibility of your, and so our, going dancing. It is predictions based on resolve, or intention, which seem to be of interest to issues of acquiescence.

Recall that I cannot treat the prediction that I will not do X at t as sufficient to rule-out the possibility of my doing X at t, and still be deliberating about whether or not to do X at t. Similarly, insofar as I regard you as a potential participant with me in a certain joint action Q, of which your doing P at t is an essential part, I should not treat the prediction that you will not do P at t, as sufficient to rule-out the possibility that you will do P at t, if I am deliberating about whether to do my part in Q. No resolve on the part of a constituent of the agent in question – in this case, you or I – is sufficient to make possible behaviour on the part of that constituent irrelevant.

It is important not to interpret this as the claim that what should be happening here is that we all deliberate together. What's at issue is still my deliberation about *my* actions, not my deliberation about *our* actions. It's just that what we can do together may bear on my deliberation about my actions. I might be interested in the larger actions of which mine, potentially, forms a part – if so, I am interested in what we could do for the indirect reason that I am interested in the character of the various things I could do. I have not, in that case, ceased to deliberate about my own actions, in favour of deliberating with you about what we could do (or of lobbying for us to do that). This is another point at which we must be careful not to assimilate these issues about acquiescence to issues of collective consequentialism.

To see how reasoning of this form might work, and assess its credibility, consider the following case. A deliberator must choose between two options, the outcomes of which depend on the behaviour of another person (the numbers represent shared evaluations[17] of each eventual outcome):

[17] The evaluations are to be understood as reason-based, rather than as disposition-based. Hence it is quite possible for the other person to choose a sub-optimal option, as is supposed in the discussion.

		The other person	
		a	b
Deliberator	A	11	10
	B	10	100

Table 2: The capricious other person

The deliberator believes that the other is resolved to choose (a), but that he will do so irrationally, out of mere caprice.[18] In this situation, the deliberator acting individually can secure an outcome of 11 by choosing (A), or an outcome of 10 by choosing (B). If she treats the other person only as someone to whom she is vulnerable, she should choose (A). I claim that she may also treat the other person as a potential participant in joint action, however. If she does this, the fact that he has resolved to choose (a) is weightless. It remains true that the outcome of 100 is available to *them*, even if she cannot secure it individually. Given the availability to them of this much better outcome, and the fact that he is resolved to choose (a) only out of caprice, their bringing about the outcome of 11 is an irrational joint act. My claim is that this may give the deliberator some reason not to choose an action which is a necessary component of that irrational joint act, which is to say that it may give her *some* reason not to choose (A).[19]

[18] In discussions of decision problems such as this, the assumption is usually made that the other person is rational. That means that common knowledge of the payoff structure brings with it predictive value: we can expect the other person to do what he has reason to do. The case I am discussing is different, however, since we are supposing that we know, on some independent grounds, that the other person will not do what he has most reason to do. He is irrational, or capricious, but his behaviour is still predictable. Such cases are very common. For discussion of the implications of the assumptions of common rationality and common knowledge, see M. Hollis and R. Sugden, "Rationality in Action", *Mind* 102:405 (1993), pp. 1-35.

[19] There is an important possible objection here. If (A/a), which yields 11, is an irrational joint act because of the possibility of their getting 100, then surely (B/a), which yields the even worse outcome 10, is an irrational joint act too. If so, there is a reason from consideration of the possibility of joint action for the deliberator both to do (B), and so avoid the irrational joint act (A/a), and not to do (B), and so avoid the irrational joint act (B/a). That looks incoherent.

In what sense is the possibility of the other person's choosing (b) relevant to the deliberator's evaluation of expected effectiveness? She needn't believe that he may change his mind, if she chooses (B). She need not suffer from any mistaken views about causation, or about his co-operativeness. It's not that she's under the illusion that she can, somehow, individually secure the payoff of 100; instead, she resists an otherwise favoured option because of its contribution to an irrational joint act. (Thinking of her reason as a reason *against* choosing (A), rather than a reason *for* choosing (B), helps to avoid the mistake of thinking that this reason must be founded on the illusion that 100 is somehow individually securable by her. It does not depend on that illusion.[20]) The possibility of his choosing (b) is relevant because it is a necessary condition of their receiving 100 – which, since it is a possibility open to them, makes her choosing (A) and his choosing (a), which leads to their receiving only 11, an irrational joint act.

Insofar as we think of ourselves as individual agents, we should think of others in our environment as persons to whom we are vulnerable in our courses of action. I do something, and you affect the success of the course of action I have chosen. But insofar as we are willing to think of ourselves as capable of sharing joint agency with others in our environment, we should think of them as potential participants in joint action. That changes, or should change, our attitude to the things we each can do individually. Individual actions have *consequences*, but they also may form *parts* of larger actions. Hence, we may think of them not just in terms of their causal relations with other acts – if I do such and such, this may cause you to co-operate – but also in terms of their constitutive relations with other acts.[21] If we all vote and the government changes, that's something we've done together. I could not

---

This appearance is deceptive. The deliberator's options (A) and (B) do not stand on an equal footing as far as more-inclusive reasoning is concerned – and not just because of the slight difference in the expected payoffs they yield if the other person chooses (a).

The deliberator's choosing (A) is a sufficient condition of their not receiving 100. Hence, it is necessarily true that, if she chooses (A), they do not get 100. In contrast, her choosing (B) is sufficient for them not to receive 100, only if he chooses (a). So it is not necessarily true that, if she chooses B, they do not get 100.

[20] Compare S. L. Hurley, *Natural Reasons*, p. 148, on the temptation to think that claims that collectively available outcomes may provide individuals with reasons for action, must depend on "magical or superstitious thinking of some kind."

have done it by myself, and my voting looks irrational if we focus only on the expected consequences of my act.[22] But if we focus on its constitutive relations with other acts – the fact that it forms part, with them, of electing a new government – it may well make sense.

In the case of the capricious person, the deliberator has a reason not to choose the course (A) which would secure the best payoff available to her individually, because doing so would involve her choosing a course of action which is part of an irrational joint act. Their choosing (A/a), considered as a joint act, is irrational, because it yields only 11, and the much better outcome of 100 is available to them. The payoff figures chosen are not particularly important to this case: what is important is that an outcome is available to *them* which is indisputably better, from a morally relevant point of view, than any outcome which is available if the deliberator chooses the course of action which secures the best outcome individually available to *her*. This structure is in fact quite common in cases where someone else is obstructive. It is present in the case of the kidnapper, and Cohen's analogy to that case supposes that it is shared too by the case of incentive-seeking by the talented.

Cases with this feature are cases in which it may be rational to acquiesce, but cases also in which there may be reasons against doing so, from what I've been calling pure worries about acquiescence. Such reasons are *pro tanto* reasons: genuine reasons for action, but capable of being defeated by opposing reasons. In fact it is likely, in most cases, that such reasons are heavily outweighed by opposing reasons having to do with an actor's vulnerability to others, which may be why they are not often noticed. As I shall explain in Chapter Six, it is rare that we are right to act according to pure worries about acquiescence, since in most cases it would be reckless to ignore our best predictions about what the other persons will do. But the existence of *pro tanto* reasons for action based on pure worries about acquiescence is

---

[21] See S. L. Hurley, *Natural Reasons*, p. 148-151, on 'constitutive consequences of acts'.
[22] An introductory discussion of Downs's puzzle about the rationality of voting, is I. McLean, *Public Choice. An Introduction* (Oxford: Basil Blackwell, 1989), Chapter Three.

sufficient to show that Cohen's criticism of the argument from incentives may have force against a deliberative form of that argument, in principle.

Finally in this section, let me illustrate the claims I've made about the significance of others as potential participants in joint action by reference to the familiar moral scenario of Jim and the Indians. Jim is a botanist who stumbles across a captain, Pedro, who is about to shoot twenty Indians, chosen at random from the local population, in order to quell protest. Since Jim is an honoured visitor, however, Pedro offers him the choice of shooting one of the Indians himself, in which case the other nineteen will be set free. Jim believes that he would be unable to overpower Pedro, and that if he does not shoot one Indian himself, all twenty will be shot.[23]

Jim is faced with a stark choice. He believes that Pedro really is resolved to shoot twenty Indians if he, Jim, does not shoot one. Also, he would not be able to overpower Pedro, and so make Pedro's resolve causally ineffective. Given these facts, only two outcomes seem to be within Jim's individual causal power to secure: the outcome in which one Indian gets shot, and the outcome in which twenty Indians get shot.

Suppose that there is *some* reason for Jim not to shoot any Indians. (It may nevertheless be right, all things considered, for him to take up Pedro's offer.) There are two common explanations for this. One of these involves Nozick's idea of a side constraint, or something like it.[24] One may not violate a side constraint even to prevent more violations of the same side constraint. Perhaps Jim operates with a side constraint against taking human life, or those of us who think that there is some reason not to shoot any Indians, think he should operate with such a side constraint. This would explain why there is something to be said in favour of his not shooting any Indians, certainly. (On the other hand, this explanation is in danger of crowding out any possible reasons *in favour* of his shooting one Indian.)

---

[23] See B. Williams, "A critique of utilitarianism", in J. J. C. Smart and B. Williams, *Utilitarianism for and against* (Cambridge: Cambridge University Press, 1973), pp. 98-100.
[24] See R. Nozick, *Anarchy, State, and Utopia* (Oxford: Blackwell, 1974), pp. 28-30.

The second common response to the demand for explanation appeals to the idea of personal integrity. Jim has some reason not to shoot any Indians because to do so would violate some of his fundamental beliefs and values, and so would undermine his personal integrity. In the form that Williams offers this explanation, it is a mistake to think that integrity is another value, whose level of promotion can be traded off by the agent against the various other values which are relevant. Instead the appeal to integrity is supposed to be about the structural relationship between, as Raz puts it, "one's own projects and the moral requirements which arise independently of them."[25] Moral requirements on agents are not wholly independent of those agents' most fundamental projects.

Either of these two explanations may be correct. However, I think there is a third explanation, which comes from considerations of the rationality of joint action, and so illustrates how such considerations may work. I noted earlier that Jim is faced with a stark choice, since he is individually capable of securing only two outcomes, both of which are repellent. That is true, but it is also true that Jim and Pedro *together* are capable of bringing it about that no Indians are shot. It is possible for both Jim and Pedro not to shoot any Indians. Of course, Jim believes that Pedro really is resolved to shoot twenty Indians if Jim shoots none, and it is this belief which supports his predictions about how Pedro would behave if Jim chose each of his alternatives. Insofar as he is concerned with his vulnerability to Pedro, Jim should use these predictions to evaluate the expected effectiveness of his options, and so make his choice.

If Jim is concerned with the rationality of his and Pedro's *joint* action, however, then the predictions based on Pedro's resolve should not play the same role in his reasoning. In particular, they should not make irrelevant the possibility that, together, they bring it about that no Indians are shot. One of the things they can do together is to shoot no Indians; a necessary part of this joint action is Jim's not

---

[25] J. Raz, *The Morality of Freedom*, p. 286. Here Raz is offering an interpretation, with which I agree, of Williams's view. The relevant passage in Williams is "A Critique of Utilitarianism", in J. J. C.

shooting any Indians. So if he were to shoot one Indian, Jim would be taking part in a joint action which is not the optimal joint action – that is, he would be taking part in the joint action in which one Indian gets shot and the others go free. Concern with his part in this joint action may give Jim some reason not to shoot any Indians himself (though that reason may well be outweighed by his hard-nosed reasons, which reflect his vulnerability to Pedro).

The idea of the rationality of individual action places a constraint on the class of possible outcomes to which reasons for action can attach. According to this constraint, a possible outcome can give an individual a reason to act only if it is reasonable for the individual to believe that it is possible for his action to secure the outcome (in some degree). The possibility of world peace may give me some reason to act in pursuit of it, but only insofar as it is reasonable for me to believe that my action can secure it to some degree. If it is unreasonable for me to believe that my action can secure it in any degree, then its possibility gives *me* no reason to act at all. In contrast, the idea of the rationality of joint action places a different constraint on the class of possible outcomes to which reasons for action can attach. If I am a member of some group which has the causal power to bring about an outcome, that is sufficient for me to have a reason for action attaching to that possibility.

This idea correlates with the more-inclusive view's central proposition: that the expected contribution of an action to promotion of the good, should not always be evaluated on the assumption that third parties' behaviour is fixed. Sometimes it should be evaluated on the assumption that others would behave in a way which, we predict, they will not. In fact, there is a close connection between the rationality of joint action and the more-inclusive view. The more-inclusive view is built to respond to the significance of others as potential participants in joint action, just as the hard-nosed view was built to respond to our vulnerability to others. For the more-inclusive view enables us to explain how, all things considered, it can be right for us to perform a certain action P rather than another Q, even though P would have better

---

Smart and B. Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press,

consequences than Q only if others do not behave as we predict. The more-inclusive view licences disregard of another person's will in cases of potential acquiescence, and this may be explained in terms of the rationality of joint action. Even though I know you will not co-operate, it may be rational for me to perform the action which makes sense only on the assumption that we act together.

There are pertinent comparisons to be made between this view and other general moral views, such as Kantianism and rule-consequentialism. I shall try to address these issues in Chapter Six. For the moment, I want briefly to explain why I have chosen to discuss 'joint action' rather than 'co-operation', and to head-off a possible objection to my argument.

5. An objection answered

The idea that others may be significant as potential participants in joint action is importantly different from an apparently similar idea proposed by Donald Regan and Susan Hurley. They have claimed that someone may have a reason to co-operate with others who are willing to do so in bringing about an outcome which he cannot secure individually, even if this involves his doing something which is irrational as an individual act.[26] Their idea contrasts *co-operative action* with individual action, whereas mine contrasts *joint action* with individual action. The difference is that further conditions must apply for a situation of possible joint action to be one of possible co-operative action. The possibility of joint action requires only a joint causal power to bring about outcomes which are not individually securable; the possibility of co-operative action requires also that some others are willing to co-operate.

1973), pp. 116-117. Raz goes on to reject Williams's view.

[26] See S. L. Hurley, *Natural Reasons*, Chapter 8; S. L. Hurley, "Newcomb's Problem, Prisoners' Dilemma, and Collective Action", *Synthese* 86 (1991), pp. 173-196; S.L. Hurley, "A new take from Nozick on Newcomb's Problem and Prisoners' Dilemma", *Analysis* 54: 2 (1994), pp. 65-72; D. Regan, *Utilitarianism and Co-operation* (Oxford: Clarendon Press, 1980), Chapter 8 especially.

The case of the capricious other person, for example, does not satisfy this latter condition, and so on Hurley's and Regan's view there is no reason from the possibility of co-operation for the deliberator to choose (B). In contrast I think that the existence of a joint causal power to bring about the outcome of 100 is sufficient to give the deliberator a reason to choose (B). It is *this* fact which makes the other person's will like the deliberator's own will in one respect, in not being a constraint on her deliberation.

Further conditions having to do with the possibility of co-operative action serve only to assimilate the rationality of joint acts to the rationality of individual acts, by picking out cases where the environment happens to be favourable, so that there is some chance of the individual actor's actions helping to cause the optimal outcome. But this is to narrow the scope of the rationality of joint action unduly. The possibility of joint action does not lose its salience in all cases where there is no prospect of co-operation. Even if I'm certain that Smith will obstruct me, I may choose a course of action with one eye on a scenario in which she is not obstructive. In doing so, I may be quite rational, since I am concerned to avoid taking part in a certain irrational joint act. I can often avoid *that* no matter how unco-operative the other person is. So considerations of this sort needn't fade away when co-operation isn't forthcoming.

In fact, the argument for the more-inclusive view from the significance of others as potential participants in joint action, is quite different in character to arguments which have been made about the inadequacy of certain forms of consequentialism. These other arguments focus on problems which arise if we insist that the appropriate unit of agency, when we're evaluating consequences, is always an individual person. As I suggested earlier, that issue is quite independent of the issue I've been discussing, though it can seem as if they are connected. The reason that I say we should evaluate the consequences of an individual's actions on the assumption that others will not behave as we predict, is not because there is a real possibility that they *will* do otherwise than we predict. Arguments about the unit of agency, of the

usual kind, make sense only on that assumption. I think we should evaluate consequences in the way I've described because others *could* do otherwise than we predict, not because they will.[27] Moreover, for *any* unit of agency, the same issues of acquiescence could arise with respect to other agents.

Nevertheless, those who favour thinking in terms of co-operative action may make an important objection here. They may say that the idea of possible joint action, at least as I've discussed it, is not specified closely enough, so that the number of possible joint actions proliferates hopelessly.[28] For my account of joint action claims that the existence of a joint causal power to secure some outcome is a sufficient condition for the possibility of joint action in pursuit of that outcome. Without specification of some extra conditions, necessary for the possibility of joint action, it seems that each of us is a potential participant in an unmanageably large number of possible joint actions, some of which will be very odd indeed. For each of us interacts, and could interact, causally with very many other persons, in subtle and often irrelevant ways.

One could make this point without putting it in terms of co-operation theory. But the point is sharpened when presented by those who advocate analysis of problems of acquiescence, for example, in terms of co-operative action, for the following reason. As I've noted, the idea of co-operative action typically includes extra conditions in addition to the existence of a joint causal power, and so looks well-suited to avoiding the problems of proliferation from which the idea of joint action suffers. It usually includes, for example, the idea that potential co-operators must be capable of recognising each other, must be willing to co-operate, and therefore must share, in part, evaluations of the possible outcomes of their actions. So, the co-operative theorist might claim, we would be better off trying to explain pure worries about acquiescence in terms of the rationality of co-operative action, and need not be concerned with the more-inclusive view.

---

[27] See below, Chapter Six, section 1, and Chapter Seven, where I discuss the problematic sense of 'could' which this claims relies upon.

[28] Susan Hurley made this objection to me.

Let me respond to this objection. The charge is that my account of joint action leaves potential joint action unconstrained in two ways: (a) with respect to a fixed group of potential actors, the range of possible joint acts is left unconstrained; and (b) the group of potential actors is left unconstrained. Problem (a) is no more of a problem for the idea of joint action than it is for our understanding of deliberation generally. Accounts of co-operative action, like accounts of individual action, do not tell us how to determine the range of things an actor could do which is practically relevant. In my terminology, they do not tell us how to determine the actor's range of options. We probably can't say much in response to this problem other than that knowledge of established patterns of action, together with some imagination or innovation, is what we use to formulate an actor's options. I do not think that this problem is any more severe in cases of joint action, than in cases of individual or co-operative action. The range of possible actions which a fixed group of potentially joint actors could perform, is narrowed in the same way, in principle, as is the range of possible actions which a single actor could perform.

Problem (b) seems more difficult to deal with. I have joint causal powers with a very large number of sets of other actors, and most of these sets are gerrymandered in odd ways. Which of them should count as genuine sets of potential joint actors, and why? Again the problem is more tractable if we know something about the deliberator's practical task and aims. This information allows us to use knowledge of established patterns of behaviour, allowing as usual for innovation, to pick out certain sets of joint actors as real candidates – in a way which is similar to the method which everyone must use, since there are no others, in dealing with problem (a).

The deliberator can narrow both the range of potential joint actors and the range of potential joint actions, then, by casting around for established patterns of action which are relevant to her task. Moreover, I suggest that a similar process of narrowing the range of possible joint actors and actions must be present in co-operative action. Recognition of shared evaluation of outcomes, for example, can take place only after some initial sifting of other agents into those who are, and those who

are not, potential co-operators. Suppose there is a fire, and the deliberator wants to know how best to co-operate with whoever else is willing to help. She does not consider whose evaluations of outcomes overlap with hers in the whole world of actors, but rather casts around to see who is a feasible co-operator, using as her guide knowledge of established ways of dealing with fires. Only once this initial sifting has taken place does it make sense to look for shared evaluation of outcomes, mutual recognition, willingness to co-operate, and so on.

So, the range of potential joint actors and actions is constrained by conditions other than those of willingness to help, shared evaluation, and so on. Moreover, the factors which must be appealed to must be appealed to also in the analysis of co-operative action, and, with respect to the range of possible actions, in the analysis of individual action. The extra conditions which are built-in to the idea of co-operative action are unnecessary to avoid unacceptable proliferation.

6. Conclusion

I have done my best to defend the idea of a distinct kind of significance of other persons for us as deliberators, one which is in addition to that which they have in virtue of our vulnerability to them, and one which makes sense of pure worries about acquiescence. This is a very difficult area, however, in which intuitions are shaky and may easily differ, so I do not expect to have convinced everyone. But at least I hope to have brought-out some logical relations between our ideas about acquiescence and our ideas about deliberation and rational action. Those who do not accept the conclusion may know better which premise to reject.

If we accept the significance of others as potential participants in joint action, we should reject the hard-nosed view, understood as the view that behaviour of third parties which, we predict, will not happen, cannot be relevant to the evaluation of the expected effectiveness of an actor's options. If so, then it may be right, in some circumstances, to choose an option which would be more effective than some other

option only if others behave in a way which, we predict, they will not.[29] Moreover, that implication does not depend on any uncertainty in our predictions. Even if we are certain about what others will do, it may be right in some circumstances to act in a way which would be more effective only if others were to do what we are sure they will not. In Chapter Six, I shall explain in more detail how reasoning of this sort could be applied and the circumstances in which it may be appropriate, and I shall compare it with some other similar moral views.

[29] If pure worries are rational, they provide *pro tanto* reasons for action, so they may not be overriding in any particular case. Indeed, I have suggested that they will often be opposed by strong hard-nosed reasons having to do with our vulnerability to others. It does not follow from the existence of *pro tanto* reasons that it is ever right, all things considered, to act upon them.

Chapter Six

More-Inclusive Reasoning

The aim of this chapter is to develop further my account of more-inclusive reasoning. I shall explain further what it might involve, and when it would be appropriate. My method will be to draw pertinent comparisons with other, apparently similar doctrines. Amongst themselves, these doctrines are surprisingly diverse. More-inclusive reasoning has some affinity to Kantian ethics – but also to rule- and other variant forms of consequentialism, and to Liam Murphy's views about the demandingness of morality.

It is interesting to speculate about what these various views have in common, which explains why they each have affinities to more-inclusive reasoning. The most plausible answer, on the face of it, is that they are all to be contrasted with act-consequentialism. If so, exploration of the rationale and character of more-inclusive reasoning should be expected to be, at the same time, exploration of possible defects of act-consequentialism.

I think that this supposition is correct. But this chapter does not offer any new arguments against act-consequentialism. Such arguments as I have, were presented in Chapter Five in support of the more-inclusive view as a view about the significance of

others for us as deliberators. In this chapter I shall be more concerned to elaborate the more-inclusive view. The discussion is mostly typological and explanatory, rather than argumentative. If there is an over-arching argument, it is that act-consequentialism is a complex doctrine, and hence that it may be rejected or modified in a number of different ways. For the moment, the possibility of progress in this area seems to lie in striving for a more accurate description of the territory, rather than in trying to resolve the various issues, or to reduce them by showing some of them to be identical to each other.

1. More-inclusive reasoning and Kantian ethics

The more-inclusive view says that:

(MIV) behaviour which will not occur, according to our best predictions, may nevertheless be relevant to assessment of expected effectiveness.

In Chapter Five I tried to explain the sense of 'relevant' which is at issue here. It has to do with the behavioural assumptions it is appropriate to make about other persons when evaluating the expected effectiveness of an option. In this section I shall explain how such evaluation could take place, in part by drawing comparisons with some of Kant's views.

The more-inclusive view, understood in the way I've suggested so far, bears some resemblance to one of Kant's ethical doctrines – namely, his view that the rightness of certain actions is independent of our expectations of the behaviour of

others. This doctrine gains its most notorious expression in Kant's insistence that it is wrong to lie, even to a murderer enquiring at the door about his next victim's whereabouts. Kant thought that his ethical views implied that lying, even in such unusual circumstances, was wrong; and he sought to explain this in terms of personal responsibility:

> After you have honestly answered the murderer's question as to whether his intended victim is at home, it may be that he has slipped out so that he does not come in the way of the murderer, and thus that the murder may not be committed. But if you had lied and said he was not at home when he had really gone out without your knowing it, and if the murderer had then met him as he went away and murdered him, you might justly be accused as the cause of his death.[1]

That is, by lying to the murderer, the agent takes on responsibility for the murderer's actions, since she has not treated him as a rational creature. Hence the lie carries with it a risk, no matter that it is small, of causing the victim's death. In contrast, in telling the truth, she treats the murderer as a rational creature, preserving his responsibility for the use he makes of her information. There is no risk of *her* causing the victim's death, in telling the truth; since if the murderer uses the information to kill, *his* will is the first cause of that action.[2]

---

[1] I. Kant, "On a Supposed Right to Lie From Altruistic Motives", quoted at p. 326 in C. M. Korsgaard, "The Right to Lie: Kant on Dealing with Evil", *Philosophy and Public Affairs* 15:4 (1986), pp. 325-349. Korsgaard's view is that the lie in question is permitted by the universalisability test, but not by the injunction to treat persons as ends in themselves.

[2] Korsgaard, "The Right to Lie", pp. 334-337.

Kant's doctrine is like the more-inclusive view, in its removal from deliberation of consideration of what other persons will do. Insofar as we are interested in the significance of others as potential participants in joint action, we ignore some information about what they will do. Similarly, what is right, according to Kant, does not depend in any direct way on what we expect others to do in response to our actions. In this respect, Kant's view and the more-inclusive view are alike.

But the more-inclusive view is subject to two kinds of restriction not shared by Kant's doctrine. First, there is an *internal restriction*, on the unpredicted possible behaviour which is allowed to count as relevant. Kant's view is that what is right is wholly independent of what we expect others to do: it is a matter of the universalisability of the maxim which we will. Thus lying is wrong, since no maxim permitting it could be universalised, regardless of what we expect others to do – even if we expect them to commit murder using the information we give them. Such radical independence (of rightness of action from the expected behaviour of other persons) amounts to what we might call a *very*-inclusive view: in effect, we are to operate on the assumption that others will do as they ought. Since, according to the Kantian view, what others ought to do does not depend on any facts about their psychology or disposition, we may disregard these contingent features of their circumstances, and conceive of them simply as rational creatures. The expected effectiveness in promoting justice of telling the truth is then always greater than the

expected effectiveness of lying, since other persons' wickedness does not enter into the calculation of expected effectiveness.[3]

In contrast, I think that the more-inclusive view should incorporate some restriction on the range of possible behaviour which is counted as relevant. It is very difficult to say exactly what the restriction should be, but we may locate it between two ends of a spectrum. At one end of the spectrum is Kant's view, according to which facts about a third party's circumstances and psychological condition should have no bearing on the range of relevant possibilities: whatever these facts are, we should deliberate on the assumption that the person in question will do as he ought. Hence we may disregard the contingencies of his situation entirely. At the other end of the spectrum is the hard-nosed view, according to which the relevant possibilities are very strongly constrained by the person's circumstances and motivations. Indeed, if we suppose that determinism is true (though the hard-nosed view does not itself presuppose this), and that our best predictions could foresee the future perfectly, then the hard-nosed view implies that the range of relevant possible behaviour is uniquely selected. What's relevant is what persons will actually do, and nothing more.

The restriction we need would characterise some middle ground between these extremes. It would specify a sense of 'could', according to which the range of things a person could do (a) extends further than the range of things they will actually do, and (b) is nevertheless determined in some way by facts about their circumstances. I do

---

[3] Kant's official doctrine is that the rightness of an action does not depend on its expected consequences at all – not even on its expected consequences calculated on very inclusive assumptions about others' possible behaviour. It depends instead on the quality of will involved. So perhaps it is better to call the view I have described in the text, quasi-Kantian. But note that the quasi-Kantian view simulates the official Kantian doctrine, in practice. If one thinks that what someone ought to do is independent of their circumstances, then deliberating on the assumption that others will do as they ought amounts to deliberating without regard for certain features of one's circumstances – namely, the dispositions and intentions of other persons.

not know how to specify this sense of 'could', and so I do not know how to specify the associated restriction on the range of relevant possibilities (though I shall make some speculations on this issue in Chapter Seven). It may turn out that no such sense of 'could' is consistent with metaphysical or physical truth, though I doubt it. But in any case, such a sense of 'could' is, I think, needed to make more-inclusive reasoning plausible. It would provide an internal restriction on more-inclusive reasoning, differentiating it from very- or ultra-inclusive views, such as the quasi-Kantian doctrine I've been discussing. (And, if I'm right, it is needed to make sense also of our views about acquiescence.)

Secondly, there is an additional restriction, not shared by Kant's view, on the applicability of more-inclusive assumptions to different deliberative problems. Kant does not think that we should begin to take an interest in what others are expected to do in cases where the expected consequences of doing otherwise are very grave. But the idea of more-inclusive reasoning naturally encourages such a restriction on scope, which I shall call an *external restriction.*[4] Often we should not employ more-inclusive assumptions about the behaviour of others, since the significance of others which is captured by the idea of our vulnerability to them outweighs their significance as potential participants in joint action. When faced with a murderer at the door, I should be more concerned with my (and his potential victim's) vulnerability to him, than I am with the potential for participating with him in the irrational joint act of my lying to him while he goes peacefully on his way.

More-inclusive reasoning is subject to an external restriction, then, since it is not always, or even often, appropriate to pay more attention to the significance of

others as potential participants in joint action, than to their significance as persons to whom we are vulnerable. One important variable in determining the suitability of more-inclusive reasoning, is the difference between the expected benefit calculated on hard-nosed assumptions of (a) the option recommended by hard-nosed reasoning, and (b) the option recommended by more-inclusive reasoning. Where this is very small, it may be appropriate to use more-inclusive reasoning, since the loss in dealing appropriately with vulnerability may be outweighed by the gain in dealing appropriately with the possibility of joint action. For example, in the case of the capricious other person discussed in Chapter Five, the agent stood to do only marginally worse by choosing the option recommended by more-inclusive reasoning, if the other person were to act as we predicted, than she would do if she were to choose the option recommended by hard-nosed reasoning. In cases like this, where there is little to choose between the options on hard-nosed grounds, more-inclusive considerations may have more prominence.

Another important factor is the relationship between the deliberator, the actor, and those who stand to do worse by more-inclusive reasoning if the other persons behave as we predict they will. In the case of Jim and the Indians, it is the Indians who stand to do worse if Jim declines to shoot and Pedro carries out his threat of shooting twenty as a result. They are third parties with respect to the actor, and also with respect to the deliberator, whether it is Jim deliberating or us doing it. This considerably weakens the force of considerations of the rationality of joint action.[5] In

---

[4] Another way of putting the same point is to say that more-inclusive reasons have *pro tanto* force (see Chapter Five, section 4).

[5] Reasons for action differ in the extent to which they are affected by a shift from first-person to third-person points of view. Your well-being provides me straightforwardly with reasons for action on your behalf; in contrast, your reason to pay someone $10 which is provided by your having promised to do so, does not translate readily into a reason for me to do so. More-inclusive reasons are like these latter

contrast, in the case of the capricious other person, it is the actor herself who stands to do worse – and then, only marginally so – if the other person in fact behaves as she predicts he will. In such cases, consideration of the significance of others as potential participants in joint action may have more weight.[6] In both of these ways, then, the idea of more-inclusive reasoning which I'm advancing incorporates restrictions which mitigate the worries about recklessness which have afflicted Kant's view.

Finally in this section, I should explain how more-inclusive reasoning might work. What we need to see is how the more-inclusive set of possible actions of the other person should be brought to bear on the assessment of expected effectiveness of the actor's options. There is a well-established method for bringing predictions to bear on the assessment of expected effectiveness, at least where meaningful probabilities can be attached to different possible outcomes.[7] This is the method of calculating the expected benefit of an option as the sum of the products of the probability and value of each possible outcome of that option. In this way a range of possible outcomes, the probability of which may depend on what other persons do, may be combined in a single assessment of expected effectiveness. The different possibilities are assessed individually for effectiveness, and then weighted according to their probability.

---

reasons. Your reasons to resist acquiescence *on grounds of pure worries* do not translate readily into reasons for me to do so on your behalf. That may be because of some conceptual link between more-inclusive reasoning and the value of autonomy. My autonomy is more resistant than my well-being to promotion by others on my behalf.

[6] In this connection it is interesting to note that at one point in his Tanner Lectures, Cohen considers the possibility of letting the worst-off decide for themselves whether the top tax rate should be kept high. Cohen does not discuss the issues I've raised here, but one possible interpretation of the rationale for letting the worst-off decide, is as follows: The worst-off are third parties with respect to the actor – the taxing agency – so the case for acting on pure worries about acquiescence is stronger if they are made the sole deliberators. See Cohen, "Incentives", pp. 373-375.

[7] See R. Nozick, *The Nature of Rationality* (Princeton: Princeton University Press, 1993), Chapter 2, for a discussion.

This method cannot be extended, except in an *ad hoc* fashion, to more-inclusive reasoning. According to more-inclusive reasoning, a possible outcome may bear on expected effectiveness even if its probability is zero.[8] But the method just outlined multiplies the value of each possibility by its probability, and so would count a possibility with probability of zero as making no contribution to expected effectiveness. One could assign nominal probabilities to such events – but this would be arbitrary, and would have the result that the probabilities would sum to more than one. So we must look for a different method of calculating expected effectiveness, to see how more-inclusive reasoning might work.

There are two ways in which this might be done. One may either ignore the predictions, here supposed to be in the form of probability information, or one may make use of them in a different way. The first course assimilates more-inclusive reasoning to decision-making under conditions of uncertainty. Suppose that the decision problem is represented by the following table:

		The other person			
		A (p=1/3)	B (p=2/3)	C (p=0)	D (p=0)
	i	5	5	4	30
The	ii	14	14	14	12
Actor	iii	16	10	17	18

Table 3: Illustration of more-inclusive reasoning

[8] The probability in question may be subjective or objective: it may, but need not, reflect the deliberator's beliefs. Even a possibility with objective probability of zero may be relevant, if the more-inclusive view is correct, since it may bear on the rationality of joint actions.

Someone who employs more-inclusive reasoning has to form a judgement about the range of possible behaviour of the other person which is relevant to the actor's decision. I have supposed that there are four relevant ways in which he could behave, labelled A, B, C, and D, with the objective probability of each shown in brackets. The agent-neutral value of each possible outcome is then shown in the main body of the table.[9]

The method which assimilates more-inclusive reasoning to decision making under conditions of uncertainty, can apply any of a range of principles which have been proposed for the latter, to select a course of action for the actor.[10] In each case, we ignore the probability information, and look only at the value of each possible outcome. Thus, one could adopt the *maximax* principle, according to which one should select the option with the highest possible value – in this case, option (i). Or, mindful of the recklessness of that principle, one could adopt the *maximin* principle, which tells us to choose the option whose worst outcome is as good as possible (favouring option [ii]), or the *minimax regret* principle, which tells us to choose the option with the minimum maximum regret. *Regret* in this context is understood as the positive difference between the outcome of an option, given the other player's move, and the best possible outcome, given the same move on the part of the other player. It is calculated by subtracting the value of each outcome from the highest value in the same column. A regret table for the decision problem we're discussing is shown below.

[9] Once again, I suppose that the other person may not choose the option which is optimal for him. (See Chapter Five, note 17.)

	A	B	C	D	maximum regret
i	11	9	13	0	13
ii	2	0	3	18	18
iii	0	4	0	12	12

Table 4: A regret table for the same decision problem

So in this case, option (iii) would be selected by the minimax regret principle.

The various rules for decision-making under conditions of uncertainty have been widely-discussed – with, I think, the general conclusion that they are each suited to different purposes and conditions. Thus, for example, Rawls famously argued for the merits of the maximin rule in circumstances where, "the person choosing has a conception of the good such that he cares very little, if anything, for what he might gain above the minimum stipend that he can, in fact, be sure of by following the maximin rule . . . [and] the rejected alternatives have outcomes that one can hardly accept. The situation involves grave risks."[11] We might think that consideration of circumstantial factors, such as these, could motivate principled choice of the appropriate decision-rule for more-inclusive reasoning in a range of different cases, following the assimilation strategy. But this strategy seems to be open to a fatal objection.

Its defect is that it wilfully ignores relevant information, in taking no notice of our predictions of the other person's behaviour. This complaint has two strands. The

---

[10] I have relied on a simple introduction to these issues: A. Colman, *Game Theory and Experimental Games. The Study of Strategic Interaction* (Oxford: Pergamon Press, 1982), pp. 22-30.

[11] J. Rawls, *A Theory of Justice*, p. 154. Rawls argues that the circumstances of parties in the original position have these features.

first point is that it just seems perverse to recommend taking a decision using less information rather than more, barring special reasons.[12] The second point is that attention to this information may undermine the rationale for choosing one rather than another of the rules which have been proposed for decision making in conditions of uncertainty. Thus, for example, Rawls's rationale for adopting maximin had to do with the relative importance, to the actor, of securing a high minimum payoff rather than attempting to secure a high maximum payoff. In our example, as we saw, the rule recommends choosing option (ii), since this has a minimum payoff of 12, which is higher than the minimum of any other option. But if we look at the information about the probabilities of each outcome, we see that this outcome has a probability of zero. It makes no sense to think of *securing* this outcome, and so the otherwise intelligible rationale for selecting maximin is undermined.

This criticism does not renege on adherence to the more-inclusive view. Outcomes which, we predict, will not occur can be relevant to the assessment of expected effectiveness. But if the question is *how* they are to be taken into account, it is useless to appeal to the prospect of *securing* one outcome rather than another, in arguing for the appropriateness of a decision-rule which has its natural home in cases where we do not have probability information. That's not a more-inclusive argument: it's an example of the magical thinking which I took pains to distinguish from more-inclusive reasoning in Chapter Five. The various rules for decision-making in conditions for uncertainty, if they are to be applied to more-inclusive reasoning, need new rationales.

---

[12] One such special reason is the connection between ignorance and impartiality which Rawls has sought to exploit. Another set of reasons has to do with the costs of decision-making – but, as I have

There may be something in that approach, but I shall not pursue it any further. Instead I want briefly to outline the second possible approach, which makes use of probability information, but in a different way than the established method for hard-nosed reasoning. Let us consider a different decision problem:

		The other person			E(V)
		A (p=1/3)	B (p=2/3)	C (p=0)	
	i	16	10	8	12
The	ii	10	10	20	10
Actor	iii	8	8	8	8
	iv	4	4	4	4

Table 5: The second strategy illustrated

The kind of method I have in mind proceeds in two stages. First it uses the information about probability in the usual way, to calculate (according to hard-nosed assumptions) the expected payoff of each option, shown in the table as E(V). Second, it uses this payoff information to establish some threshold or target for options to meet, assessed on more-inclusive grounds. There would be many ways of doing this. For example, one could focus only on unpredicted outcomes, in column C in our example, and treat the maximum value of E(V) as a simple threshold for options to meet. One then chooses any option with an unpredicted outcome greater than the

done throughout this work, I ignore these issues here. They could not be expected to justify the assimilation strategy in general as the appropriate method for calculating expected effectiveness.

maximum value of E(V). In this case, one would choose option (ii) on more-inclusive reasoning, since it is the only option with an unpredicted outcome of value greater than 12.[13]

I have not worked-out the details of a two-stage method along these lines. The significance of the proposal is that it reflects the thought that the significance of others which is a result of our vulnerability to them, does not disappear when we consider the possibility of joint action with them. More-inclusive reasoners are not reckless, since they do not ignore information about what others will do; they just set this information alongside consideration of what others *could* do, in the context of thoughts about possible joint action. They weigh the two kinds of significance against each other.

There are no doubt many possible ways of doing this, some of which are more intuitionistic than others. The method I have suggested, of using hard-nosed reasoning to establish a threshold, and using more-inclusive reasoning to choose between options which, in some sense, meet this threshold, may be no improvement on the simple injunction to weigh-up the two kinds of significance in the case at hand, taking into account the kinds of factor I have mentioned. My intention has not been to argue in a detailed way for a more-inclusive decision rule; but rather to suggest that more-inclusive reasoners are not committed to ignoring information about what other persons will do. One feature of the threshold method I have sketched is that it will recommend exactly the same option as hard-nosed reasoning in many cases, and for the same reason – in these cases, no option has an unpredicted outcome which meets

[13] The threshold need not be set at maximum E(V). For example, it could be set at 0.5 maximum E(V) – in which case, in our example, more-inclusive reasoning would be indifferent between options (i), (ii), and (iii), merely ruling-out option (iv).

the threshold. Hence there are no countervailing more-inclusive considerations to set against the hard-nosed considerations. But that is the right answer. We should not expect the possibility of joint action to be salient in every case. In many cases, especially when dealing with other persons who are just, there is no scope for pure worries about acquiescence.

2. Variant forms of consequentialism

The more-inclusive view has affinities with Kantianism, but also with some consequentialist views. These consequentialist views share, in common with Kantianism, opposition to what I shall call *individual act-consequentialism.* Consequentialism, if it is a single doctrine at all, falls far short of being a full moral view. For that reason it calls for further specification, and some of the ways in which it may be further specified are in opposition to the version of consequentialism which is most discussed, which is individual act-consequentialism. In this section I shall draw some comparisons between these variants of consequentialism, and more-inclusive reasoning.

Consequentialism may be defined in a variety of ways, but I shall understand it to be the doctrine that actions are right insofar as they produce impartially good states of affairs.[14] Understood in this way, it is a doctrine about right action, with two main components: first, an impartial conception of the good, which conceives of value as independent of particular persons' projects and commitments. According to

[14] This is similar to the way Rawls, following Frankena, defines teleological theories: "[in teleological theories] the good is defined independently from the right, and then the right is defined as that which

this conception, whether or not something is good does not depend on facts about what people are committed to: the conception of the good is agent-neutral or impartial. The second component doctrine says that actions are right insofar as they promote good states of affairs, thus understood.[15]

The doctrine about right action can be understood either subjectively or objectively. Understood subjectively, it is a *rule* for action, which instructs each agent to promote the good. As such, it provides a basis for evaluating agents, which makes their *doing* such and such grounds for commendation or criticism in light of their beliefs at the time. Understood objectively, it is a *criterion* of right action, which says that what makes an action right is its promotion, in fact not in belief, of good states of affairs. As such, it provides a basis for evaluating actions themselves. Where an agent's beliefs are faulty through no fault of his own, therefore, we may condemn his action without condemning him.[16]

As it stands, consequentialism is under-specified, since it does not tell us what promoting good states of affairs amounts to. *Individual Act-Consequentialism* (IAC) spells it out as follows:

(IAC) each agent should act in the way which maximises the expected benefit of that particular act, given what others will do.

---

maximizes the good." J. Rawls, *A Theory of Justice*, p. 24. Compare Scheffler's definition, in S. Scheffler, *The Rejection of Consequentialism* (Oxford: Clarendon Press, 1982), pp. 1-2.

[15] Raz lists seven features of consequentialism. He intends this not as a complex definition, but as an indication of the variety of doctrines which are discussed under that heading. My account combines his third feature, agent-neutrality in the conception of the good, with a diluted version of his fourth feature, according to which the goal of *maximisation* is replaced, for the purposes of defining consequentialism generally, with the goal of *promoting* good states of affairs. See J. Raz, *The Morality of Freedom* (Oxford: Clarendon, 1986), pp. 268-269.

[16] In this paragraph I follow D. Parfit, *Reasons and Persons* (Oxford: Clarendon, 1984), p. 25. See also D. Regan, *Utilitarianism and Co-Operation* (Oxford: Oxford University Press, 1980), Chapter 2.

The great virtue of IAC is its scrupulous attention to the particular circumstances of the actor. It incorporates a kind of practical realism, requiring agents to be willing to act contrary to habit or prejudice, if doing so would be beneficial in the circumstances. In that sense, it is a very critical doctrine, which is difficult to reconcile with any rival moral rules or principles.[17] It is a generalisation (carried to extreme lengths) of the pragmatic attitude: 'that may work elsewhere, but does it work here?'.

Unfortunately, IAC suffers from a number of well-known problems.[18] One problem, which we may leave aside, is that *aiming* to follow IAC may be a poor way of *following* it – just as aiming to promote one's self-interest may be a poor way of promoting that.[19] A different problem, is that everyone's actually following IAC may be corrosive of social practices which are valuable, such as the keeping of agreements. In some circumstances it may be true that no particular agent can ensure, by himself, that agreements are generally kept, and that each does better, no matter what the others do, to break his agreements. IAC, with its focus on the individual acts of each agent, recommends to each agent that he break the agreement. Yet the net effect of this may be that everyone is worse off than they would have been if a significant number of agreements had been kept.

IAC is susceptible to difficulties like this in some cases of strategic interaction, and also in cases where each person's contribution to some good is more

---

[17] Morality based on IAC is intolerant of other rules or principles, but it is consistent with a wide range of values.

[18] I shall not discuss objections to IAC's agent-neutral conception of value, which apply equally to all other versions of consequentialism, as I have defined it. For discussion, see J. Dancy, *Moral Reasons* (Oxford: Blackwell, 1993), chapters 10 and 11.

costly than the marginal increase in production of that good which her contribution achieves.[20] These difficulties have been considered sufficiently severe, amongst consequentialists, to motivate departures from IAC, as I shall indicate in a moment. But it is worth noting first a separate kind of difficulty. IAC's insistence that agents should be willing to disobey habits has been thought to run foul of facts of human psychology. One claim is that humans are not psychologically capable of choosing how to act on a purely case-by-case basis. For to do so would, in many cases, involve dropping deep personal attachments and projects, in order to promote agent-neutral value.[21] This can be construed as a point about what humans are capable of, or instead as a point about what it is right to demand of them. I shall return to the latter issue in section 3 below, in discussing Liam Murphy's views.

A different kind of point is that IAC not only makes demands of us which we are incapable of meeting, but is incompatible with the correct theory of prudential value. Griffin has argued that:

> We do not just look from an act outwards, towards its further and further consequences; the [correct] prudential value theory shows that we should also look from the whole of life inwards, to find the acts that give it is general contours. This means abandoning the narrow perspective that sees acts as making ripples, some of which admittedly last a long while . . . Some of our most important acts are not like that; they are important because they alter the

---

[19] Parfit discusses this at length, in D. Parfit, *Reasons and Persons*, Chapter 1.

[20] See J. Griffin, *Well-Being*,Chapter 10.

[21] The classic source for discussions of these issues is B. Williams, "A critique of utilitarianism", in J. J. C. Smart and B. Williams, *Utilitarianism for and against* (Cambridge: Cambridge University Press, 1973), pp. 116-117. See also Raz's critical discussion, in J. Raz, The Morality of Freedom (Oxford: Clarendon Press, 1986), pp. 284-287.

character of deliberation; they relegate act-by-act deliberation to a relatively small segment of life.[22]

If this argument is correct, IAC may be shown to have too narrow a focus, even if we ignore worries about the demands it places on us, and the practical problems it may land us in. This is a theoretical objection to IAC: according to the objection, IAC is incompatible with a true theory.

It is possible for advocates of IAC to shrug-off the practical problems which their theory may land them with. They may say that the collective irrationality of the result of acts which are individually rational by IAC's lights, has *no* implications for our understanding of individual rationality. According to this view, it is just a lamentable fact of life that individually rational acts may combine in this way to produce undesirable outcomes – the fault is not with the theory of rationality (IAC), but with human circumstance. I have some sympathy with this view.[23] But our prime concern is in investigating the territory, and seeing how more-inclusive reasoning fits in; it is not with pressing the case for one or another account of practical rationality as hard as possible. In any case, the response I have just mentioned has no force against the complaints about demandingness or compatibility with correct value theory. So I shall just assume that these objections serve to show the interest of variant forms of consequentialism, and try to sketch the possible alternatives.

---

[22] Griffin, *Well-Being*, p.202.

[23] I agree with its general thrust, if not with its application in detail to this case. The general thrust is that we should not require of a theory of practical rationality that it be able in all cases to insure us against regret. That seems right, since circumstances may just be tragic. But I am not sure whether this is the correct view to take of the practical difficulties which beset IAC.

If one is moved by these objections to IAC, there are a number of directions in which one could travel. In particular, one could broaden consequentialism's focus in any of three ways. IAC focuses on individual agents, performing individual actions, and asks about the consequences of these. One could ask instead about the consequences of this agent performing a certain pattern of action across a period of time (resolute choice, or ReC), or some larger group of agents performing a certain pattern of action at this time (collective-consequentialism, or CC), or some larger group of agents performing a certain pattern of action across a period of time (rule-consequentialism, or RC). Of these theories, I shall discuss CC and RC. ReC is of interest in situations of dynamic choice, where individual agents can get into difficulty if the choices they make over time are rationally unconstrained by each other.[24] Edward McClennen has examined these issues, and notes a connection between problems of dynamic choice, understood as problems of intra-personal interaction, and inter-personal problems of interaction. But the issues are too esoteric for me to discuss here.[25]

First, then, consider CC. This theory claims that:

---

[24] That is, if the fact that my choosing X at a certain time is rational on the understanding that I will later choose Y, is not allowed to provide a reason for my later choosing Y. Diets may be like this. Eating a single cream cake will not put me off course, so I may eat a cake now on the understanding that I do not eat one tomorrow. If that rationale does not constrain my choice tomorrow, the same reasoning applies again, so I may eat another cake – but the upshot is that I go off course over time.

[25] In McClennan's work, they centre on the merits of rival versions of axiomatic decision theory. In brief, his argument is that a certain kind of pragmatic argument for two axioms, the 'weak ordering principle' and the 'independence principle', fails, because adherence to such principles is not necessary to avoid the pragmatic difficulties which they are supposed to enable us to avoid, so long as we are willing to adopt what McClennen calls resolute choice. See E. F. McClennen, *Rationality and Dynamic Choice. Foundational explorations* (Cambridge: Cambridge University Press, 1990), Chapter 1 especially. He discusses the analogue with inter-personal choice problems at pp. 256-264. Griffin's discussion of 'global' versus 'local' application of consequentialism is also of relevance. See J. Griffin, *Well-Being*, pp. 195-206.

(CC) Each agent should act in the way which is assigned to her by the pattern of acts, performed by some group of agents, with the maximum expected benefit given what outsiders will do.

CC is related to IAC in the following way: for each individual, it separates *other agents* into two categories, *insiders* and *outsiders*, and, in effect, treats the insiders as a single agent, whose action is to be guided by IAC. Each individual insider then must perform the action which is assigned to her by the pattern chosen for the group as a whole. Consider the following example:

		Other person	
		a	b
The	A	10	4
Actor	B	4	7

Table 6: Collective Consequentialism illustrated

(The numbers show agent-neutral payoffs for the different outcomes.) IAC faces a well-known difficulty in cases like this, where there is more than one equilibrium. The problem is one of *co-ordination*: there are two optimal outcomes (A/a and B/b), in the sense that there are two outcomes which result from each player choosing the best strategy given the other's strategy. If the other person chooses (b), the actor does

best to choose (B); if the other person chooses (a), the actor does best to choose (A). The problem with IAC is that it seems incapable of giving the agents reasons to settle on one equilibrium solution rather than another – because the actions of each are selected independently. It is an *indeterminate* theory, in the sense that it fails to select a unique strategy.[26]

In contrast, CC is obviously capable of selecting not just one equilibrium rather than the other, but the optimal equilibrium, in cases like this. For, according to CC, each actor should choose the option which is assigned to her by the pattern of acts (amongst the relevant group) with the maximum expected benefit. In this case, the pattern of acts with the maximum expected benefit, treating the other person as part of the relevant group, is (A/a); hence, the actor should choose (A), and the other person, if she follows CC too, should choose (a). In this case at least, CC seems to do better than IAC.

But CC is a crude theory. It departs from IAC in a gross way, and sacrifices that theory's virtue of sensitivity to particular circumstances. Communities in which everyone follows CC do better in co-ordination games than do communities in which everyone follows IAC. But this superiority has a ludicrously narrow scope: not because there are few situations which correspond to the classic structure of co-ordination games, but because it applies only where *everyone* in some community follows one or the other theory. *That*'s a much more unrealistic assumption than the assumption that real situations have the appropriate structure. In almost every real case, some will follow one theory, some another, and in these cases, there is no reason to think that followers of CC will do better than followers of IAC. It is only in that

[26] See M. Hollis and R. Sugden, "Rationality in Action", *Mind* 102 (1993), pp. 10-17.

range of worlds which are picked-out by universal performance of one theory or another, that CC is guaranteed to do better than IAC.

In cases where others are not reliable CC followers, CC is vulnerable to the charge of recklessness which was levelled at Kant's ethics. The charge in this case is that there is no reason to assume that other persons will be followers of CC. There is a version of CC which answers this charge of recklessness, however, and that is *Co-operative Consequentialism*, or CoC.[27] CoC provides a rationale for CC, and at the same time narrows its application, by considering the conditions in which co-operation is possible. In CoC, potential co-operators are the insiders, other agents the outsiders, and in following CoC each actor must endeavour to identify the members of these different groups.

This modification reclaims some sensitivity towards the particular circumstances of the agent. Agents are enjoined to co-operate where possible, and not to do so where there is no prospect of others co-operating. In particular, there must be some willingness on the part of potential co-operators to do what is necessary for co-operation, and they must all have sufficient knowledge and rationality to calculate the best group pattern, and to converge on a shared strategy. Of course, satisfaction of these conditions does not guarantee that someone who follows CoC would do better than someone who follows IAC, in any particular situation. To the extent that she would do predictably worse, CoC is still vulnerable to the charge of recklessness. But at least CoC tailors its application to the particular circumstances of agents, in seeking and using information about the possibility of co-operation with other agents.

---

[27] On CoC, see S. L. Hurley, *Natural Reasons*, Chapter 8; S. L. Hurley, "Newcomb's Problem, Prisoners' Dilemma, and Collective Action", *Synthese* 86 (1991), pp. 173-196; S.L. Hurley, "A new take from Nozick on Newcomb's Problem and Prisoners' Dilemma", *Analysis* 54: 2 (1994), pp. 65-72; D. Regan, *Utilitarianism and Co-operation* (Oxford: Clarendon Press, 1980), Chapter 8 especially.

CoC differs from more-inclusive reasoning in two main ways. First, it is equivalent to IAC in cases where others are not willing to co-operate – since the group of insiders consists, in those cases, merely of the actor, and the group of outsiders includes all the other agents who are practically relevant. In general, CoC differs from IAC in discriminating between two classes of other agents: those who are potential co-operators, and those who are not. With regard to the former group, CoC finds prediction of their behaviour inappropriate, since the assumption that they are potential co-operators amounts, as we've seen, to the assumption that they would do what is necessary for effective co-operation. So predictions about these other agents' behaviour are suspended, insofar as we consider co-operating with them.[28] But with respect to outsiders, CoC treats other agents as IAC does. We decide how to act by picking the option with maximum expected effectiveness, given our best predictions about their behaviour. In contrast, more-inclusive reasoning does not tie the appropriateness of prediction to willingness to co-operate, or membership of some relevant group. It supposes only that we weigh the importance of our vulnerability to others – *all* others, regardless of their willingness to co-operate – against the importance of the possibility of acting jointly with them.

Secondly, even where the possibility of co-operation is present, CoC differs from more-inclusive reasoning. For under CoC, the class of outcomes which is considered as relevant is still restricted according to hard-nosed assumptions. It is

[28] In fact, there are logical difficulties in using predictions about potential co-operators' behaviour to decide upon a course of action, as Regan shows. D. Regan, *Utilitarianism and Co-Operation*, pp. 139-140 and Chapter 9. The difficulty is that what counts as co-operating depends on the membership of the group of co-operators (since the latter helps determine the best pattern of behaviour amongst the group), whilst the reverse is true also. This interdependency, between optimum strategy and membership of the group of co-operators, makes prediction of the others' behaviour, based on what they have reason to do, a logically inappropriate basis for determining the membership of the group.

restricted to those we think might be realised. In cases where co-operation is possible, CoC differs from IAC because it considers as relevant things which the group of co-operators might do, together, and the group may be able to secure outcomes which no individual could (such as co-ordination). But the outcomes which CoC considers to be relevant are still restricted to those which, we believe, might come about – whereas more-inclusive reasoning may treat as relevant even outcomes which we're *certain* will not occur.

The practical difference between CoC and more-inclusive reasoning in this respect may emerge in two kinds of case. In the first kind of case, we predict that some outsiders will not perform a certain action, but more-inclusive reasoning considers its occurrence to be relevant anyway. Cases of this kind are common in political life – in international negotiations, for example, or in wage bargaining with employers. We, the members of a trades union, may be the group of insiders as far as I'm concerned, whilst the employer is an outsider. According to both CoC and more-inclusive reasoning, I should suspend my predictions of other insiders' behaviour when deliberating; but according to CoC I should use my best predictions of the employer's behaviour. Yet these are just the kinds of case in which considerations of acquiescence are likely to be relevant, and in which pure worries about acquiescence may have weight. A more-inclusive reasoner can explain the relevance of the possibility of the employer paying a fair wage, when we are certain that he will not do so. Someone who follows CoC cannot do this, since she treats outsiders as IAC treats all other agents, according to hard-nosed assumptions.

For what they have reason to do depends on the membership of the group of co-operators. It is not an independent variable which can be used to solve the membership issue.

In the second kind of case, some of the insiders are what we might call *imperfectly willing potential co-operators*. They have a general willingness to co-operate, but they are irrational or weak-willed, or perhaps they take themselves to be under certain moral constraints, such that we know that there are some things they won't do for co-operation's sake. In cases like this, as in all other cases, more-inclusive reasoners will weigh-up their significance in terms of our vulnerability, against their significance in terms of the possibility of joint action. But adherents of CoC cannot be so flexible. They must assign these people to one of their two categories of other persons: they are either insiders or outsiders. There is no provision in CoC for imperfect partners. (I shall return to this point about flexibility later in this section.)

Finally in this section, let us consider briefly rule-consequentialism, or RC. RC claims that:

(RC) Each agent should act according to the rule which has the best consequences on the assumption of universal compliance.

Like CC, RC differs from IAC in evaluating acts not one-by-one, according to their individual consequences, but instead according to their consequences when considered along with some other acts. But RC generalises in a different way than CC does. CC asks: in this particular situation, which *pattern* of acts amongst the insider group would produce the best outcome, given the behaviour of outsiders? RC asks: in this situation and *all others like it* in relevant ways, which *type* of act would produce the best outcome? Thus RC does not depend on a division of the world of *agents* into

potential co-operators and the rest, but instead on a division of the world of *acts* into those which are of the same type as this one, and those which are not.

What kind of theory RC turns out to be depends largely on the specification of types of act. A two-part argument has often been made to the effect that (i) fine distinctions between types of act are needed to avoid unwanted implications of RC across a tolerably realistic variety of cases, and (ii) once these are admitted, they necessarily proliferate in such a way that RC collapses into IAC, since it ends up treating each particular act as *sui generis*.[29]

But though the first half of this argument is no doubt right, the second half seems to be mistaken. The thought that RC must inevitably collapse into IAC seems to rest on one or both of two thoughts: either the thought that nominalism is generally correct, so that reference to *types* of anything is merely reference to *names*; or the thought that someone who holds RC must really hold IAC, deep down, so that he cannot genuinely resist the pressure to maximise the benefits of each individual act. But neither of these thoughts is compelling. Nominalism goes wrong by failing to allow that some differences between things are more significant than others: it sees all differences as on an equal footing, at some philosophical level. And the second thought just fails to take RC seriously enough. I shall speculate in a moment why that may be so.

Let us assume that RC is theoretically viable (able to stand on its own feet, independently of IAC). Nonetheless, it is practically objectionable. For, like CC, RC departs from IAC in too gross a way, and sacrifices too much sensitivity to the agent's particular circumstances. In effect, it imports into deliberative contexts the

assumptions of ideal theory about the range of relevant possibilities. The assumption of universal compliance is appropriate to ideal theory – to thoughts about perfectly just states, or perfectly democratic societies, for example. As such, it is appropriate also to applications of ideal theory, as when we judge that some existing state is not ideal in some respect. But it is not appropriate to deciding what to do.

As I stressed in Chapter Four, the range of relevant possibilities is likely to be much broader in ideal theory than in deliberation. That's because a possibility must be irrelevant for some *general* reason, if it is irrelevant to ideal theory; but in deliberation we typically have a good deal of information about the actor's specific circumstances, including information about what others are likely to do. It's unacceptably cavalier simply to ignore that information, and to act on the assumption of universal compliance with a rule. More-inclusive reasoning doesn't do that – it recognises two distinct types of significance of other actors, and does not try to resolve them into a single type. Hence it does not ignore our predictions about what others will do; it sets that information alongside information pertinent to their significance for us a potential joint actors.

This feature of RC generates unacceptable implications in many situations. A plausible rule, judged on the assumption of universal compliance, is that *parents should look after all and only their children.* The value of familial love would be served by universal adherence to this rule. But in cases where we know that many parents are not, in fact, going to adhere to it, it is morally wrong to adhere to this rule. Sometimes parents, and others, should look after other people's children, even if that means undermining the family unit, because not to do so poses too great a danger to

---

[29] For discussion of a version of this argument, see J. L. Mackie, *Ethics: Inventing Right and Wrong*

the children's welfare. Consider a different case. *Never harm the innocent intentionally* is a very plausible injunction, on RC grounds – but if we know that some powerful group is going to depart from it, it may be morally required for us to do so too.[30]

The real problem with RC is not, I suspect, any theoretical instability caused by the indistinctness of its rationale from that of IAC or the truth of nominalism. It is the practical problem that its implications are unacceptable because it is insufficiently sensitive to the particularities of agents' circumstances. It is this which puts pressure on adherents of RC to make fine distinctions between types of act, each of which is subject to a different rule, in an attempt to reduce the unacceptability of RC's implications. For this practical reason, I suspect, RC looks as if it is fated to collapse into IAC. And I suspect that it is this practical problem which has led critics to suppose wrongly that RC is not really theoretically distinct from IAC.

Let me sum up the claims of this section. From the perspective of a more-inclusive reasoner, CC and RC share two faults, despite their differences from each other. The first fault is that they are not sufficiently different from IAC, since they do not treat as relevant any outcomes which we are certain will not occur. Someone who recognises the significance of others as potential participants in joint action will think that some such outcomes should be treated as relevant. Thus, for example, CoC (a variant of CC), despite its sophistication in dividing other agents into insiders and outsiders, treats our best predictions as sufficient to rule-out possible behaviour on the part of outsiders as irrelevant. But in the case of wage bargaining and other cases like

---

(Harmondsworth: Penguin, 1977), pp. 136-140.

it, we may well want to treat outcomes which are ruled-out by our best predictions as nonetheless relevant. The issue about which outcomes we treat as relevant remains even if we are willing to treat *us as a group* as the relevant unit of agency in deliberation.

The second fault is these theories' simplistic response to the two different kinds of significance of other persons. I have not argued that the variant forms of consequentialism are motivated by recognition of the significance of others as potential participants in joint action. But a more-inclusive reasoner is likely to think that that is part of their motivation; that one of the things which contributes to their plausibility is, for example, concern with issues of acquiescence which have their real root in the possibility of joint action. If that is true, then RC and CC (and CoC) look too simple. For they try to achieve what cannot be achieved, which is to deal with the two-fold significance of others using a single decision-procedure. In the case of RC, the procedure is to divide actions into types, and consult rules for each type, where the rules are evaluated for their consequences on the assumption of universal compliance. In the case of CC, including CoC, the procedure is to divide agents into two types, insiders and outsiders, and to employ our best predictions for one type whilst suspending them for the other. This method faces pressure from two directions: it is bound to ignore our vulnerability to insiders, and to ignore the possibility of joint action with outsiders.

These claims ignore the motivations for the variant forms of consequentialism which are independent of considerations of acquiescence, and so they do not amount

---

[30] Nagel discusses a case in which the only way of extracting some vital information from a prisoner is to torture him. See T. Nagel, "War and Massacre", in his *Mortal Questions* (Cambridge: Cambridge University Press, 1979), p. 54.

to a refutation of those theories' merits. But it seems to me to be helpful to see these theories in this light, and to think of them as the result of valiantly trying to devise a single decision procedure capable of responding to the two kinds of significance of others. For then we have a sympathetic diagnosis of their failings, such as they are. They fail because they do not recognise the two different ways in which others may be significant for us as deliberators, and so cling onto the hope that a sufficiently sophisticated division between types of actor or action will provide the theoretical structure necessary to explain our considered judgements in cases of acquiescence.

Hence the line I've been following is one of pluralism. There is, admittedly, a danger in moral and political philosophy of resorting to pluralism too often or too readily – it is the same danger that nominalists fall prey to, of treating every difference as a difference in kind. But sometimes things really are of different kinds.

3. Murphy's views on demandingness

The final exploratory contrast I want to make is between more-inclusive reasoning and some recent discussion of the demandingness of morality. One of the criticisms which has been levelled at consequentialism, in its different forms, is that it is too demanding of agents, requiring too much of them. The critics complain that one way a moral theory can go wrong is by requiring that we behave in ways which are quite alien to us. Sophisticated versions of this complaint are not mere grumbling at the heavy burdens of consequentialist morality, but rather say something like this: moral theory mustn't ignore the most basic facts about human capacity and motivation, or it

will not describe a *human* morality, but at best a morality for some other kind of agent.[31]

Consequentialists can respond to such charges in two broad ways. One thing they can do is to deny the validity of that kind of criticism, by denying that the demandingness of a moral theory can be evidence of its truth or falsity. According to such a view, what morality may properly demand is not up for argument; it is something to be settled by tracing the implications of the true moral theory, where truth of moral theories is assessed on independent grounds. I think that this response is well-aimed if the complaint is simply that it would be burdensome to accept the truth of a demanding moral theory. But I just suggested that sophisticated versions of the complaint aren't like that – they express, perhaps inarticulately, the worry that we can go wrong in moral theory if we simply ignore facts about human motivation and capacity, and end up discussing what morality would require of a different type of creature than us. (This may be thought of as a very high-level worry about the range of relevant possibilities. The thought is that consequentialism goes wrong by treating as relevant some possibilities which really are irrelevant, since they are incompatible with some very basic facts about humans.)

The second thing consequentialists can do is to try to show that their favoured version of consequentialism doesn't demand too much, or make the wrong kind of demands. This is the strategy pursued by Liam Murphy, in his recent discussions of demandingness.[32] Murphy agrees that consequentialism is vulnerable to worries about

---

[31] On the general topic of demandingness, see S. Scheffler, *Human Morality* (New York and Oxford: Oxford University Press, 1992). Scheffler distinguishes between different worries about the demandingness of morality, such as worries about its *scope*, or worries about its *overridingness*, or its *stringency*. See *ibid.*, Chapter Two.

[32] See L. B. Murphy, "The Demands of Beneficence", *Philosophy and Public Affairs* 22:4 (1993), pp. 267-92; see also Tim Mulgan's response, T. Mulgan, "Two Conceptions of Benevolence", *Philosophy*

demandingness, but seeks to develop an alternative 'principle of beneficence' – in our terms, a variant of consequentialism – which escapes the objection.[33] He first considers what he calls a *Limited Principle* of benevolence, according to which morality's demands are constrained by some limit. This limit may be described in two ways: either by a Schefflerian prerogative, in which each agent is permitted to give her own interests disproportionately greater weight than those of others, when assessing the value of possible outcomes; or by means of some fixed absolute threshold, beyond which morality cannot make demands.[34]

According to the former method, the maximum legitimate demands of morality increase as the amount of good to be done increases; according to the second, they are fixed at some level or other, independent of the amount of good to be done. Murphy finds both ways of describing a limit objectionable:

> We seem to believe, on the one hand, that the demands of a principle of beneficence cannot increase indefinitely, as the amount of good to be done increases. This is why we reject a fixed multiplying factor in favor of a simple upper limit. On the other hand, however, the multiplying factor is attractive precisely because it does not decouple the extent of the demands of a principle of beneficence from the amount of good to be done.[35]

---

and Public Affairs 26:1 (1997), pp. 62-79; and Murphy's reply, L. B. Murphy, "A Relatively Plausible Principle of Beneficence: Reply to Mulgan", *Philosophy and Public Affairs* 26:1 (1997), pp. 80-86.

[33] See Murphy, "The Demands of Beneficence", pp. 272-277. Murphy frames his discussion in terms of 'principles of beneficence', but he notes that what he calls the *Simple Principle* of beneficence is equivalent to consequentialism (in particular, what I have called IAC). *Ibid.*, p. 268.

[34] Murphy, "The Demands of Beneficence", pp. 274-277. Scheffler's idea of an agent-centred prerogative is presented in S. Scheffler, *The Rejection of Consequentialism* (Oxford: Clarendon Press, 1982), Chapter Two especially.

According to Murphy, we seem to think both that the maximum demands should increase with the amount of good to be done, and that they should not.

His tactic in the face of this apparent contradiction is to distinguish two sources of high demands: (i) impersonal causes of imperfections in the world, and (ii) imperfections due to partial compliance with the principles of morality on the part of other persons. His suggestion is that an acceptable principle of beneficence would be sensitive to (i) but not to (ii) – that is, it would require more of persons for reasons to do with impersonal sources of harm and suffering, but not for reasons of others' immoral behaviour. Thus the thought that there is some *limit* to what morality may require of us is misplaced. According to Murphy, the simple principle's real mistake lies in supposing that we may be required to take on other people's shares of the demands of beneficence.[36]

The result of this analysis is what Murphy calls the *Cooperative Principle* of beneficence, according to which, crudely, what agents are required to do is limited, in terms of the sacrifice it imposes on them, to "the level of sacrifice that would be optimal if the situation were one of full compliance".[37] This is a 'co-operative' conception of beneficence because it conceives of beneficence as a joint project and joint responsibility, such that the demands of beneficence are subject to considerations of fair shares.

Murphy's suggestion has some similarities with both more-inclusive reasoning and the variants of consequentialism we examined in section 2. All of these

---

[35] Murphy, "The Demands of Beneficence", p. 276.

[36] He writes: "We should do our fair share, which can amount to a great sacrifice in certain circumstances; what we cannot be required to do is other people's shares as well as our own." *Ibid.*, p. 278.

[37] *Ibid.*, p. 280.

theories make use of counterfactual assumptions about the behaviour of other persons in specifying what agents should do.[38] But whereas more-inclusive reasoning, RC, and CC specify an *action* as part of a pattern of group behaviour which may involve unpredicted behaviour on the part of others, Murphy's conception specifies *a maximum required level of sacrifice* on the part of the agent whose action is being deliberated about. The difference is that between deciding how much sacrifice to make, and deciding what to do.

Murphy's theory is open to strong objections. As Mulgan points out, it delivers strongly counter-intuitive results in some situations. Mulgan asks us to consider different possible causes of famine, some of which are due to non-compliance with correct moral principles on the part of other persons, and others of which are due to natural causes. He points out that Murphy's principle will require us to help in the latter kind of case, but not in the former. Surely, he says, our reactions should not vary in this way between the cases.[39] What should command our attention is the suffering involved – which, by hypothesis, is the same in each case. In light of this, the causes of the suffering are relatively unimportant.

Murphy himself provides what may be a stronger counter-example to his theory. Suppose two people are walking through a park on the way to the airport. Two children are in danger of drowning in the lake, and each passer-by could save one child and still get to the airport on time, but if either were to save two, she would miss her flight. Now just suppose that one of the passers-by is unwilling to save any

---

[38] A relevant discussion, particularly in connection with justice, is C. Bertram, "Principles of Distributive Justice, Counterfactuals and History", *Journal of Political Philosophy* 1:3 (1993), pp. 213-228.

[39] Mulgan, "Two Conceptions of Benevolence", p. 71. Mulgan notes that the example must be modified to take account of Murphy's insistence that the cooperative principle deals only with *future* full compliance, but this need not concern us.

children; what should the other do? Murphy's principle requires her to save just one child, since that action corresponds with the optimal sacrifice under full compliance, even though she could save the other at the cost of missing her flight. He concedes that this is the wrong answer.[40]

Murphy's response is to invoke a distinction between "immediate and distant sources of demands." He suggests that the co-operative principle remains correct in general, as applied to potential beneficiaries with whom we are not directly acquainted – but that in the case of the drowning children, we are under special obligations to do more than is required by the co-operative principle, because of their proximity to us. Hence the counter-example shows only the existence of these additional special obligations.[41] But I think that the problem with Murphy's principle is quite general.

The reason is that Murphy's laudable ambition – of separating the two kinds of source of high moral demands – cannot be carried through. I think it was very useful for Murphy to point out that there are these two sources, one having to do with impersonal imperfections in the world, one having to do with other persons' failings. I think he's right too that we would want the demands morality places on us to vary in *somewhat* different ways with these two variables. But I do not think the hygienic cut between the two kinds of factor, which his principle requires, can be made.

The reason is the familiar one, that it is difficult to sort facts about the world into those which are products of natural causes, on one hand, and those which are products of human actions or omissions, on the other. This is clearly true in cases of famine. There is no clear distinction between famines which are caused by natural

---

[40] Murphy, "Two Principles", pp. 290-292.

disasters, and famines which have human causes -- because most natural disasters aren't sufficient to guarantee famine. They must be augmented by human failures. I'm not arguing that there is *no* contrast to be made between high demands which are the result of natural causes, and those which are the result of human actions or omissions. But I think Murphy's proposal would require a very clear distinction to be made, and that such a level of clarity is not available. Our moral intuitions, which should be accorded *some* weight on any plausible meta-ethical view, do not reflect any such clear distinction (in the case of the drowning children, for example). For that reason, a theory whose rationale is its sharply different sensitivity to the two kinds of cause, will inevitably run up against those intuitions.

4. Conclusion

The hard-nosed view about the range of alternatives which are relevant to political argument, is built-in to Individual Act-Consequentialism. Rejection of the hard-nosed view is thus an attack on IAC itself, and leaves more-inclusive reasoning, as an idea, in the company of anti-consequentialist theories, such as Kant's, as well as variant forms of consequentialism. These theories share in common the view that prediction does not have the role in practical (including political) reasoning which IAC assigns to it.

But these theories are also importantly different from *each other*. I have tried to develop my account of more-inclusive reasoning by discussion of them and their differences, and by indicating where they diverge, practically and theoretically, from

[41] Murphy, "Two Principles", pp. 290-292.

more-inclusive reasoning. I have not tried to develop a conclusive criticism of any of the theories discussed. But I have claimed some advantages for more-inclusive reasoning along the way. The first advantage, comparing it to CC, CoC, and RC, is that it departs from IAC in a *sufficiently radical* way. Those other theories all agree that IAC has too narrow a view of the evaluation of options, but they fail to recognise some possible outcomes which are relevant to that evaluation. This is just to repeat the claim of Chapter Five, that outcomes we're certain won't happen may be relevant – and to insist that the familiar variants of consequentialism don't capture this thought.

The second advantage, comparing more-inclusive reasoning to Kantianism, CC, CoC, RC and Murphy's view, is that these other theories depart from IAC in *too simple* a way. Staking this claim depends on a willingness to see these theories as all responding, in part at least, to the worries about acquiescence which have been our concern. If that's right, then they can be criticised for failing to come to terms with the fact that other persons are significant for us as deliberators in two essentially distinct ways: as persons to whom we are vulnerable in our activity, and as persons with whom we may act jointly. Each of the theories discussed may be interpreted as responding to this latter kind of significance, even if they do not do so explicitly. But they all go wrong in sacrificing too much of IAC's sensitivity to particular circumstances. They do that because they try to develop a single decision procedure, and one can't respond to both kinds of significance at the same time in every case.

The proper approach, I've suggested, is to give explicit recognition to the two different kinds of significance, which must then be weighed against each other. In section 1, I sketched a method for doing this, which involved setting thresholds for

options to meet. That method may be too intuitionistic, and there may be a better way to spell-out more-inclusive reasoning. But its real importance lay in emphasising that more-inclusive reasoners needn't simply ignore information about what other persons will do. In thinking about the range of relevant possibilities, they insist that we distinguish between two realms of possibility: that which will happen, which is pertinent to thoughts about vulnerability, and is addressed by our best predictions, and that which could happen, which is pertinent to thoughts about joint action, and is addressed by thoughts about what other persons could do.

Chapter Seven

Conclusion

This thesis has examined a cluster of related topics. First, we investigated the relationship between the concepts of *desert*, *responsibility*, and *justice*. Then we examined the more specialised question, of whether what justice requires depends on what people could do, in an agency-implicating sense. This led to an examination of the rationality of acquiescence, which in turn led to discussion of the nature of deliberation. What connects these topics is a preoccupation with the relationship between our understanding of persons as agents, on one hand, and our understanding of the nature of practical (including political) reasoning, on the other.

In this final chapter I would like to do two things. In section 1 I shall provide a brief summary of the arguments made and the conclusions reached. Then in section 2 I shall describe a major limitation of the work, which is its reliance on an undeveloped account of the nature of action. Throughout I have invoked the idea that there is a contrast between the understanding of persons and actions which our best predictive theories provide, and the picture of agency which is presupposed, I have suggested, by many of our attitudes towards action, as well as by our ordinary thoughts about

persons as agents. In section 2 I shall indicate some of the work which would need to be done in order to flesh-out that vital contrast.

1. Summary

There are a number of views one can take about the conceptual relationship between responsibility and justice. One of these, which I have not discussed at any length, says that the concepts are mutually independent; which is to say that neither features essentially in the true theory of the other. Utilitarians seem likely to accept this view, insofar as they are interested in justice. But I have been more interested in views which find a connection between the two concepts.

It is important which view one takes. It is important not only because one's view of justice and responsibility will depend, obviously, on one's view of the relationship between them, but also for theory-building reasons. One of the things which someone who sets out to develop a theory of justice might be interested in, for purposes of general orientation, is the relationship between justice and responsibility. And, of course, the same is true of someone who sets out to develop a theory of responsibility.

Responsibility is a puzzling concept. On a common-sense view, it combines an interest in agency, naturally oriented towards causal explanations of action, with elements of moral commitment. In order to explain what responsibility is, on this view, we should expect to have to understand both causality and human action, on one hand, and moral principles, having to do with excusing conditions, mitigating factors,

and so on, on the other. We might express this common-sense view, crudely, in terms of a formula: responsibility equals causality plus morality.

If we accept the common-sense view, broadly, we can explain why the concepts of justice and responsibility seem to be connected. One element of our thoughts about justice has to do with *fairness*. Fairness has to do with responding to people even-handedly; and a large part, though not all, of our response to persons consists of responding to what *they do*. So if we want to be fair, we must form some picture of what it is that people do – some picture, that is, of which events and facts about their circumstances are attributable to them, rather than to other causes. But in doing so, we will also want to make a division *amongst* those facts and events which are attributable to them – between those which are, and those which are not, subject to special excusing conditions. Fairness discriminates between those harms I cause to myself, depending on whether or not I satisfied full conditions of agency at the time. This dual concern, with causality and excusing conditions, takes us straight into the territory of responsibility.

Accepting the common-sense view, then, we get the following picture: responsibility is conceptually independent of justice, but justice depends, in part, on responsibility. For judgements of responsibility are involved in judgements of fairness. Justice is not necessarily all about responsibility; but the concept of responsibility brings some independent content to the concept of justice. Hence, in thinking about justice, we are bound to be led into thoughts about the nature of human action, since these are largely what judgements of responsibility are about. Justice is in hock to metaphysics.

But recent currents in thoughts about justice, and to some extent, thoughts about responsibility, reject this picture of political philosophy's indebtedness to a metaphysical concept of responsibility.[1] They assert instead the opposite dependency: concepts which we thought had essential metaphysical content, such as desert and responsibility, are really wholly moral notions. Someone's being responsible for something, on this *naturalist* view, is a matter of their being the legitimate bearer of its costs (or recipient of its benefits), where that is decided by reference to a theory of justice or fairness. Similarly, someone's being deserving is a matter of their having behaved in a certain way against the background of (more or less) just institutions. Naturalism differs from the common-sense view in the explanation it gives of our thoughts about justice. Common sense says that we get an independent grip on what costs it is legitimate to let persons bear, by understanding the extent to which those persons are the authors of their circumstances. Naturalism says that we have an independent theory about which costs it is legitimate to let persons bear, which we can use, if we wish, to reconstruct talk about desert and responsibility.

Naturalism, in this sense, is a very attractive doctrine, but its attractions do not amount to arguments for it. It is attractive because it offers a new way of getting to

---

[1] I have tried to describe this in Chapter Two. The most prominent exponent of the strategy of trying to develop a theory of justice which is as independent as possible from metaphysical controversies, is of course Rawls. This tendency is perhaps more marked in *Political Liberalism* than in *A Theory of Justice*. But it has become commonplace to express the hope, if not the conviction, that political philosophy can do without metaphysics. Again, this is for understandable reasons, as I tried to explain in Chapter Two. On the other side, the most notable advocate of the view that responsibility is largely a moral, and not a metaphysical, notion, is Scanlon. But see also Ripstein's discussion of the concept -- written as a rejoinder to Scheffler's observation that recent political theory shuns reliance on agency-implicating concepts. Ironically, Ripstein's understanding of responsibility places most emphasis on 'standards of care', which is, I take it, a version of the idea that we can explain responsibility in terms of broadly moral notions, rather than attempting to do the opposite. See A. Ripstein, "Equality, Luck, and Responsibility", *Philosophy and Public Affairs* 23:1 (1994). There he writes (p. 5): "Responsibility and desert do have a role to play in moral thought . . . Indeed, their role is central. But the reason they are central is not that they are somehow more basic than some other set of moral notions that they support. Instead, they are central to moral thought because they are themselves moral notions that make no sense apart from other moral notions."

grips with age-old problems concerning responsibility and desert, as well as a way of freeing theories of justice from involvement with these controversies. But these are pragmatic reasons – reasons for hoping that naturalism is true, not reasons for believing it to be so. Such pragmatic reasons have some force, I suggested, insofar as we adopt a theory-building stance towards these issues. But they do not serve to justify belief in naturalism.

Chapters Two and Three searched for reasons for and against belief in naturalism. I looked first at some simple arguments, which were quickly dismissed (2.2). Then I turned to the idea that the common-sense view is vindicated by our intuitions about equality, as discussed in the debate about what egalitarians should try to equalise (2.3 - 2.5). But this idea was not supported by examination of that debate. In this case as in others, our intuitions seem to be capable of explanation in a number of different ways; hence, they do not provide clear support for one theory rather than another.

Chapter Three examined Rawls's views on desert and responsibility. Rawls is, arguably, the leading naturalist, and our purpose in examining his attempt to develop a naturalist theory of justice was to get a better appreciation of the resources of naturalism, as well as to look for evidence of its flaws. Once again, however, we found no conclusive reasons in favour or against naturalism. But we did discover some of the different ways in which a naturalist theory may be constructed. In particular, we should distinguish between *limited* and *thoroughgoing* naturalism (3.4).

Limited naturalism asserts that agency-implicating concepts such as desert or responsibility are conceptually *interdependent* with justice. On this view, the concept of justice has some content independent of agency-implicating concepts such as desert

and responsibility, which themselves have some content independent of the concept of justice. But justice and these concepts help refine each other. (I discussed how Scanlon's remarks could be elaborated as a form of limited naturalism.) In contrast, thoroughgoing naturalism asserts that the concept of justice is wholly independent of agency-implicating concepts, though they are conceptually dependent on it. It is thoroughgoing naturalism which promises the greatest theoretical benefits, in terms of finessing philosophical problems in our understanding of agency.

Chapter Four offered an argument against thoroughgoing naturalism, however, which drew on Cohen's discussion of incentives. Any theory of justice which is minimally beneficent, I claimed, inherits the essentially comparative nature of the concept of benefit. Someone is benefited by an arrangement only if they do better under it than they do under relevant alternatives. Hence, in making judgements of benefit we rely implicitly on the existence of a range of relevant alternatives (4.3).

If we ask how the range of relevant alternatives should be identified, in any case, we discover the importance of distinguishing between two modes or types of normative political argument (4.4). In *ideal theory*, the range is likely to be very broad, since we do not want our specification of ideals to be strongly constrained by facts about the actual world (4.5). In *deliberation*, however, it is practically important to make good use of information about what is possible in the actor's particular circumstances. Hence, the range of relevant possibilities should be more strongly constrained by facts about the actual world in deliberative argument, than it is in ideal theory (4.6).

The distinction between ideal theory and deliberation is a very important one for our general understanding of political argument. In particular, we should not make

the mistake of thinking that deliberation is just a matter of working our how to travel from here to the ideal state of affairs by the shortest route. That map-reading metaphor obscures the fact that the assumptions about the range of relevant possibilities we should be operating with in deliberation, are very different from those we should be operating with in ideal theory – which makes the two kinds of exercise fundamentally different (they have, after all, different objectives). The map-reading metaphor does not correctly portray the relationship between describing ideals and deliberation.

The distinction between ideal theory and deliberation proves crucial in the argument against naturalism, too. It is, I think, fairly obvious that our best predictions about what will happen do not pick-out the range of possibilities which is relevant to ideal theory; rather, those possibilities are picked-out, in part, by reference to judgements about what persons *could do*, in an agency-implicating sense (4.5). We can see this in the particular case which Cohen discusses. The reason that the incentives argument fails, in its ideal theoretic form, is that it fails to compare how well the worst off do under a regime of incentives, with how well they do under a relevant alternative, in which the talented work just as hard for less. That alternative is relevant just because we judge that talented persons *could* work just as hard for less – but that judgement cannot be understood as a prediction about what will happen. It's a different kind of judgement, one which implicates our understanding of agency. And the same applies generally, I claimed. Any ideal theoretic argument which relies on comparison of how well people do under different arrangements will rely on agency-implicating judgements about what persons could do.

Naturalism is far harder to refute when it is applied to deliberation. Chapter Five attempted to extend the anti-naturalist conclusion to deliberative contexts. The

strength of naturalism in these contexts is its expression of a certain kind of practical realism, which is an undeniable virtue of deliberative arguments. The sense of realism in question insists on the importance of evaluating courses of action in terms of our best predictions about how the actor's environment, including other persons, would behave if those actions were carried out. If we do not evaluate courses of action in this way, we ignore the particularities of the actor's circumstance, and recommend actions on the grounds, apparently, that they have worked elsewhere. But we know that what works in one circumstance needn't work in another (5.3).

I think that we should accept the importance of predicting the behaviour of the actor's environment, including the behaviour of other persons, when evaluating courses of action. But I argued that other persons can have a different kind of significance for us when we deliberate. They can be significant as potential participants in joint action (5.4). If we are responsive to this kind of significance, we evaluate courses of action on assumptions about the possible behaviour of other persons which are *more-inclusive* than our best predictions about the future. In particular, we need not treat someone's resolve or intention not to do something, as sufficient grounds for discounting that possibility, *even if* we are certain that the resolve will be causally efficacious, and they won't perform the action in question.

If it is right to evaluate courses of action on these more-inclusive grounds, in some cases, then naturalism is false even as a doctrine about deliberation. For we must rely on agency-implicating judgements about what other persons could do, rather than just our best predictions, in deliberation as well as in ideal theory.

The argument for the more-inclusive view relied heavily on the plausibility of certain intuitions about the range of reasons we have for resisting acquiescent courses

of action. I distinguished four types of worry about acquiescence (5.1), and claimed that the more-inclusive view is presupposed by the rationality of one kind, which I called *pure worries*. I accept that this argument is vulnerable, on two kinds of ground. First, one may dispute the plausibility of the intuitions in question – and claim, for example, that there is no reason not to acquiesce in the behaviour of the capricious person (5.4). Or, alternatively, one could accept the intuitions, but seek to explain them on grounds which do not invoke the more-inclusive view. There may be some such alternative explanation, though I am not aware of it.

Finally, Chapter Six turned to elaboration of the nature of deliberation on the basis of the more-inclusive view. I noted that *more-inclusive reasoning* shares opposition to *individual act-consequentialism* with other, diverse, views. I sought to explain how it might work, and when it might be appropriate, by exploring similarities and contrasts with these other views. Like Kantianism, more-inclusive reasoning sometimes recommends a course of action which would be more successful than some other option only if other persons were to behave as we predict they will not – and this remains true even in cases where we are certain of our predictions (6.1).

This has been thought to be a kind of recklessness, and a regrettable feature of Kantianism, so it may be thought to reflect badly on more-inclusive reasoning. But I noted some important contrasts between the two kinds of view, which seem to lessen the charge of recklessness as applied to more-inclusive reasoning (6.1). First, more-inclusive reasoning does not recommend that we assume that other persons will behave just as they ought, which is (effectively) what Kantianism recommends. It accepts some restriction on the range of relevant possible actions of others, due to contingent facts about their circumstances. Second, more-inclusive reasoners do not

lose sight of our *vulnerability* to other persons, and so will often judge more-inclusive assumptions about them to be inappropriate.

I made essentially the same point in discussing Liam Murphy's views on demandingness (6.3), as well as other variant forms of consequentialism (6.2). These views may be interpreted as responding, in part, to the issues of the rationality of acquiescence which motivate the more-inclusive view. But, unlike the more-inclusive view, they each try to develop a single procedure for responding to other persons. By adopting this approach, they fail in two respects to respond appropriately to other persons. They sacrifice too much of the sensitivity to the vulnerability of the actor, which is individual act-consequentialism's strength. And they also fail to recognise the relevance of some possibilities which we're certain won't be realised. Both faults have the same cause: failure to distinguish the two different kinds of significance which other persons can have for us as deliberators.

The distinction which I made, in Chapter Four, between ideal theory and deliberation leaves Chapter Five and Chapter Six on an argumentative branch of their own. The arguments of the previous chapters do not depend on the claims I make about the more-inclusive view, or the rationality of acquiescence. One is led naturally by the distinction between ideal theory and deliberation, and an interest in naturalism, to wonder whether naturalism is false not just about ideal theory but also about deliberation. I think that it is, for the reasons I have given. But the argument is less clear-cut. In the final section, I shall indicate one approach we could take to making it more secure.

2. Developing the account of agency

Throughout this thesis I have invoked the idea that there is a contrast between the understanding of persons which is provided by our best predictive theories, and the picture of them which is presupposed by our ordinary thoughts about agency, and (if I'm right) by our attitudes towards acquiescence. I have supposed, in particular, that our best predictive theories pick-out a different, and narrower, range of possibilities than do our agency-implicating judgements. Without that supposition, the contrast between the hard-nosed view and the more-inclusive view would be illusory. And it is required also in the argument of Chapter Four, that the range of possibilities relevant to ideal theory is picked-out by agency-implicating judgements rather than predictive theories.

But the contrast may seem to rest on a highly controversial metaphysical picture. Consider the following critical argument. The arguments I've made ought not to depend on the falsity of our best predictive theories. I do not think they do, so let's assume for the sake of argument that our best predictive theories are true as they stand. If they are true, and all events are causally determined, our predictive theories are capable of telling us what the future holds. Suppose they tell us that Smith will *R* at time *t*. Then it seems we cannot suppose that that Smith could *not-R* at *t*. It's causally necessary that he will *R*, and (worse than that) the fact that we *know* that it's causally necessary, makes our supposition that he could do something else unsustainable. Hence, either the arguments made here presuppose imperfect predictive theories, or they presuppose some unusual sense of 'could'.

I've already indicated that I do not want to concede that the arguments presuppose imperfect predictive theories. But there are two points to be made against the critical argument. The first is that the assumption of causal necessity is itself up for dispute – a claim of causal necessity needn't be entailed by the assumption that we have *certain predictions*. Epistemic necessity, that is, needn't presuppose causal necessity.[2] So I could affirm the consistency of my arguments with the former, without undertaking to show their consistency with the latter. It may be that the future is certain, but not causally necessary – in which case the claim that someone could do something which, we know, they will not, needn't involve an odd sense of 'could'.

The second point is that even if epistemic necessity can be shown to presuppose causal necessity, the sense of 'could' which would be involved isn't that unusual. It is a compatibilist sense of 'could', according to which the range of things a person could do (a) includes things they will never actually do, and (b) is nevertheless constrained by facts about them and their circumstances.[3] Perfect predictive theories, I'm assuming, satisfy (b) but not (a). The idea of logical possibility, on the other hand, provides us with a sense in which things are possible for persons which satisfies (a) but not (b). The sense I need specifies a realm of possibility more-inclusive than the set of all actual events, but less inclusive than logical possibility.[4]

I do not know how to specify such a compatibilist sense of 'could', beyond saying that it must satisfy these two conditions. I do not have a solution to this

---

[2] This is a general claim, proposed on the basis of an alien example: mathematical certainty does not presuppose causal necessity. But, though the example is alien, the burden of proof lies with the critic, to show why, in the case of prediction, epistemic necessity *does* presuppose causal necessity.

[3] See Chapter Six, section 1.

[4] Physical possibility, as it is usually understood, fits this description. Something is physically possible if it is not contrary to any universal physical law. That specifies a realm of possibility between logical possibility and actuality. But it is too inclusive for our purposes: winning an Olympic gold medal is a possibility beyond my reach, but there is no universal law of nature which forbids it. The appropriate realm seems to be constrained by some, but not all, facts about the actual world.

problem – but it's worth pointing out that it is a problem for others as well. Our ordinary thoughts about agency also presuppose a sense of 'could' which satisfies these conditions. It is not generally considered to be a contradiction to say that Jones went for a walk on Wednesday afternoon but could have done otherwise. And usually we do not accept that it suffices to refute the claim that 'he could have done otherwise' to point out that he did not in fact do otherwise.

It may be, of course, that our ordinary thoughts about agency are just wrong, in which case pointing out that they are in the same boat as the arguments I've made doesn't help the latter much. But it's not just our common-sense theory of agency which is in the same boat. It is a presupposition of practical reasoning, in general, that the actor could do something other than the action she will in fact perform. There might still be some purpose in deliberating, if it were true that one could do only what one will in fact do (it might make us feel better) – but it would not be the purpose which practical reasoning claims for itself. Practical reasoning too presupposes a sense of 'could' which satisfies conditions (a) and (b).

Maybe all practical reasoning rests on a mistake about what is possible. That might be true, but it seems to me that a more plausible hypothesis is that we don't understand the nature of judgements about possibility well enough to make the argument about the inconsistency of the required sense of 'could' with perfect predictive theories. So I think that it is not inconsistent to say that it is certain that Smith will *R* at *t* and that he could *not-R* at *t*. That doesn't explain the relevant sense of 'could', of course; it just pleads the case for its possible existence.

The topics I've discussed in this dissertation are united by a single common interest in the relationship between our understanding of practical reasoning and our

understanding of agency. The central claim I've made is that we shouldn't try to make political arguments without invoking agency-implicating concepts, because we cannot avoid doing so. But the arguments I've made have relied on a picture of action – in particular, on the thought that predictive theories do not get to grips with agency – which I've simply assumed. The obvious way in which they could be fleshed-out further, therefore, would be to fill-in some details of this picture.

A final thought. We do not understand practical reasoning very well if we conceive of its singularity as consisting only in its concern with ends (desires or values) and principles. In order to understanding political and other forms of practical reasoning, we need to understand also the complex ways in which we bring ends and principles to bear on the circumstances of agents, in order to arrive at recommended options. Discussions of the nature of practical reasoning should give more weight to its *descriptive* components than they typically do, for within such descriptions we often find lurking covert and contestable specifications of the range of relevant possible arrangements or actions.[5] Much disagreement, and much of the rationality of what we do, turns on our sense of what is possible.

[5] One discussion which puts the descriptive elements at centre-stage – indeed, even claims that moral reasoning consists *entirely* of appropriate description of the agent's circumstances – is J. Dancy, *Moral Reasons* (Oxford: Blackwell, 1993), pp. 111-119.

Bibliography

R. J. Arneson, "Equality and Equal Opportunity for Welfare", *Philosophical Studies* 56 (1989).

R. J. Arneson, "Equality", in R. E. Goodin and P. Pettit (eds.), *A Companion to Contemporary Political Philosophy* (Oxford: Basil Blackwell, 1993).

B. Barry, *Theories of Justice* (Hemel Hempstead: Harvester Wheatsheaf, 1989).

B. Barry, *Political Argument. A Reissue with a New Introduction* (Hemel Hempstead: Harvester Wheatsheaf, 1990).

B. Barry, *Justice as Impartiality* (Oxford: Clarendon Press, 1995).

C. Bertram, "Principles of Distributive Justice, Counterfactuals and History", *Journal of Political Philosophy* 1 (1993).

G. A. Cohen, "On the Currency of Egalitarian Justice", *Ethics* 99 (1989).

G. A. Cohen, "Equality of What? On Welfare, Goods, and Capabilities", in M. Nussbaum and A. Sen (eds.), *The Quality of Life* (Oxford: Clarendon Press, 1993).

G. A. Cohen, "The Pareto Argument for Inequality", *Social Philosophy and Policy* 12 (1995).

G. A. Cohen, "Incentives, Inequality, and Community", in S. Darwall (ed.), *Equal Freedom. Selected Tanner Lectures on Human Values* (Ann Arbor: University of Michigan Press, 1995).

G. A. Cohen, "Where the Action Is: On the Site of Distributive Justice", *Philosophy and Public Affairs* 26 (1997).

A. Colman, *Game Theory and Experimental Games. The Study of Strategic Interaction* (Oxford: Pergamon Press, 1982).

G. Cupitt, *Justice as Fittingness* (Oxford: Clarendon Press, 1996).

J. Dancy, *Introduction to Contemporary Epistemology* (Oxford: Blackwell, 1985).

J. Dancy, *Moral Reasons* (Oxford: Blackwell, 1993).

N. Daniels, "Wide Reflective Equilibrium and Theory Acceptance in Ethics", *Journal of Philosophy* 76 (1979).

S. Darwall (ed.), *Equal Freedom. Selected Tanner Lectures on Human Values* (Ann Arbor: University of Michigan Press, 1995).

R. Dworkin, "What is Equality? Part 1: Equality of Welfare", *Philosophy and Public Affairs* 10 (1981).

R. Dworkin, "What is Equality? Part 2: Equality of Resources", *Philosophy and Public Affairs* 10 (1981).

R. Dworkin, "Foundations of Liberal Equality", in S. Darwall (ed.), *Equal Freedom. Selected Tanner Lectures on Human Values* (Ann Arbor: University of Michigan, 1995).

J. Feinberg, *Doing and Deserving. Essays in the Theory of Responsibility* (Princeton, NJ: Princeton University Press, 1970).

D. Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986).

J. Griffin, *Well-Being. Its Meaning, Measurement, and Moral Importance* (Oxford: Clarendon Press, 1986).

J. Griffin, *Value Judgement. Improving our Ethical Beliefs* (Oxford: Clarendon Press, 1996).

G. Hawthorn (ed.), *The Standard of Living* (Cambridge: Cambridge University Press, 1987).

T. Hobbes, *Leviathan*, edited by M. Oakeshott (Oxford: Basil Blackwell, 1946).

M. Hollis and R. Sugden, "Rationality in Action", *Mind* 102 (1993).

S. L. Hurley, *Natural Reasons. Personality and Polity* (New York and Oxford: Oxford University Press, 1989).

S. L. Hurley, "Newcomb's Problem, Prisoners' Dilemma, and Collective Action", *Synthese* 86 (1991).

S. L. Hurley, "Justice without Constitutive Luck", in A. Phillips Griffiths (ed.), *Ethics* (Cambridge: Cambridge University Press, 1993).

S. L. Hurley, "A new take from Nozick on Newcomb's Problem and Prisoners' Dilemma", *Analysis* 54: 2 (1994).

S. L. Hurley, "Egalitarianism without constitutive luck: Incentives and Responsibility", unpublished manuscript.

S. L. Hurley, "Cohen on Incentives", unpublished manuscript.

I. Kant, *Critique of Pure Reason*, translated by N. Kemp Smith (Basingstoke: Macmillan, 1929).

I. Kant, *Foundations of the Metaphysics of Morals*, second edition, translated with an introduction by L. White Beck (New York: Macmillan, 1990).

I. Kant, *Critique of Practical Reason*, third edition, translated by L. White Beck (New York: Macmillan, 1993).

C. M. Korsgaard, "The Right to Lie: Kant on Dealing with Evil", *Philosophy and Public Affairs* 15 (1986).

C. Larmore, *Patterns of Moral Complexity* (Cambridge: Cambridge University Press, 1987).

J. L. Mackie, *Ethics: Inventing Right and Wrong* (Harmondsworth: Penguin, 1977).

I. McLean, *Public Choice. An Introduction* (Oxford: Basil Blackwell, 1989).

E. F. McClennen, *Rationality and Dynamic Choice. Foundational explorations* (Cambridge: Cambridge University Press, 1990).

S. M. McMurrin (ed.), *Liberty, Equality, and Law* (Salt Lake City: University of Utah Press, 1987).

D. Miller, *Social Justice* (Oxford: Clarendon Press, 1976).

D. Miller, *Market, State, and Community. Theoretical Foundations of Market Socialism* (Oxford: Clarendon Press, 1989).

T. Mulgan, "Two Conceptions of Benevolence", *Philosophy and Public Affairs* 26:1 (1997).

L. B. Murphy, "The Demands of Beneficence", *Philosophy and Public Affairs* 22:4 (1993).

L. B. Murphy, "A Relatively Plausible Principle of Beneficence: Reply to Mulgan", *Philosophy and Public Affairs* 26:1 (1997).

T. Nagel, "War and Massacre", in T. Nagel, *Mortal Questions* (Cambridge: Cambridge University Press, 1979).

T. Nagel, *Mortal Questions* (Cambridge: Cambridge University Press, 1979).

T. Nagel, *The View from Nowhere* (New York: Oxford University Press, 1986).

T. Nagel, *Equality and Partiality* (New York and Oxford: Oxford University Press, 1991).

R. Nozick, *Anarchy, State, and Utopia* (Oxford: Basil Blackwell, 1974).

R. Nozick, *The Nature of Rationality* (Princeton: Princeton University Press, 1993).

M. Nussbaum and A. Sen (eds.), *The Quality of Life* (Oxford: Clarendon Press, 1993).

D. Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984).

D. Parfit, "Equality or Priority?", *The Lindley Lecture,* University of Kansas (1995).

M. Phillips, "Reflections on the Transition From Ideal to Non-Ideal Theory", *Nous* XIX (1985).

A. Phillips Griffiths (ed.), *Ethics* (Cambridge: Cambridge University Press, 1993).

J. Rawls, *A Theory of Justice* (Oxford: Oxford University Press, 1972).

J. Rawls, "Social Unity and Primary Goods", in A. Sen and B. Williams (eds.), *Utilitarianism and beyond* (Cambridge: Cambridge University Press, 1982).

J. Rawls, "The Priority of Right and Ideas of the Good", *Philosophy and Public Affairs* 17 (1988).

J. Rawls, *Political Liberalism* (New York: Columbia University Press, 1993).

J. Raz, *The Morality of Freedom* (Oxford: Clarendon Press, 1986).

J. Raz, *Ethics in the Public Domain* (Oxford: Clarendon Press, 1994).

D. Regan, *Utilitarianism and Co-operation* (Oxford: Clarendon Press, 1980).

N. Rescher, *The Coherence Theory of Truth* (Oxford: Clarendon Press, 1973).

A. Ripstein, "Equality, Luck, and Responsibility", *Philosophy and Public Affairs* 23:1 (1994).

J. E. Roemer, "Equality of Talent", *Economics and Philosophy* 1 (1985).

J. E. Roemer, "Equality of Resources implies Equality of Welfare", *The Quarterly Journal of Economics* (1986).

J. E. Roemer, "Egalitarianism, Responsibility, and Information", *Economics and Philosophy* 3 (1987).

J. E. Roemer, "A Pragmatic Theory of Responsibility for the Egalitarian Planner", *Philosophy and Public Affairs* 22 (1993).

M. J. Sandel, *Liberalism and the Limits of Justice* (Cambridge: Cambridge University Press, 1982).

T. M. Scanlon, "Equality of Resources and Equality of Welfare: A Forced Marriage?", *Ethics* 97 (1986).

T. M. Scanlon, "The Significance of Choice", in S. Darwall (ed.), *Equal Freedom. Selected Tanner Lectures on Human Values* (Ann Arbor: University of Michigan Press, 1995).

S. Scheffler, *The Rejection of Consequentialism* (Oxford: Clarendon Press, 1982).

S. Scheffler, *Human Morality* (New York and Oxford: Oxford University Press, 1992).

S. Scheffler, "Responsibility, Reactive Attitudes, and Liberalism in Philosophy and Politics", *Philosophy and Public Affairs* 21 (1992).

S. Scheffler, "Individual Responsibility in a Global Age", *Social Philosophy and Policy* 12:1 (1995).

A. Sen, "Equality of What?", in S. M. McMurrin (ed.), *Liberty, Equality, and Law* (Salt Lake City: University of Utah Press, 1987).

A. Sen, "The Standard of Living: Lecture 1, Concepts and Critiques", in G. Hawthorn (ed.), *The Standard of Living* (Cambridge: Cambridge University Press, 1987).

A. Sen, *Inequality Reexamined* (Oxford: Clarendon Press, 1992).

A. Sen, "Capability and Well-Being", in M. Nussbaum and A. Sen (eds.), *The Quality of Life* (Oxford: Clarendon Press, 1993).

A. Sen and B. Williams (eds.), *Utilitarianism and beyond* (Cambridge: Cambridge University Press, 1982).

G. Sher, *Desert* (Princeton, NJ: Princeton University Press, 1987).

J. Stanyer and G. Stoker (eds.), *Contemporary Political Studies 1997*, volume two (Nottingham: Political Studies Association, 1997).

H. Steiner, *An Essay on Rights* (Oxford: Blackwell, 1994).

H. Steiner, "Choice and Circumstance", unpublished manuscript.

G. Strawson, "Consciousness, Free Will, and the Unimportance of Determinism", *Inquiry* 32 (1989).

P. F. Strawson, "Freedom and Resentment", in his *Freedom and Resentment and other essays* (London: Methuen and Co., 1974).

L. Temkin, "Inequality", *Philosophy and Public Affairs* 15:2 (1986).

P. Van Parijs, *Real Freedom for All. What (if anything) Can Justify Capitalism?* (Oxford: Clarendon Press, 1995).

D. Wiggins, "Deliberation and Practical Reason", in his *Needs, Values, Truth* (Oxford: Basil Blackwell, 1987).

B. Williams, "A critique of utilitarianism", in J. J. C. Smart and B. Williams, *Utilitarianism for and against* (Cambridge: Cambridge University Press, 1973).

C. Woodard, "Responsibility, Desert, and Liberal Theories of Justice", in J. Stanyer and G. Stoker (eds.), *Contemporary Political Studies 1997*, volume two (Nottingham: Political Studies Association, 1997).

R. Young, "Egalitarianism and Personal Desert", *Ethics* 102 (1992).