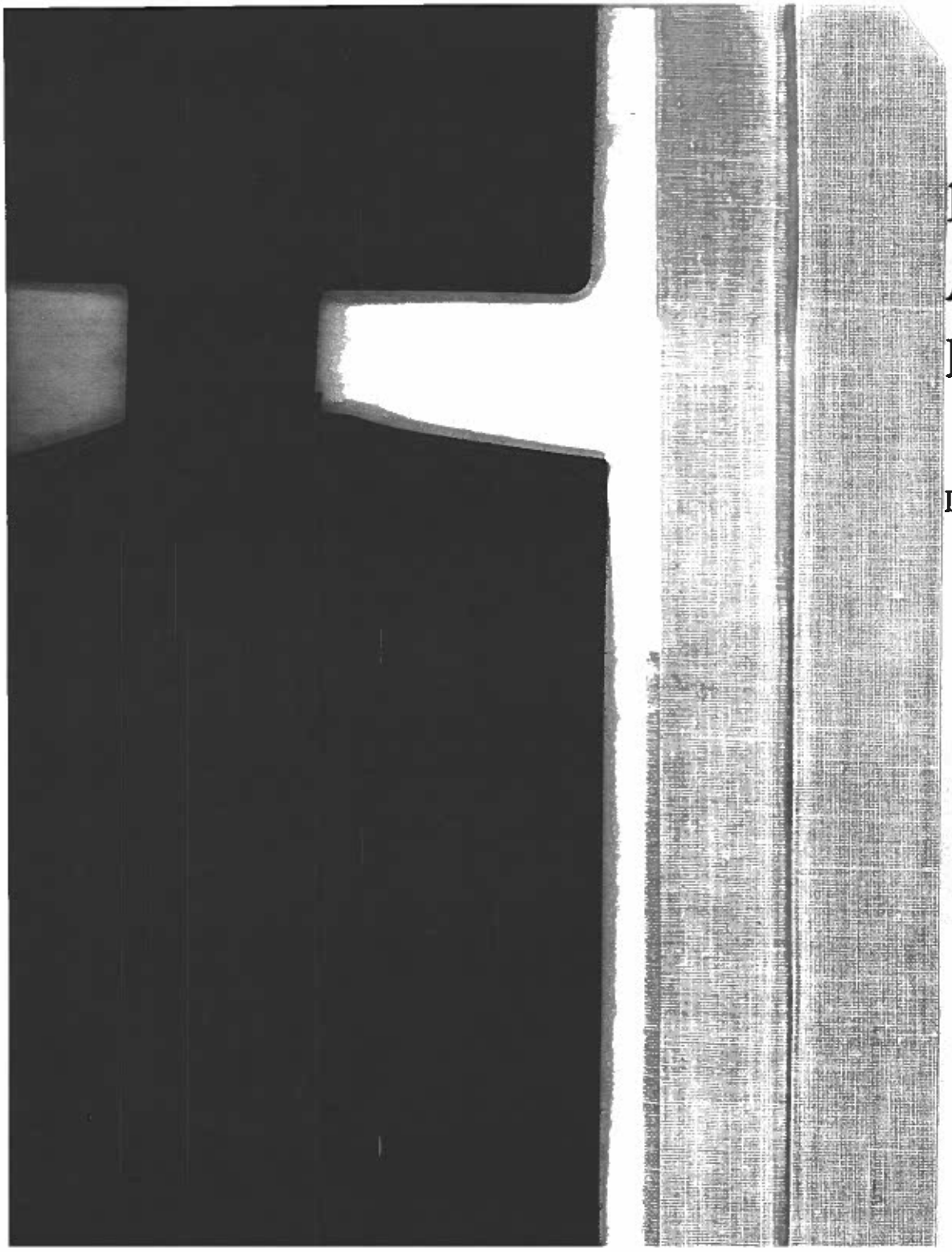


PRUDENCE, MORALITY, AND THE PRISONER'S DILEMMA

DEREK PARFIT

PRUDENCE, MORALITY, AND THE PRISONER'S DILEMMA

BY

DEREK PARFIT

ANNUAL PHILOSOPHICAL LECTURE
Henriette Hertz Trust
1978

FROM THE PROCEEDINGS OF THE
BRITISH ACADEMY, LONDON, VOLUME LXV (1979)
OXFORD UNIVERSITY PRESS

ISBN 0 85672 211 1

Printed in Great Britain
at the University Press, Oxford
by Eric Buckley
Printer to the University

PHILOSOPHICAL LECTURE

PRUDENCE, MORALITY, AND THE PRISONER'S DILEMMA

By DEREK PARFIT

Read 16 November 1978

THERE are many theories about what we have reason to do. Some of these theories are, in certain cases, directly self-defeating. What does this show?

I

Consider first *the Prisoner's Dilemma*. You and I are questioned separately about some joint crime. The outcomes would be these:

		You: confess	You: keep silent
I	confess	Each gets 10 years	I go free, you get 12 years
	keep silent	I get 12 years, you go free	Each gets 2 years

It will be better for each if he[1] confesses. This is so whatever the other does. But if both confess that will be worse for each than if both keep silent.

Let us simplify. It will be worse for each if each rather than neither does what will be better for himself. One case occurs when

Positive Condition: each could either (1) give himself some benefit or (2) give the other some greater benefit,

and

Negative Condition: neither's choice would be in other ways better or worse for either.

When the Positive Condition holds, the outcomes would be these:

		You: do (1)	You: do (2)
I	do (1)	Each gets the lesser benefit	I get both benefits, you get neither
	do (2)	I get neither benefit, you get both	Each gets the greater benefit

If we add the Negative Condition, the diagram becomes:

		You: do (1)	You: do (2)
I	do (1)	Third-best for both	Best for me, worst for you
	do (2)	Worst for me best for you	Second-best for both

Part of the Negative Condition cannot be shown in this diagram. There must be *no reciprocity*: it must be true that neither's choice would cause the other to make the same choice. It will then be better for each if he does (1) rather than (2). This is so whatever the other does. But if both do(1) that will be worse for each than if both do (2).

When could there be no reciprocity? Only when each must make a final choice before learning what the other chose. This is not common. Nor would it ensure the Negative Condition. There might, for instance, be delayed reciprocity. Either's choice might affect whether he is later benefited by the other. We can therefore seldom know that we face a Two-Person Prisoner's Dilemma.

We can often know that we face a Many-Person version. One can be called *the Samaritan's Dilemma*. Each of us could sometimes help a stranger at some lesser cost to himself. Each could about as often be similarly helped. In small communities, the cost of helping might be indirectly met. If I help, this may cause me to be later helped in return. But in large communities this is unlikely. It may here be better for each if he never helps. But it would be worse for each if no one ever helps. Each might gain from never helping, but he would lose, and lose more, from

Another case occurs when

> *Positive Condition*: each of us could, at some cost to himself, give to the others a greater total sum of benefits,[1]

and

> *Negative Condition*: there would be no indirect effects cancelling out these direct effects.

The Positive Condition often holds. If we are numerous, so does the Negative Condition. What each does would here be unlikely to affect what the others do.

The commonest examples are *Contributor's Dilemmas*. These involve *public goods*: outcomes which benefit even those who do not help to produce them. It can be true of each person that, if he helps, he will add to the sum of benefits. But his share of what he adds may be very small. It may not repay his contribution. It may thus be better for each if he does not help. This can be so whatever others do. But it would be worse for each if fewer others help. And if none help that would be worse for each than if all do.

Some public goods need financial contributions. This is true of roads, the police, or national defence. Others need co-operative efforts. When in large firms wages depend on profits, it can be better for each if others work harder, worse for each if he does. The same can be true for peasants on collective farms. A third kind of public good is the avoidance of an evil. This often needs self-restraint. Such cases may involve

> *Commuters*: Each goes faster if he drives, but if all drive each goes slower than if all take buses;
>
> *Soldiers*: Each will be safer if he turns and runs, but if all do more will be killed than if none do;
>
> *Fishermen*: When the sea is overfished, it can be better for each if he tries to catch more, worse for each if all do;
>
> *Peasants*: When the land is overcrowded, it can be better for each if he has more children, worse for each if all do.

There are many other cases. It can be better for each if he adds to pollution, uses more energy, jumps queues, and breaks agreements; but if all do these things that can be worse for each than if none do. It is very often true that, if each rather than none

[1] Or *expected* benefits (possible benefits multiplied by the chances that his act will produce them). In many of my later claims, 'benefit' could mean

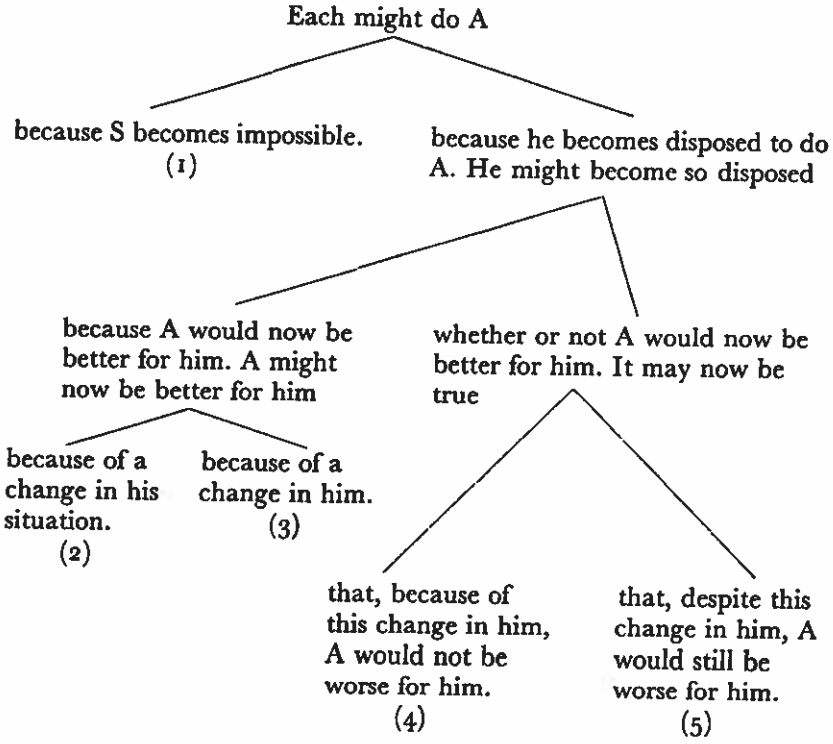
does what will be better for himself, that will be worse for everyone.

II

Each may be disposed to do what will be better for himself. There is then a practical problem. Unless something changes, the actual outcome will be worse for everyone.

Let us use labels. Each has two alternatives: *S* (self-benefiting), *A* (altruistic). If all do S that will be worse for each than if all do A. But, whatever others do, it will be better for each if he does S. The problem is that, for this reason, each is now disposed to do S.

The problem will be partly solved if most do A, wholly solved if all do. A solution may be reached in one or more of these ways:



(1) to (4) abolish the Dilemma. The altruistic choice ceases to be worse for each. These are often good solutions. But they are sometimes inefficient, or unattainable. We then need (5). This solves the practical problem. But it does not abolish the Dilemma. A theoretical prob-

In solution (1), the self-benefiting choice is made impossible. This is sometimes best. In many Contributor's Dilemmas, there should be inescapable taxation. But (1) would often be a poor solution. Fishing nets could be destroyed, soldiers chained to their posts. Both have disadvantages.

(2) is a less direct solution. S remains possible, but A is made better for each. There might be a system of rewards. But if this works all must be rewarded. It may be better if the sole reward is to escape some penalty. If this works, no one pays. If all deserters would be shot, there may be no deserters.

(1) and (2) are political solutions. What is changed is our situation. (3) to (5) are psychological. It is we who change. This change may be specific, solving only one Dilemma. The fishermen might grow lazy, the soldiers might come to prefer death to dishonour. Here are four changes of a more general kind:

> We might become *trustworthy*. Each might then agree to do A on condition that the others join in this agreement.
>
> We might become *reluctant to be 'free-riders'*. If each believes that many others will do A, he may then prefer to do his share.
>
> We might become *Kantians*. Each would then do only what he could rationally will everyone to do. None could rationally will that all do S. Each would therefore do A.
>
> We might become *more altruistic*. Given sufficient altruism, each would do A.

These are moral solutions. Because they might solve any Dilemma, they are the most important psychological solutions.

They are often better than the political solutions. This is in part because they do not need to be enforced. Take the Samaritan's Dilemma. It cannot be made impossible not to help strangers. Bad Samaritans cannot be easily caught and fined. Good Samaritans could be rewarded. But for this to be ensured the law might have to intervene. Given the administrative costs, this solution may not be worth while. It would be much better if we became directly disposed to help strangers.

It is not enough to know which solution would be best. Any solution must be introduced. This is often easier with the political solutions. Situations can be changed more easily than people. But we often face another Contributor's Dilemma. Few political solutions can be introduced by a single person. Most

good, benefiting each whether or not he does his share in bringing it about. In most large groups, it will not be better for each if he does his share. His own contribution will not make enough difference.

This problem may be small in well-organized democracies. It may be sufficient here to get the original problem widely understood. This may be difficult. But we may then vote for a political solution. With a responsive government, there may even be no need to hold a vote.

The problem is greater when there is no government. This is what worried Hobbes. One example is the spread of nuclear weapons. Without world government, it may be hard to achieve a solution.

The problem is greatest when its solution is opposed by some ruling group. This is *the Dilemma of the Oppressed.*

Such Contributor's Dilemmas often need moral solutions. We often need some people who are directly disposed to do their share. If these can change the situation, so as to achieve a political solution, this may be self-sustaining. But without such people it may never be achieved.

The moral solutions are, then, often best; and they are often the only attainable solutions. We therefore need the moral motives. How could these be introduced? Fortunately, that is not our problem. They exist. That is how we solve many Prisoner's Dilemmas. Our need is to make these motives stronger, and more widely spread.

With this task, theory helps. Prisoner's Dilemmas need to be explained. So do their moral solutions. Both have been too little understood.

One solution is, we saw, a conditional agreement. For this to be possible, it must first be true that we can all communicate. If we are self-interested, this would seldom make a difference. In most large groups, it would be pointless to agree that we will make the altruistic choice, since it would be better for each if he breaks this agreement. But suppose that we are trustworthy. Each could now promise to do A, on condition that everyone else makes the same promise. If we know that we are all trustworthy, each will have a motive to join this conditional agreement. Each will know that, unless he joins, the agreement will not take effect. Once we have all made this promise, we will all do A.

In cases that involve only a few people, such a joint condi-

involve large numbers it has little use. It will take some effort both to enable all to communicate and then to reach a joint agreement. But the agreement is a public good, benefiting each whether or not he helps to produce it. In most large groups, it will not be better for each if he helps. To this Contributor's Dilemma, trustworthiness provides no solution.

If we are reluctant to be free-riders, this problem is reduced. There is now no need for an actual agreement. All that is needed is an assurance that there will be many who do A. Each would then prefer to do his share. But a reluctance to free-ride cannot by itself create this assurance. So there are many cases where it provides no solution.

The Kantian Test could always provide a solution. This Test has its own problems. Could I rationally will either that none practise medicine, or that all do? If we refine the Test, we may solve such problems. But in Prisoner's Dilemmas they do not arise. These are the cases where we naturally say, 'What if everyone did that?'

The fourth solution is sufficient altruism. This has been the least understood. Each altruistic choice benefits others. But in Contributor's Dilemmas the benefit to each of the others may be very small. It may even not be perceptible. Some believe that such benefits make no moral difference. If that were so, rational altruists would not contribute.

It cannot be so. Consider *the Donor's Paradox.* Many wounded men lie out in the desert. Each of us has one pint of water, which he could carry to some wounded man. But if our pints are carried separately, much of the water would evaporate. If instead we pour our pints into a water-cart, there would be no evaporation. For rational altruists, this would be a better way of giving. Each wounded man would receive more water. But the pint that each of us contributes would now be shared between all these many men. It would give to each man only a single drop. Even to a wounded man, each drop of water is a very tiny benefit. If we ignore such benefits, we shall be forced to conclude that each of our contributions is now wasted.[1]

Let us next subdivide the moral solutions. When some moral motive leads someone to do A, what he does may either be, or not be, worse for him. This distinction raises deep questions.

[1] I follow J. Glover, 'It Makes No Difference Whether Or Not I Do It', *Proceedings of the Aristotelian Society, Suppl. Vol. 49* (1975). A similar argument could show that, when our acts may benefit or harm large numbers of

But I shall simply state what my arguments assume. What is in our interests partly depends on what our motives are. If we have moral motives, it may therefore not be true that doing A is worse for us. But this might be true. Even if we know it is, we might still do A.

I am here dismissing four claims. Some say that no one does what he believes to be worse for him. This has been often refuted. Others say that what each does is, by definition, best for him. In the economist's phrase, it will 'maximize his utility'. Since this is merely a definition, it cannot be false. But it is here irrelevant. It is simply not about what is in a person's long-term self-interest. Others say that virtue is always rewarded. Unless there is an after-life, this has also been refuted. Others say that virtue is its own reward. This is too obscure to be easily dismissed—or discussed here.

To return to my own claims. Many Prisoner's Dilemmas need moral solutions. We must become directly disposed to make the altruistic choice. These solutions are of two kinds. Some abolish the Dilemma. In such cases, because of this change in us, it is no longer true that it will be worse for each if he does A. But in other cases this is still true. Even in such cases, we might do A. Each might do, for moral reasons, what he knows to be worse for him.

We often need moral solutions of this second kind. Call them *self-denying*. They solve the practical problem. The outcome is better for everyone. But they do not abolish the Dilemma. A theoretical problem remains.

III

It is this. We may have moral reasons to do A. But it will be better for each if he does S. Morality conflicts with self-interest. When these conflict, what is it rational to do?

On one view, it is the self-benefiting choice which is rational. This view lacks a good name. Call it *prudence*. If we accept this view, we will be ambivalent about self-denying moral solutions. We will believe that, to achieve such solutions, we must all act irrationally.

Many writers resist this conclusion. Some claim that moral reasons are no weaker than prudential reasons. Others claim, more boldly, that they are stronger. On their view, it is the altruistic choice which is rational.

This debate may seem unresolvable. How can these two kinds of reason be weighed against [illegible]? Moral

are, of course, morally supreme. But prudential reasons are prudentially supreme. Where can we find a neutral scale?

Some believe we do not need a neutral scale. They claim that, in Prisoner's Dilemmas, prudence is *self-defeating*. Even in prudential terms, morality wins.

Is this so? Call prudence

> *individually self-defeating* when it would be worse for someone if he is prudent,

and

> *collectively self-defeating* when it would be worse for each if all rather than none are prudent.

Prudence might be individually self-defeating. Either of these might be true:

> (1) It might be worse for someone if he acted prudently. When there is uncertainty, the prudent act may not be the one which turns out best.
>
> (2) It might be worse for someone if he was disposed to act prudently. This might be worse for him even if he always did what would be best for him. One example is the 'paradox of hedonism': happiness, if aimed at, may be harder to achieve.

In Prisoner's Dilemmas, neither of these is true. The bad effects are here produced by acts, not dispositions. And there is no uncertainty. It will be better for each if he acts prudently. It is the self-benefiting choice which is prudent; and, whatever others do, it will be better for each if he makes this choice. So prudence is not here individually self-defeating. But it is collectively self-defeating. If all act prudently, that will be worse for each than if none do.

Does this show that, if we all act prudently, we are irrational? We can start with a smaller question. Do our own assumptions show us this? Is our prudence failing even in its own terms?

We might answer: 'No. The prudence of each is better for him. It succeeds. Why is our prudence here collectively self-defeating? Only because the prudence of each is worse for others. That does not make it unsuccessful. It is not benevolence.' If we are prudent, we will of course deplore Prisoner's Dilemmas. These are not the cases loved by classical economists, where each gains from universal prudence. We might say: 'In those cases, prudence both works and approves the situation. In

his own prudence. But since each loses even more from the prudence of others, prudence here condemns the situation.'

This may seem an evasion. When it is worse for each if we are all prudent, it may seem that our prudence should condemn itself. Suppose that in some other group, facing the same Dilemmas, all make the altruistic choice. They might say to us: 'You think us irrational. But we are better off than you. We do better even in prudential terms.'

We could answer: 'That is just a play on words. You "do better" only in the sense that you are better off. Each of you is *doing* worse in prudential terms. He is doing what is worse for him.' We might add: 'What is worse for each of us is that, in our group, there are no fools. Each of you has better luck. His own irrationality is worse for him, but he gains even more from the irrationality of others.'

They might answer: 'You are partly right. Each of us *is* doing worse in prudential terms. But, though *each* is doing worse, *we* are doing better. That is not a play on words. Each of us is better off because of what we *do*.'

This suggestion looks more promising. Return to the simpler Two-Person Case. Each could either benefit himself (S) or give to the other some greater benefit (A). The outcomes would be these:

		You: do S	You: do A
I: do S		Third-best for each	Best for me, worst for you
I: do A		Worst for me, best for you	Second-best for each

To ensure that neither's choice can affect the other's, suppose that we cannot communicate. If I do A rather than S, that will then be worse for me. This is so whatever you do. And the same holds for you. If we both do A rather than S, each is therefore doing worse in prudential terms. The suggestion is that *we* are doing better.

What makes this promising is that it contrasts 'each' with 'we'. In some claims, these are equivalent. It cannot be true that each is old but we are young. But in other claims they are not equivalent. It might be true that each is weak but we are strong. We *together* might be strong. Our suggestion is of this second kind. It might be true that, though each is doing worse in prudential terms, we together are doing better.

Is this true? Let us use this test. Our prudence gives to each a certain aim. Each does better, in prudential terms, if he more effectively achieves this aim. *We* do better, in the same terms, if we more effectively achieve the aim of each. This test seems fair. It might show that, if each does the best he can, we together could not do better.

What is the aim that our prudence gives to each? We might say, 'to act prudently'. This is true, but misleading. Some aims are fundamental. Others are derived from these. Call the former *goals*. When we are measuring success, only goals count. Suppose that we are trying to scratch our own backs. The goal of each might be that he cease to itch. We would then do better if we scratched each other's backs. But we might be contortionists: the goal of each might be that his back be scratched *by himself*. If we scratched each other's backs, we would then do worse.

If we are prudent, what is the goal of each? Is it that his interests be advanced, or that his interests be advanced *by himself*? If it was the second, we would not be prudent. Perhaps we are Nietzscheans, whose ideal is 'the fiercest self-reliance'. If we both do A rather than S, we would be doing worse in these terms. The interests of each would be better advanced. But neither's would be advanced by himself. Neither's goal would be achieved.

This Nietzschean ideal is not prudence. Both give each the aim of self-advancement. But only for Nietzscheans is this the goal. For the prudent, any act is a mere means. The goal is always the effect—whether this be pleasure, or some other benefit. (Nietzsche's 'blond beasts' were, it is said, lions. But, for them too, acting is a means. They prefer to eat what others kill.)

The goal of each person's prudence is the best possible outcome for himself. If we both do A rather than S, we make the outcome better for each. We cause the goal of each to be better achieved. We are therefore doing better in prudential terms. This confirms the suggestion made above. The prudent act is S. If we both act prudently, we are doing worse than we could even in prudential terms.

Does this show that our prudence here condemns itself? It may seem so. And it is tempting to contrast prudence with morality. We might say: 'Prudence breeds conflict, telling each to work against others. That is how universal prudence can be bad for all. Where prudence divides, morality unites. It tells

provided by self-interest, morality therefore wins. This is what we learn from Prisoner's Dilemmas. If we exchange prudence for morality, we do better even in prudential terms.'

This is too swift. *We* do better, but *each* does worse. If we both do A rather than S, *we* make the outcome better for each, but *each* makes the outcome worse for himself. Whatever the other does, it would be better for each if he did S. In Prisoner's Dilemmas, the problem is this. Should *each* do the best he can for himself? Or should *we* do the best we can for each? If *each* does what is best for himself, *we* do worse than we could for each. But *we* do better for each only if *each* does worse than he could for himself.

This is just a special case of a wider problem. Consider any theory about what we have reason to do. There might be cases where, if each does better in this theory's terms, we do worse, and vice versa. Call such cases *Each–We Dilemmas*.

Some theories cannot produce such Dilemmas. We shall later see why, for certain theories, this is so. If a theory does produce Each–We Dilemmas, it is not obvious what this shows. Reconsider prudence. This tells each to do the best he can for himself. We are discussing cases where, if we all act prudently, we are doing what is worse for each. Prudence is here collectively self-defeating. But it is not obvious that this is a fault. Why should a theory be collectively successful? Why is it not enough that, at the individual level, it works?

We might say: 'But a theory cannot apply only to a single individual. If it is rational for me to act prudently, it must be rational for everyone to do so. Any acceptable theory therefore must be successful at the collective level.'

This involves a confusion. Call a theory *universal* if it applies to everyone, *collective* if it claims success at the collective level. Some theories have both features. One example is a Kantian morality. This tells each to do only what he could rationally will everyone to do. The plans or policies of each must be tested at the collective level. For a Kantian, the essence of morality is the move from *each* to *we*.

At the collective level—as an answer to the question, 'How should we all act?'—prudence *would* condemn itself. Suppose that we are choosing what code of conduct will be publicly encouraged, or taught in schools. It would here be prudent to vote against prudence. If we are choosing a collective code, the prudent choice would be morality.

it is not a collective code. It is a theory of individual rationality. This answers the smaller question that we asked above. In Prisoner's Dilemmas, where it is only collectively self-defeating, prudence does not condemn itself.

IV

Many bad theories do not condemn themselves. So the larger question remains open. In such cases, what it is rational to do?

It may help to introduce another common theory. This tells each to do what will best achieve his present aims. Call this *the instrumental theory*. Suppose that, in some Prisoner's Dilemma, my aim is the outcome which is best for me. On the instrumental theory, it is then the prudent choice which is rational. If my aim is to benefit others, or to apply the Kantian Test, it is the altruistic choice which is rational. If my aim is to do what others do—perhaps because I do not wish to be a free-rider—it is uncertain which choice is rational. This depends on my beliefs about what others do.

As these remarks show, the instrumental theory may conflict with prudence. What will best achieve my present aims may be against my own long-term self-interest. Since the two theories may conflict, those who believe in prudence must reject the instrumental theory.

They might point out that, even at the individual level, it can be self-defeating. It can produce intertemporal Dilemmas. These will be most common if I care less about my further future. Suppose that, at different times, I have conflicting aims. At each time I could either (1) do what will best achieve my present aims or (2) do what will best achieve, or enable me to achieve, all of my aims over time. On the instrumental theory, I should always do (1) rather than (2). Only so will I at each time do the best I can in instrumental terms. But over time I may then do worse, in these same terms. Over time, I may be less successful in achieving my aims at each time. (Here is a trivial example. At each time I will best achieve my present aims if I then waste no energy on being tidy. But if I am never tidy this may cause me at each later time to achieve less.)

Those who believe in prudence may appeal to such cases. They might say: 'The instrumental theory is here self-defeating. Even in this theory's terms, prudence is superior. The prudent act is (2). If you always do (2) rather than (1), you will more effectively achieve your aims at each time. If you are prudent,

This is again too swift. I do better *over time*. But *at each time* I do worse. If I always do (2), I am at each time doing what will less effectively achieve the aims that I then have. (1) is what will best achieve these. Remember the interpersonal Dilemma. For the word 'we' substitute 'I over time', and for the word 'each' substitute 'I at each time'. In the interpersonal Dilemma, we do better only if each does worse than he could. In the intertemporal Dilemma, I do better over time only if at each time I do worse than I then could.

We must again distinguish two levels. The instrumental theory is here *intertemporally* self-defeating. But it does not claim to be successful at the intertemporal level. So it does not condemn itself. It is not a failure in its own terms.

Those who believe in prudence must claim that, none the less, it should be rejected. They might say: 'Any acceptable theory must be intertemporally successful. It is no defence that the instrumental theory does not claim such success. That merely shows it to be structurally flawed. If a theory is intertemporally self-defeating, this is enough to show that it should be rejected.'

This is a dangerous argument. If it refutes the instrumental theory that it is intertemporally self-defeating, why does it not refute prudence that it is collectively self-defeating? And if it is a good reply that prudence does not claim to be collectively successful, why can the instrumental theorist not make a similar reply?

As this shows, prudence can be challenged from two directions. This makes it harder to defend. Answers to either challenge may undermine answers to the other.

One challenge comes from moral theories. The other challenge need not come from the instrumental theory. It can come from theories which are more plausible. The instrumental theory has two features. It is *time-relative*: appealing to the agent's aims at the time of acting. And it is *purely instrumental*: it discusses only means, taking the agent's aims as given. According to this theory, no aim is irrational. Any aim can provide reasons for acting.

Other theories are time-relative, but not purely instrumental. One example is *the deliberative theory*. This appeals, not to the agent's actual aims at the time of acting, but to the aims he would then have, if he knew the facts and was thinking clearly. According to this theory, if an aim would not survive such deliberation, it does not provide good reasons for

A deliberative theorist may add further claims. He may say that, even if they would survive this test, certain kinds of aim are intrinsically irrational.

Since it is time-relative, the deliberative theory may conflict with prudence. Someone may be thinking clearly, yet have aims which he knows to be against his own long-term self-interest. And we may deny that all such aims are thereby shown to be irrational. We may believe that there are many aims which are not less rational than the pursuit of self-interest. Some examples might be: benefiting others, discovering truths, or creating beauty. On a time-relative theory, what it is rational for me to do now depends on which among these many aims are the ones that I have now.

Those who believe in prudence must reject such theories. They must claim that reasons for acting cannot be time-relative. They might say: 'The force of a reason extends over time. Since I *will* have reason to promote my future aims, I have reason to do so *now*.' This claim is at the heart of prudence.

Many moral theorists make a second claim. They believe that certain reasons are not agent-relative. They might say: 'The force of a reason may extend, not only over time, but over different lives. Thus, if *you* have reason to relieve your pain, this is a reason for me too. *I* have a reason to relieve *your* pain.'

Prudence makes the first claim, but rejects the second. It may be hard to defend both halves of this position. In reply to the moralist, the prudent man may ask, 'Why should *I* give weight to aims which are not *mine*?' But he can then be asked, 'Why should I give weight *now* to aims which are not mine *now*?' He may answer by appealing to the intertemporal Dilemmas, where time-relative theories are intertemporally self-defeating. But he can then be challenged with the interpersonal Dilemmas, where his own theory is collectively self-defeating. The moralist might say: 'The argument for prudence carries us beyond prudence. Properly understood, it is an argument for morality.'

This is a tempting line of thought. But something else should be discussed first. At the interpersonal level, the contrast is *not* between prudence and morality.

V

It will help to draw some more distinctions. We have been considering different theories about rationality. We can describe

According to all these theories, we should try to act rationally. Call this our *formal* aim. We can ignore this here. By 'aims' we can mean *substantive* aims. We can describe moral theories in the same way. According to all these theories, we should try to act morally. Different moral theories give us different substantive aims.

We can next distinguish two ways in which a theory might be substantively self-defeating. Call this theory *T*, and the aims it gives us *our T-given aims*. Say that we *successfully follow T* when each succeeds in doing what, of the acts available, best achieves his T-given aims. Call T

> *indirectly self-defeating* when we will best achieve our T-given aims only if we do not try to do so,

and

> *directly self-defeating* when we will best achieve our T-given aims only if we do not successfully follow T.

Consider first a moral theory: Act Consequentialism, or *AC*. This gives to all one common aim: the best possible outcome. If we try to achieve this aim, we may often fail. Even when we succeed, the fact that we are disposed to try might make the outcome worse. AC might thus be indirectly self-defeating. What does this show? A consequentialist might say: 'It shows that AC should be only one part of our moral theory. It should be the part that covers successful acts. When we are certain to succeed, we should aim for the best possible outcome. Our wider theory should be this: we should have the aims and dispositions having which would make the outcome best. This wider theory would not be self-defeating. So the objection has been met.'

Could AC be *directly* self-defeating? Could it be true that we will make the outcome best only if we do not successfully follow AC? This is not possible. We successfully follow AC when each does what, of the acts available, makes the outcome best. This does not ensure that our acts jointly produce the best possible outcome. But, if they do, we must be successfully following AC. So AC cannot be directly self-defeating.

We can widen this conclusion. When any theory *T* gives to all agents *common* aims, it cannot be directly self-defeating. If we cause these common aims to be best achieved, we must be successfully following T. So it cannot be true that we will best achieve our T-given aims only if we do not successfully follow T.

What if T gives to *different* agents *different* aims? There may then be no way in which we can *best* achieve [illegible]

of each. So we must change our definition. And we need our earlier distinction. Call T

> *directly individually self-defeating* when it is certain that, if someone successfully follows T, he will thereby cause his T-given aims to be worse achieved,

and

> *directly collectively self-defeating* when it is certain that, if all rather than none successfully follow T, we will thereby cause the T-given aims of each to be worse achieved.

Suppose that T gives to you and me different aims. And suppose that each could either (1) promote his own T-given aim or (2) more effectively promote the other's. The outcomes would be these:

		You: do (1)	You: do (2)
I	do (1)	The T-given aim of each is third-best achieved	Mine is best achieved, yours worst
	do (2)	Mine is worst achieved, yours best	The T-given aim of each is second-best achieved

Suppose finally that neither's choice will affect the other's. It will then be true of each that, if he does (1) rather than (2), he will thereby cause his T-given aim to be better achieved. This is so whatever the other does. So we both successfully follow T only if we both do (1) rather than (2). Only then is each doing what, of the acts available, best achieves his T-given aim. But it is certain that if both rather than neither successfully follow T—if both do (1) rather than (2)—we will thereby cause the T-given aim of each to be worse achieved. Theory T is here directly collectively self-defeating.

If for 'T' we substitute 'prudence', we have just described a Prisoner's Dilemma. As this shows, nothing depends on the content of prudence. Such cases may occur when

> (*a*) theory T is *agent-relative*, giving to different agents different aims,
>
> (*b*) the achievement of each person's aim partly depends on what others do,

and

These conditions may hold if for 'T' we substitute 'common-sense morality'.

VI

Most of us believe that there are certain people to whom we have special obligations. These are the people to whom we stand in certain relations—such as our children, parents, pupils, patients, members of our own trade union, or those whom we represent. We believe we ought to help these people in certain ways. We should try to protect them from certain kinds of harm, and should try to give them certain kinds of benefit. Common-sense morality largely consists in such obligations.

Carrying out these obligations has priority over helping strangers. This priority is not absolute. We may not believe that I ought to save my child from some minor harm rather than saving a stranger's life. But I ought to protect my child rather than saving strangers from *somewhat* greater harms. My duty to my child is not overridden whenever I could do somewhat greater good elsewhere.

When I try to protect my child, what should my aim be? Should it simply be that he is not harmed? Or should it rather be that he is saved from harm by me? If you would have a better chance of saving him from harm, I would be wrong to insist that the attempt be made by me. This suggests that my aim should take the simpler form. Let us assume that this is so.

Consider *the Parent's Dilemma*. We cannot communicate. But each could either (1) save his own child from some harm or (2) save the other's child from another somewhat greater harm. The outcomes would be these:

		You: do (1)	You: do (2)
I	do (1)	Both our children suffer the greater harm	Mine suffers neither harm, yours both
	do (2)	Mine suffers both, yours neither	Both suffer the lesser harm

Since we cannot communicate, neither's choice will affect the other's. If the aim of each should be that his child not be harmed, each should here do (1) rather than (2). Each would

the other does. But if both do (1) rather than (2) both our children will be harmed more.

Consider next those benefits which I ought to try to give my child. What should my aim here be? Should I insist that it be *I* who benefits my child, if I knew that this would be worse for him? Some would answer, 'No'. But this answer may be too sweeping. It treats parental care as a mere means. We may think it more than that. We may agree that, with some kinds of benefit, my aim should take the simpler form. It should simply be that the outcome be better for my child. But there may be other kinds of benefit, which my child should receive *from me*.

With both kinds of benefit, we can face Parent's Dilemmas. Consider *Case Two*. We cannot communicate. But each could either (1) benefit his own child or (2) benefit the other's child somewhat more. The outcomes would be these:

		You: do (1)	You: do (2)
I	do (1)	Third-best for both our children	Best for mine, worst for yours
	do (2)	Worst for mine, best for yours	Second-best for both

If my aim should here be that the outcome be better for my child, I should again do (1) rather than (2). And the same holds for you. But if both do (1) rather than (2) that will be worse for both our children. Compare *Case Three*. We cannot communicate. But I could either (1) enable myself to give my child some benefit or (2) enable you to benefit yours somewhat more. You have the same alternatives with respect to me. The outcomes would be these:

		You: do (1)	You: do (2)
I	do (1)	Each can give his child some benefit	I can benefit mine most, you can benefit yours least
	do (2)	I can benefit mine least, you can benefit yours most	Each can benefit his child more

gain do (1) rather than (2). And the same holds for you. But f both do (1) rather than (2) each can benefit his child less. Note he difference between these two examples. In Case Two we are oncerned with what happens. The aim of each is that the outome be better for his child. This is an aim that the other can irectly cause to be achieved. In Case Three we are concerned ith what we *do*. Since my aim is that *I* benefit my child, you annot, on my behalf, do so. But you might enable me to do so. 'ou might thus indirectly help my aim to be achieved.

Two-Person Parent's Dilemmas are unlikely to occur. But e often face many-person versions. It is often true that, if all ather than none give priority to our own children, that will ither be worse for all our children, or will enable each to enefit his children less. Thus there are many outcomes which ould benefit our children whether or not we help to produce em. It can be true of each parent that, if he does not help, at will be better for his own children. He can spend what he ves—whether in money, time, or energy—directly on them. ut if none help, that will be worse for all our children than if l do. In another common case, each could either (1) add to is own earnings or (2) (by self-restraint) add more to the earnigs of others. It will here be true of each that, if he does) rather than (2), he can benefit his children more. This is so hatever others do. But if all do (1) rather than (2) each can enefit his children less. These are only two of the ways in hich such cases can occur. There are many others.

Similar remarks apply to all similar obligations—such as ose to pupils, patients, clients, or constituents. With all such bligations, there are countless many-person versions of my ree examples. They are as common, and as varied, as prudntial Each–We Dilemmas. As we have just seen, they will ten have the same cause. Here is another way in which this ight be true. Suppose that, in the original case, it is our wyers who must choose. This is the *Prisoner's Lawyer's Dilemma*. both lawyers give priority to their own clients, that will be orse for both clients than if neither does. Any prudential ilemma may thus yield a moral Dilemma. If one group face e former, another may in consequence face the latter. This n be so if we believe that each member of the second group ght to give priority to some members of the first. The problem mes from the giving of priority. It makes no difference hether this is given to oneself or certain others.

can arise for other parts of common-sense morality. It can arise whenever this morality gives to different people different duties. Suppose that each could either (1) carry out some of his own duties or (2) enable others to carry out more of theirs. If all rather than none give priority to our own duties, each may be able to carry out fewer. Deontologists can face Each–We Dilemmas. But I shall not discuss these here.

VII

What do such cases show? Common-sense morality is the moral theory most of us accept. According to this theory, there are certain things that each of us ought to try to achieve. These are what I call our 'moral aims'. We successfully follow this moral theory when each does what, of the acts available, best achieves his moral aims. In my cases it is certain that, if all rather than none successfully follow this theory, we will thereby cause the moral aims of each to be worse achieved. Our moral theory is here directly collectively self-defeating. Is this an objection?

Let us start with a smaller question. Could we revise our theory, so that it would not be self-defeating? If there is no such revision, ours may be the best possible theory. Since we believe our theory, we should ask what is the smallest such revision. So we should first identify the part of our theory which is self-defeating.

It will help to bring together two distinctions. One part of a moral theory may cover *successful acts*, on the assumption of *full compliance*. Call this part *ideal act theory*. This says what we should all try to do, simply on the assumptions that we all try, and all succeed. Call this *what we should all ideally do*.

Note next that, in my examples, what is true is this. If *all* of us *successfully* follow our moral theory, it will be self-defeating. It is our ideal act theory which is self-defeating. If we ought to revise our theory, this is the part that must certainly be revised.

The revision would be this. Call our theory *M*. In such cases we should all ideally do what will cause the M-given aims of each to be better achieved. Thus in my Parent's Dilemmas we should all ideally do (2) rather than (1). That will make the outcome better for all our children, and will enable each to benefit his children more.

Call this revision *R*. Note first that R applies only to those cases where M is self-defeating. If we decide to adopt R, we will need to consider how such cases can be recognized. I believe that

Note next that R is restricted to our ideal act theory. It does not say what we ought to do when there are some others who do not follow R. Nor does it say what our aims should be when our attempts may fail. Nor does it say what dispositions we should have. Since these are the questions with most practical importance, it may seem that adopting R would make little difference. But this is not likely. If we revise this part of our theory, we shall probably revise the rest. Take the case of a public good which would benefit our children. One such good is the conservation of a scarce resource. Suppose that we are fishermen, trying to feed our children. We are faced with declining stocks. It is true of each that, if he does not restrict his catch, that will be better for his own children. This is so whatever others do. But if none restrict their catches that will be worse for all our children than if all do. According to R, we should all ideally restrict our catches. If some fail to do so, R ceases to apply. But it would be natural to make this further claim: each should restrict his catch provided that enough others do so too. We would need to decide what counts as enough. But, whatever we decide, adopting R would have made a difference. Failure to restrict our catches would now be at most a defensive second-best. Consider next the relation between acts and dispositions. Suppose that each could either (1) save his own child from some lesser harm or (2) save another's child from some greater harm. According to R, we should all ideally do (2). Should we be *disposed* to do (2)? If the lesser harms would themselves be great, such a disposition might be incompatible with love for our own children. This may lead us to decide that we should remain disposed to do (1). This would mean that, in such cases, our children would be harmed more; but, if we are to love them, this is the price they must pay. Such remarks cannot be made whenever M is self-defeating. It would be possible to love one's children and contribute to most public goods. Nor could such remarks cover all similar obligations—such as those to pupils, patients, clients, or constituents. It is therefore likely that, if we adopt R, we will be led to change our view about some dispositions.

We can now return to the main question. Ought we to adopt R? Is it an objection to our moral theory that, in certain cases, it is self-defeating? If it is, R is the obvious remedy. R revises M only where M is self-defeating. And the only difference is that R is not.

The problem is not that, in our attempts to follow M, we are somehow failing. That might be no objection. The problem is that we all *successfully* follow M. Each succeeds in doing what, of the acts available, best achieves his M-given aims. This is what makes M self-defeating. And this does seem an objection. If there is any assumption on which a moral theory should *not* be self-defeating, it is surely the assumption that it is universally successful followed.

Remember next that by 'aims' I mean substantive aims. I have ignored our formal aim: the avoidance of wrongdoing. This may seem to remove the objection. Take those cases where, if we follow M, either the outcome will be worse for all our children, or each can benefit his children less. We might say: 'These results are, of course, unfortunate. But how could we avoid them? Only by failing to give priority to our own children. That would be wrong. So these cases cast no doubt on our moral theory. Even to achieve our other moral aims, we should never act wrongly.'

These remarks are confused. It is true that, in these cases, M is not formally self-defeating. If we follow M, we are not doing what we believe to be wrong. On the contrary we think it wrong *not* to follow M. But M is substantively self-defeating. Unless we all do what we now think wrong, we will cause our M-given aims to be worse achieved. The question is: Might this show that we are mistaken? Ought we perhaps to do what we *now think* wrong? We cannot answer, 'No—we should never act wrongly.' If we are mistaken, we would *not* be acting wrongly. Nor can we simply say, 'But, even in these cases, we *ought* to give priority to our own children.' This just assumes that we are not mistaken. To defend our theory, we must claim more than this. We must claim that it is no objection to our theory that, in such cases, it is substantively self-defeating.

This would be no objection if it simply did not matter whether our M-given aims will be achieved. But this does matter. The sense in which it matters may be unclear. If we have not acted wrongly, it may not matter morally. But it matters in a way which has moral implications. Why should we try to achieve our M-given aims? Part of the reason is that, in this other sense, their achievement matters.

Someone might say: 'You call M *self*-defeating. So your objection must appeal *to M*. You should not appeal to some rival theory. This is what you have now done. When you claim

are merely claiming that, if they are not, the outcome would be worse. This assumes consequentialism. So you beg the question.'

This is not so. When our aims are held in common, call them *agent-neutral*. Other aims are agent-relative. Any aim may be concerned either with what happens or with what is done. So there are four kinds of aim. Here are some examples:

	Concerned with what happens	Concerned with what is done
agent-neutral	that children do not starve	that children are cared for by their own parents
agent-relative	that my children do not starve	that I care for my children

When I claim that it matters whether our M-given aims will be achieved, I am not assuming consequentialism. Some of these aims are concerned with what we *do*. Thus parental care may not be for us a mere means. More important, I am not assuming agent-neutralism. Since our moral theory is, for the most part, agent-relative, this would beg the question. But it need not be begged.

There are here two points. First, I am not assuming that what matters is the achievement *of M-given aims*. Suppose that I could either (1) promote my own M-given aims or (2) more effectively promote yours. According to M, I should here do (1) rather than (2). I would thereby cause M-given aims to be, on the whole, worse achieved. But this does not make M self-defeating. I would cause *my* M-given aims to be *better* achieved. In my examples the point is not that, if we all do (1) rather than (2), we cause M-given aims to be worse achieved. The point is that we cause *each of our own* M-given aims to be worse achieved. We do worse not just in agent-neutral but in agent-relative terms.

The second point is that this can matter in an agent-relative way. It will help to remember prudence, or *P*. In Prisoner's Dilemmas, P is directly self-defeating. If all rather than none successfully follow P, we will thereby cause the P-given aim of each to be worse achieved. We will make the outcome worse for everyone. If we believe in prudence, will we think this matters? Or does it only matter whether each achieves his formal aim: the avoidance of irrationality? The answer is clear. According to prudence, acting rationally is a matter of all

that matters is the achievement of our substantive P-given aims. What concerns us here is this. The achievement of these aims matters in an agent-relative way. To think it an objection that our prudence is self-defeating, we need not appeal to its agent-neutral form: Utilitarianism. Prudence is not a moral theory. But the comparison shows that, in discussing common-sense morality, we need not beg the question. If it matters whether our M-given aims will be achieved, this, too, can matter in an agent-relative way.

Does this matter? Note that I am not asking whether this is all that matters. I am not suggesting that the achievement of our formal aim—the avoidance of wrongdoing—is a mere means. Though assumed by consequentialists, this is not what most of us believe. We may even think that the achievement of our formal aim always matters most. But this is here irrelevant. We are asking whether it casts doubt on M that it is substantively self-defeating. Might this show that, in such cases, M is incorrect? It may be true that what matters most is that we avoid wrongdoing. But this truth cannot show M to be correct. It cannot help us to decide what *is* wrong.

Can we claim that our formal aim is all that matters? If that were so, my examples would show nothing. We could say, 'To be substantively self-defeating is, in the case of common-sense morality, *not* to be self-defeating.' Can we defend our moral theory in this way? In the case of some M-given aims, perhaps we can. Consider trivial promises. We might believe both that we should try to keep such promises, and that it would not matter if, through no fault of ours, we fail. But we do not have such beliefs about all of our M-given aims. If our children suffer harm, or we can benefit them less, this matters.

Remember finally that, in my examples, M is collectively *but not individually* self-defeating. Could this provide a defence?

This is the central question I have raised. It is because M is individually successful that, at the collective level, it is here *directly* self-defeating. Why is it true that, if we all do (1) rather than (2), we *successfully* follow M? Because *each* is doing what, of the acts available, *best* achieves his M-given aims. Is it perhaps no objection that *we* thereby cause the M-given aims of each to be *worse* achieved?

It will again help to remember prudence. In Prisoner's Dilemmas, prudence is collectively self-defeating. If we were choosing a collective code, something that we will all follow,

to vote against prudence. But those who believe in prudence may think this irrelevant. They can say: 'Prudence does not claim to be a collective code. To be collectively self-defeating is, in the case of prudence, *not* to be self-defeating.'

Can we defend our moral theory in this way? This depends on our view about the nature of morality. On most views, the answer is 'No'. But I must here leave this question open.[1]

[1] Many other questions need to be discussed. How, for instance, is revision R related to agent-neutralism? I hope to say more in a book on self-defeating theories (to be written for the OUP). In preparing this Lecture I have been greatly helped by R. M. Adams, R. M. Dworkin, J. L. Mackie, D. Regan, and J. J. Thomson; also by B. Barry, S. Blackburn, D. Braybrooke, P. Bricker, L. J. Cohen, N. E. Davis, D. Dennett, M. G. J. Evans, P. Foot, J. P. Griffin, G. Harman, M. Hollis, S. Kagan, R. Lindley, P. Maddy, T. Nagel, R. Nozick, C. Peacocke, J. Raz, J. Sartorelli, T. Scanlon, F. Schick, A. K. Sen, J. H. Sobel, H. Steiner, and L. Temkin. My sections III and IV owe a great deal to T. Nagel, *The Possibility of Altruism* (Oxford, 1970). My section V owes much to D. Regan, *Utilitarianism and Cooperation*, (Oxford, 1980), D. Lyons, *Forms and Limits of Utilitarianism* (Oxford, 1965), and R. M. Adams, 'Motive Utilitarianism', *Journal of Philosophy*, 12 August 1976. My section II owes much to E. Ullman-Margalit, *The Emergence of Norms* (Oxford, 1977), D. Braybrooke, 'The Insoluble Problem of the Social Contract', *Dialogue*, March 1976, and F. Miller and R. Sartorius, 'Population Policy and Public Goods', *Philosophy & Public Affairs*, Winter 1979. The other publications to which I owe most are: K. Baier, 'Rationality and Morality', *Erkenntnis*, 1977; B. Barry, *Sociologists, Economists, and Democracy* (London, 1970); J. M. Buchanan, *The Demand and Supply of Public Goods* (Chicago, 1969); D. Gauthier, 'Morality and Advantage', *The Philosophical Review*, 1967, and 'Reason and Maximization', *Canadian Journal of Philosophy*, March 1975; G. Hardin, 'The Tragedy of the Commons', *Science*, 13 December 1968; R. M. Hare, 'Ethical Theory and Utilitarianism', in H. D. Lewis (ed.), *Contemporary British Philosophy* (London, 1976); M. Olson Jr., *The Logic of Collective Action* (Cambridge, Mass., 1965); A. Rapoport, *Fights, Games, and Debates* (Ann Arbor, 1960); T. Schelling, 'Hockey Helmets, Concealed Weapons, and Daylight Saving', *The Journal of Conflict Resolution*, September 1973; A. K. Sen, 'Choice, Orderings, and Morality', in S. Körner (ed.), *Practical Reason* (New Haven, 1974); J. H. Sobel, 'The Need for Coercion', in J. Pennock and H. Chapman (eds.), *Coercion* (Chicago, 1972); and J. Watkin, 'Imperfect Rationality', in R, Borger and F. Cioffi (eds.), *Explanation in the Behavioural Sciences* (Cambridge, 1970).

RECENT PHILOSOPHICAL LECTURE

HENRIETTE HERTZ TRUST

DESCRIPTIVISM, by *R. M. Hare*. 1963.

PREDICTING AND DECIDING, by *David Pears*. 1964.

IMAGINATION AND THE SELF, by *Bernard Williams*. 1966.

THE OBJECTS OF THE FIVE SENSES, by *J. O. Urmson*. 1968.

WHAT'S REALLY WRONG WITH PHENOMENALISM?, by *J. L. Mac* 1969.

MORALITY AND ART, by *Philippa Foot*. 1970.

INTENTION AND UNCERTAINTY, by *H. P. Grice*. 1971.

IN DEFENCE OF OBJECTIVITY, by *Mary B. Hesse*. 1972.

THE JUSTIFICATION OF DEDUCTION, by *M. A. E. Dummett*. 1973.

TIMES, BEGINNINGS AND CAUSES, by *G. E. M. Anscombe*. 1974.

PROBABILITY—THE ONE AND THE MANY, by *L. Jonathan Cohen*. 19

TRUTH, INVENTION, AND THE MEANING OF LIFE, by *David Wig* 1976.

Published by THE BRITISH ACADEMY
Price £2·00 net
ISBN 0 85672 211 1