**Is Common-Sense Morality Self-Defeating?**

Derek Parfit

*The Journal of Philosophy*, Vol. 76, No. 10, Seventy-sixth Annual Meeting of the American Philosophical Association, Eastern Division (Oct., 1979), 533-545.

Stable URL:
http://links.jstor.org/sici?sici=0022-362X%28197910%2976%3A10%3C533%3AICMS%3E2.0.CO%3B2-J

*The Journal of Philosophy* is currently published by Journal of Philosophy, Inc..

Marxist "instrumentalism": that is, the dominant economic class creates and imposes the non-economic conditions for and instruments of its continued economic dominance. The only dispute—as I claimed above, page 531—is whether or not to *call* these functional prerequisites, explained in the way just outlined, "socially primary." There is disagreement over the use of terms like 'primary' and 'primacy', but no substantive disagreement; that is, no disagreement about there being non-economic social conditions required for social reproduction, and no disagreement concerning the economic determination of these conditions.

What is required—and what will be sketched in the full reply—is a framework for historically oriented social theory that can do justice to the explanatory importance of politics, without relying on different "senses" or "forms" of priority. This framework has three main elements:

(1) An account of the basic interests that are pursued through social action.

(2) A theory of the alliances through which individuals and social groups pursue these interests.

(3) An account of the autonomy of politics which emerges from the pursuit of interests through the formation of alliances.

JOSHUA COHEN

Massachusetts Institute of Technology

## IS COMMON-SENSE MORALITY SELF-DEFEATING? *

WHEN is a moral theory self-defeating? I suggest the following. There are certain things we ought to try to achieve. Call these our *moral aims*. Our moral theory would be self-defeating if we believed we ought to do what will cause our moral aims to be worse achieved. Is this ever true? If so, what does it show?

* To be presented in an APA symposium on Collective Irrationality, December 29, 1979. Judith Jarvis Thomson will comment; see this JOURNAL, this issue, 545–547. This paper is a shortened version of the last part of my "Prudence, Morality, and the Prisoner's Dilemma," due to appear in the *Proceedings of the British Academy*, LXV (1979), and to be published separately by the Academy. In preparing this version I have been helped by R. M. Dworkin, D. Regan, J. L. Mackie, J. P. Griffin, T. Nagel, R. M. Adams, J. H. Sobel, P. Foot, S. Blackburn, N. Davis, J. Sartorelli, P. Bricker, M. Hollis, and C. A. B. Peacocke.

0022-362X/79/7610/0533$01.20 

I

We ought to try never to act wrongly. Call this our *formal* aim. Could a moral theory be formally self-defeating? I shall not discuss this possibility. By 'aims' I shall mean substantive aims.

There are two ways in which a theory might be substantively self-defeating. Call this theory $T$, and the aims it gives to each our *T-given aims*. Say that we *successfully obey* $T$ when each succeeds in doing what, of the acts available, best achieves his T-given aims. Call T

> *indirectly self-defeating* when it is true that, if we try to achieve our T-given aims, these aims will be worse achieved,

and

> *directly self-defeating* when it is certain that, if we successfully obey T, we will thereby cause our T-given aims to be worse achieved.

Consider first 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 trying 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 certain that, if we successfully obey AC, we will thereby make the outcome worse? There is one kind of case in which this may seem possible. These are coordination problems, where what each ought to do depends upon what others do. In such cases even if we all successfully obey AC this does not ensure that our acts jointly produce the best possible outcome. But it cannot ensure that they do not. If they do, we must be successfully obeying AC. So AC cannot be directly self-defeating. It cannot be *certain* that, if we successfully obey this moral theory, we will thereby cause the aim that it gives us to be worse achieved.[1]

We can widen this conclusion. When any theory T gives us com-

[1] I summarize Donald Regan's *Utilitarianism and Cooperation*, forthcoming from Oxford University Press.

mon aims, it cannot be directly self-defeating. If we cause these common aims to be best achieved, we must be successfully obeying T. So it cannot be certain that, if we successfully obey T, we will thereby cause our T-given aims to be worse achieved.

When T gives to different people different aims, this can be certain. But we need a new distinction. Call T

> *directly individually self-defeating* when it is certain that, if someone successfully obeys 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 obey T, we will thereby cause the T-given aims of each to be worse achieved.

It is the second that is possible. Suppose that T gives to you and me different moral 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) | Our T-given aims are third-best achieved | Mine is best achieved, yours worst |
| | do (2) | Mine is worst achieved, yours best | Our T-given aims are second-best achieved |

Suppose finally that neither's choice will affect the other's. It will then be certain that, if I do (1) rather than (2), my T-given aim will be better achieved. This is so whatever you do. And the same holds for you. So we both successfully obey 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 obey T—if both do (1) rather than (2)—we will thereby cause the T-given aims of each to be worse achieved. Theory T is here directly collectively self-defeating.

Such cases can occur whenever

(a) our moral theory gives to each a different aim,

(b) the achievement of each person's aim partly depends on what others do, and

(c) what each does will not affect what these others do.

On the moral theories most of us accept, (a) and (b) often hold. In a case involving only two people, (c) may be unlikely. It may hold only if we cannot communicate. But in cases that involve large numbers of people (c) often holds. What each does would here be unlikely to affect what others do. Partly for this reason, it is the many-person cases which have practical importance. But it will be simpler to discuss two-person versions.

II

Consider first *self-referential altruism.* Most of us believe that there are certain people to whose interests we should give extra weight. Thus each ought to give priority to his children, parents, pupils, patients, members of his own trade union, those whom he represents, or fellowcountrymen. This priority should not be absolute. It would be wrong to save my child's toy rather than a stranger's life. But I ought to save my child from harm rather than save a stranger's child from a somewhat greater harm. I have special duties to my child, which cannot be overridden simply because I could do somewhat greater good elsewhere.

When I try to save my child from harm, 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 *Case One.* We cannot communicate. But each could either (1) save his own child from some lesser 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 we believe we ought to give priority to our own children, we must believe that each should do (1) rather than (2). Each would thus ensure that, whatever the other does, his own child will be harmed less. But if both do (1) rather than (2) both our children will be harmed more.

Besides trying to protect my child, I should try to give him certain kinds of benefit. What should my aim here be? Should it simply be that he receive these benefits, or should it rather be that he receive these benefits *from me*? Could I be right to 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 it should be my aim that *I* give my child.

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). That will be better for my child, whatever you do. And the same holds for you. But if both do (1) rather than (2) that will be worse for both our children. Consider next *Case Three*. We cannot communicate. But I could either (1) enable myself to benefit my child 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 benefit his child | 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 |

If my aim should here be that I benefit my child, I should again do (1) rather than (2). I can then, whatever you do, benefit my child more. And the same holds for you. But if both do (1) rather than (2) each can benefit his child less. Note the difference between these two examples. In Case Two we are concerned with what happens. The aim of each is that the outcome be better for his child. This is an aim that the other can directly cause to be achieved. In Case Three we are concerned with what we *do*. Since my aim is that *I* benefit my child, you cannot, on my behalf, do so. But you might enable me to do so. You might thus indirectly cause my aim to be achieved.

These two cases are unlikely to occur. But we often face many-person versions. It is often true that, if all rather than none give priority to our own children, that will either be worse for all our children, or will enable each to benefit his children less. One common case involves a *public good*: an outcome that benefits our children whether or not we help to produce it. It can be true of each parent that, if he helps, his contribution adds to the total benefit. But his own children's share of what he himself adds would, in a large community, be small. Nor would his example be widely copied. It may thus be better for his children if he does not contribute. He could spend what he saves—whether in money, time, or energy—directly on them. If we ought to give priority to our own children, it may thus be true of each that he should not contribute. Each would then be doing what is better for his own children, whatever others do. But if none contribute that would be worse for all our children than if all do. Consider next those benefits which it should be the aim of each that *he* give his children.

Whether each can do so may in part depend on how much he earns. It is often true that each could either (1) add to his own earnings or (2) add more to the earnings of others. (Choice 2 typically involves some kind of self-restraint.) It will here be true of each that, if he does (1) rather than (2), he can benefit his children more. This is so whatever others do. But if all do (1) rather than (2) each can benefit his children less. These are only two of the ways in which such cases can occur. There are many others. Similar remarks apply to all similar obligations—such as those to parents, pupils, members of our own trade union, or fellowcountrymen. So there are countless many-person cases with the structure of my two examples.

Consider finally those things which we should aim *not* to do—such as infringing people's rights, or harming the innocent. Should we have the common aim that *we* do not do these things, or should each have the aim that they are not done *by him*? We should here distinguish two questions. What should each do when we all do our duty? This assumes what is called *full compliance*. What should each do when there are some others who act wrongly? This assumes *partial compliance*. These two questions may need different answers. Suppose you threaten that, unless I harm one innocent person, you will harm both him and several others. Some claim that, even if I believe your threat, I should here be concerned only with what *I* do. I should refuse to harm the innocent, even if the outcome is that you do so on a larger scale. But this is not the kind of case we are discussing. We are asking what might happen if we all obey our moral theory. So we must change the example. Suppose that, through no fault of yours, it has become true that you must harm certain innocent people. There are two ways in which this might be true. If it would prevent some catastrophe, harming these people might be your duty. Even if we deny this, we must admit that you might have no alternative. It might be true that, whatever you do, you will harm these people. In either case, the question is: Should I harm one of these people, if that would enable you not to harm the others? Or should I again be concerned only with what *I* do?

Suppose we take the second view. Consider *Case Four*. Through no fault of ours, it has become true that each must harm three innocent people. We cannot communicate. But each could now improve our moral situation. Each could either (1) enable himself to harm one fewer or (2) enable the other to harm two fewer. The

outcomes would be these:

| | | You: do (1) | You: do (2) |
|---|---|---|---|
| I | do (1) | Each must harm two innocent people | I must harm none, you three |
| | do (2) | I must harm three, you none | Each must harm one |

If the aim of each should be that *he* does not harm innocent people, each should again do (1) rather than (2). Each would thus, whatever the other does, enable himself to harm fewer. But if both do (1) rather than (2) each must harm more. This case is again unlikely to occur. But its many-person version has some practical importance.

III

We can now ask what, if anything, such cases show. We believe that each should have certain moral aims. We successfully obey our moral theory when each succeeds in doing what, of the acts available, best achieves his moral aims. In my cases it is certain that, if both rather than neither successfully obey our moral 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 of the distinctions drawn above. 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, 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* obey 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 examples we should both ideally do (2) rather than (1). That will make the outcome better for both our children, and will enable each both to benefit his child more and to harm fewer innocent people.

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 they are common. But I have no space to discuss this here.

Note next that R is restricted to ideal act theory. It does not say what we ought to do when there are some who do not obey R. Nor does it say what our aims should be when there is a serious chance that our attempts will 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 does not follow. If we revise this part of our theory, we may be led to revise the rest. Return, for instance, to the public good that would benefit our children. According to R, we should all ideally contribute. If some do not contribute, R ceases to apply. But it would be natural to make this further claim: each should still contribute 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. We would now regard noncontribution as at most a defensive second-best. Consider next the relation between acts and dispositions. In Case One each could either (1) save his own child from a lesser harm or (2) save the other's child from a greater harm. According to R, we should both ideally do (2). Should we be *disposed* to do (2)? If the lesser harm would itself be great, such a disposition might be incompatible with love for our children. This may lead us to decide that we should remain disposed to do (1). This would mean that, if the case arose, our children would be harmed more; but, if we are to love them, this is a risk they must run. These remarks cannot be plausibly extended to all other cases where M is self-defeating. It would be possible to love one's children and contribute to most public goods. Nor would the remarks apply to all similar obligations—such as those to pupils, patients, those whom we represent, or our fellowcountrymen. It therefore seems likely that, if we adopt R, we would 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.

Remember first that, in these cases, M is *directly* self-defeating. The problem is not that our attempts are failing. That might be no objection. But in my examples all of us successfully obey M. Each succeeds in doing what, of the acts available, best achieves his M-given aim. This is what makes M self-defeating. And this would seem to be an objection. If there is any assumption on which a theory should *not* be self-defeating, it is the assumption that it is universally successfully obeyed.

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. Thus consider those cases where, if we obey 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 obey M, we are not doing what we believe to be wrong. On the contrary, we think it wrong *not* to obey 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 that has moral implications. Why should we try to achieve our M-given aims? Part of the reason must be 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 that it

matters whether our M-given aims will be achieved, you are merely claiming that, if they are not, the outcome will be worse. This assumes consequentialism. So you beg the question."

This is not so. It will help to introduce two more labels. When our aims are held in common, call them *agent-neutral*; when they are different for different agents call them *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 | what is done |
| agent-neutral | that no one starve | that no one steal |
| agent-relative | that my children do not starve | that I do not steal |

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*. 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 M-given aims or (2) more effectively promote yours. According to M, I should here do (1) rather than (2). I will thereby cause M-given aims to be, on the whole, worse achieved. But this does not make M self-defeating. I will cause *my* M-given aims to be better achieved. In my examples the point is not that, if we both do (1) rather than (2), we will cause M-given aims to be worse achieved. The point is that we will cause *each of our own* M-given aims to be worse achieved. We will 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. This can be shown with a comparison. Consider the account of rationality which gives to each agent this overriding aim: that the outcome be better for himself. Call this theory *prudence*, or *P*. Suppose that P was indirectly self-defeating. Suppose that, when each tries to make the outcome better for himself, he fails. If we believe theory P, would 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 theory P, acting rationally is a mere means. All that matters is the achievement of our substantive P-given aims. But the important point is this. The achievement of these aims matters in an agent-relative way. To think it an objection that P is self-defeating, we need not appeal to the agent-neutral form of P: Utilitarianism. P is not a moral theory. But the example shows that, in discussing M, 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 *all* that matters is our formal aim? If that were so, my examples would show nothing. We could say, "To be substantively self-defeating is, in the case of M, *not* to be self-defeating." Can we defend our theory in this way? In the case of some M-given aims, perhaps we can. One example might involve 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 believe this about all of our M-given aims. If we can benefit our children less, or must harm the innocent, this matters.

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

This is the central question raised by these examples. 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 both do (1) rather than (2), we *successfully* obey M? Because *each* is doing what, of the acts available, *best* achieves his M-given aim. 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 so-called "Prisoner's Dilemmas" P is directly collectively self-defeating. If all rather than none successfully obey P, that will here be worse for everyone. The

P-given aim of each will be worse achieved. If we were choosing a collective code, something that we will all follow, P would here tell us to reject itself. It would be prudent to vote against prudence. But someone who believes in P might call this irrelevant. He might say: "P is not 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 upon our view about the nature of morality. On most views, the answer is "No." But I must here leave this question open.[2]

DEREK PARFIT

All Souls College, Oxford

## COMMON-SENSE MORALITY *

Suppose a moral theory M says "One ought always harm as few as possible." Then M is self-defeating, as Parfit's Case Four shows. What he calls "Case Four" is really only a case-schema, but cases can be constructed from it.

It is of considerable interest that cases of Case Four make trouble for M. But we already knew that there was trouble for M—since we already knew that "One ought always harm as few as possible" is false. Suppose a trolley is headed for six, and I can deflect it in

[2] It remains open in my "Prudence, Morality, and the Prisoner's Dilemma" (*op. cit.*). But I there discuss certain other questions which I have ignored here. One is the question whether, if we could all communicate, M would still be self-defeating. Would it not tell us to promise to each other that, in return for the same promise, we will all do (2) rather than (1)? The answer is in theory "Yes." If we are all trustworthy, joining this conditional agreement would be the best way for each to promote his own M-given aims. In a two-person case, this solution could often be achieved. But in many-person cases, which are those with practical importance, this is not so. [We can now redefine my proposed revision. We should all ideally do what, if we could make this joint promise, we ought to promise to do. In this redefinition, R need not explicitly refer to those cases where M is self-defeating. Only here would M tell us that we ought to make this promise. And this redefinition makes R more plausible. We believe that, if we could, we ought to promise to each other that we will all do (2). Does this not suggest that, even when we cannot make this promise, this is what we should all ideally do?]

* Abstract of a paper to be presented in an APA symposium on Collective Irrationality, December 29, 1979, commenting on a paper by Derek Parfit; see this JOURNAL, this issue, 533–545.

0022-362X/79/7610/0545$00.50 