

THIS IS A VERY ROUGH DRAFT INTENDED ONLY FOR DISTRIBUTION AT THE
COLLOQUIUM

The Structure of Joint Intention

The topics of joint intention and response dependence are not normally thought to be connected. But it is my belief that there are problems concerning the very possibility of joint intention that can only be satisfactorily resolved by providing a response dependent account of what it is. Our having a joint intention will in part be realized by our responding to it *as* a joint intention.

I shall begin by outlining the problems and then consider various attempted solutions to them. None of them turns out to be satisfactory. I shall then outline a theory of response dependent concepts and show how it is able to solve the problems and provide a satisfactory account of joint attention. I conclude by briefly considering the application of the theory of response dependence to other topics in social philosophy, including the prisoners' dilemma and the concept of law.

The paper was hurriedly written and very rough. Many important topics are barely discussed and some are not even mentioned. But I hope I have said enough to make clear how a number of issues concerning joint intention can be put on a firmer footing and how a theory of response dependence can be of help in resolving them.

§1 The Question

There are many cases from everyday life in which we decide or intend to do something. We may decide to clean the house or to go for a walk or sign a contract. Or, on a larger scale, we may implicitly agree to adopt certain conventions, or abide by certain laws, or live together in a society. But what is involved in our having 'joint' intentions of this sort?

In asking this question, one might be after a set of necessary and sufficient conditions for joint intention or a significant set of necessary or of sufficient conditions. But I am after something rather different. I am interested in certain *desiderata* on joint intention - requirements that are desirable in some way and plausibly taken to be possible. I then want to show that there are real difficulties in seeing how these desiderata might be met and that the only satisfactory resolution of the difficulties is by way of a response dependent account.

It is not essential to our immediate purposes that these desiderata should be individually necessary or collectively sufficient for joint intention. But there is some plausibility to the thought that they serve to single out a core concept of joint intention - one that is common to certain central instances of the concept. And if this is so, then our results will have somewhat wider import; for they will show that there are difficulties in the very conception of joint intention, not merely in the satisfaction of the desiderata, and that this conception must itself be understood in terms of some form of response dependence.

Let us take the joint intention of me and my partner to clean the house as a running example. Then the situation in which we have this intention will consist in you and I having a complex of intentions and other mental states and perhaps also in certain 'objective' facts (such as the house being dirty) which pertain neither to my state of mind nor to yours. In order to simplify the discussion, I shall make three assumptions:

Common Knowledge We have common knowledge of the situation (so each of us knows how it is, we each know that each of us knows how it is, etc. etc.)

Rationality Each of us is rational (and so will draw the appropriate inferences from what we know and intend)

Resolve Each of us acts on our intentions; if we intend to do something then we will do it.

I wish each of these three assumptions itself to be included in the scope of Common Knowledge. Thus it will be taken to be common knowledge that we have the intentions that we do and that¹we are rational and act on our intentions. I shall often speak in this context of ‘commitment’ rather than ‘intention’; for given that there is common knowledge of our intentions and of the fact that we act on them, an intention will essentially amount to a commitment. There is some interest in considering to what extent our conclusions depend upon these assumptions, but it will greatly simplify the discussion if we take them to hold.

Now it is clearly desirable that the situation in which we intend to clean the apartment should be one in which you intend to clean the apartment and I intend to clean the apartment. For how else could we expect, through our having the joint intention, that the apartment would be cleaned? Generalizing a little, we may say that with any joint intention will be associated certain individual intentions to the effect that each of us should do his or her ‘bit’. Call these the *component* intentions. What we then require is:

¹Often our component intentions will involve a common action type, such as my intending to clean the apartment and your intending to clean the apartment. But this is not always so. When we form the joint intention to fight one another, for example, my intention is to vanquish *you* while your intention is to vanquish *me*.

Efficacy The situation should be one in which each of us has the appropriate component intentions.

There may be some disagreement as to what we should take the component intentions to be. Bratman ([***) would take them in our running example to be that each of us intend that *we* clean the apartment and not merely that *he* or *she* clean the apartment. There may even be some disagreement as to whether the component intentions are required. Gilbert ([***) thinks that a group may have a joint intention without any of its members having the associated component intentions. However, for present purposes, we need not take a stand on these issues, since everyone will agree that in cases of joint intention it is *possible* (and often desirable) that the agents should have the corresponding component intentions and that the component intentions should at least include the intention for each individual to perform an individual action, even if they also include an intention to perform a joint action.

The obvious way in which Efficacy might be satisfied is if the situation simply consists in my intending to clean the apartment and in your intending to clean the apartment. This gives us a weak ‘distributive’ sense in which we intend to clean the apartment through each of us separately intending to clean the apartment. But it has been commonly supposed that there is a more full-blooded sense in which we might intend to clean the apartment, one in which we form the intention to clean the apartment *together* rather than in isolation. Somehow we operate as a single collective mind rather than as two individual minds on separate but parallel paths.

In attempting to get at this more full-blooded sense of joint intention, it is natural to appeal to the way in which our intentions are interdependent. What I intend depends upon what you intend and perhaps upon what we intend; and what you intend depends upon what I intend

and perhaps also upon we intend. Bratman ([1993], p. 105) has given expression to this peculiar form of interdependence by means of the following necessary condition on joint intention:

We intend to J only if:

1. (a) I intend that we J and (b) you intend that we J
2. I intend that we J because of 1a and 1b; you intend that we J because of 1a and 1b.²

Unfortunately, this condition seems to make joint intention more peculiar than it could possibly be. For it is hard to see what sense of ‘because’ might allow me to intend that we J because, among other things, I intend that we J. Something cannot be the case because it is the case; and no more can it be the case because it is the case and something else is the case.

Perhaps there *is* a non-trivial way in which I may intend that J because, among other things, I intend that J but, absent any explanation, we are left in the dark as to how the peculiar form of interdependence among joint intentions is to be understood.

One might attempt to avoid the vicious circularity of Bratman’s account by making each of our intentions dependent only upon the other person’s intentions. Thus in place of Bratman’s conditions, we have:

1. (a) I intend that we J and (b) you intend that we J
- 2’. I intend that we J because of 1b, viz. that you intend that we J; you intend that we J

because of 1a, viz. that I intend that we J.

But this is little better. For if I intend that we J because you intend that we J and you intend that

²He also proposes a necessary and sufficient condition for joint intention; and his purpose in introducing this condition is not to get at the sense in which our intentions are interdependent but to get round the ‘Mafioso’ example.

we J because I intend that we J, then presumably, by the transitivity of 'because', 'I intend that we J because I intend that we J'. And we are back to the original difficulty.

I believe that the present difficulty arises from his working with too few intentions. We want the component intentions - that you and I intend that we J - to depend upon other intentions that we have. Bratman takes these other intentions to be the component intentions themselves. But there really is no good reason why this should be so. For there are perhaps some underlying intentions that we have and the component intentions could depend upon these rather than upon themselves. We therefore arrive at the following modification of the conditions:

1. (a) I intend that we J and (b) you intend that we J

2''. I intend that we J because of intentions that you and I have; you intend that we J because of intentions that you and I have.

This is alright as far as it goes. But it leaves open what the other intentions might be. It clearly would not do, for example, if my intention to clean depended upon your intention to invite some guests while your intention to clean depended upon my intention to paint the walls. The account also says nothing about the *way* in which our component intentions might depend upon the other intentions (this was also a defect in Bratman's original account). The first of these problems is very serious and will occupy us for a large part of the paper. But the second problem is more manageable and we can make some progress on it right away.

There are at least three ways in which one intention may depend upon other factors - causal, epistemic and rational:

(1) Causal. Your intention to clean the apartment may cause you to raise your eyebrows in an expectant manner which may then cause me to intend to clean the apartment (perhaps

without me being aware of what is going on).

(2) Epistemic. I may know that you intend to clean the apartment and, in the light of that knowledge, I may form the intention to clean the apartment.

(3) Rational. You may intend to clean the apartment and I may intend to clean the apartment if you clean the apartment. Given that I know of your intention, it is then rational for me to form the intention to clean the apartment.

In the first case, your raising your eyebrows (and, indirectly, your intending to clean the apartment) provides a causal basis for my intending to clean the apartment; in the second case, my knowledge that you intend to clean the apartment provides an epistemic basis for my intending to clean the apartment; and in the third case, your intending to clean the apartment provides a rational basis for my intending to clean the apartment, given my conditional intention.

Although these three forms of dependence are all different, I believe we are justified in confining our attention to the rational form. This is not because all cases of causal and epistemic dependence are cases of rational dependence - far from it. Rather it is because we wish to be able to see the agent as clear-eyed and as willing to endorse the way in which their intentions may depend upon prior conditions, so that if certain conditions are in fact a basis - causal, epistemic or otherwise - for his intention to do something, then he should be willing to form the conditional intention to do the thing under those conditions. In the causal case, for example, we would like to see me as being prepared to form the intention to clean the apartment conditionally upon your raising your eyebrows; and in the epistemic case, it should be possible to see me as being prepared to form the intention to clean the apartment conditionally upon my knowing that you will clean the apartment.

Once we allow each person to anticipate how he would wish to react to the intentions and attitudes of the other person, we can suppose that he forms his intentions in ignorance of the intentions and attitudes of the other person. It is as if each of us forms our intentions behind a veil of ignorance, under which the mental states of the other person are hidden from view (the assumption of common knowledge is temporarily suspended, though we form our intentions in the expectation that it will be lifted). The question then is whether these intentions formed in isolation from one another, will somehow 'engage' and exhibit the kind of interdependence that we regard as characteristic of joint intention, once the veil is lifted and the respective intentions and mental states are brought into the common light of day.

There are two other simplifications we can make. We have allowed that each of our component intentions might depend upon the mental states of the other person (and perhaps of myself) and not just on our *intentions*. But if we can anticipate a state of common knowledge, then it is not clear that any mental states besides our intentions need come into play. Suppose, for example, that I intend to clean the apartment conditionally upon my knowing that you intend to clean the apartment. Now in a state of common knowledge, there is no relevant distinction between my knowing that you intend to clean the apartment and your intending to clean the apartment; and so we might just as well assume that my intention to clean the apartment is conditional upon your intending to clean the apartment (and not upon my knowing that you intend to clean the apartment).

In the second place, we have allowed that our component intentions might depend upon the 'objective' facts. But consider a situation in which our joint intention to clean the apartment depends upon our recognizing the objective fact that it is dirty. Then there is

presumably a related situation in which we have a joint intention to clean the apartment *conditional* upon its being dirty and in which our having this conditional intention does *not* depend upon our recognizing the apartment to be dirty. By ‘internalizing’ the objective facts in this way, it is plausible to suppose that our knowledge of them will become irrelevant to our ability to form a joint intention.

We therefore arrive at the following ‘reduction’ of the original situation. Each of us has certain underlying intentions which may be either categorical or conditional in form, with the conditional intentions themselves being conditional upon other intentions of ours; and similarly with regard to those other intentions - to the extent that are conditional they will be conditional upon further intentions of ours; and so on all the way down the line. It is over a structure of underlying intentions of this sort that the component intentions constituting a joint intention will then be formed.³

§2 Interdependence

I have already noted that we want more of a joint intention than that my component intentions should depend upon some of your intentions and your component intention should depend on some of mine. Thus in the case in which my intention to clean the apartment depends upon your intention to invite some guests while your intention to clean the apartment depends upon my intention to paint the walls, our respective intentions simply pass one another by and there is no significant sense in which they are integrated into a single common whole.

³We have something similar here to the reduction of a game in ‘extended form’ to one in ‘normal form’ and, just as in the game-theoretic case, the dynamic features of the process by which we form a joint intention will thereby be lost.

The obvious way to achieve the desired integration is with:

Dependence Each of our component intentions should wholly depend upon the same underlying intentions.

Thus our respective component intentions will be ‘locked’ together through chains of dependence, with any intention upon which my component intention depends being one on which your component intention depends and any intention upon which your component intention depends being one on which my component intention depends. It should be noted that this requirement will be trivially satisfied if each of our component intentions is independent of any other intentions (since they will depend upon the same null set of underlying intentions) and, in order to avoid a trivialization of this sort, we may take an intention to depend improperly upon itself in those cases in which it would otherwise be independent.

This requirement provides us with a sense in which the participants to a joint intention are of ‘one mind’; for in regard to the pattern of dependence there is no distinction to be made between my mind or yours. Your component intention will depend upon certain intentions; and my component intention will depend upon the very same intentions. Thus our collective mind, as constituted by our underlying intentions, will be a common source of our component intentions; and each of us will take the underlying intentions of the other person, and not just of ourselves, to be part of the basis for forming our own intentions.

Bratman’ conditions, which we previously rejected, can also be seen as an attempt to secure a common mind. For he also took there to be a common source for our component intentions. His mistake, if I may call it that, was to identify the component intentions with the underlying intentions, so that each intention would properly depend in part upon itself. But once

we draw the distinction between the two kinds of intention, the difficulties over circularity disappear and the component intentions can be taken to have a common source in something other than themselves.

I previously suggested that we may limit our attention to those cases of dependence in which one intention provides a rational basis for the other; and so let us say a little more about how this is to be achieved. Suppose that someone has the intention to φ conditional upon condition C (we may write this as $I_a[\varphi | C]$, where $\varphi | C$ denotes the act of doing φ conditional upon C and $I[\alpha]$, or I_a , denotes the intention to perform the action α); and suppose he also accepts that C . He may then derive the categorical intention to φ ($I_a\varphi$) via the rule of ‘Rule of Detachment’⁴:

$$\begin{array}{c} C \quad I_a[\varphi | C] \\ \hline I_a\varphi \end{array}$$

and the two premisses thereby provide a rational basis for the conclusion. Suppose, for example, that:

- (a) I intend to clean the apartment if you intend to clean the apartment; and
- (b) you intend to clean the apartment.

It then follows by Detachment that I intend to clean the apartment. Of course, as with any rule of

⁴For our purposes it does not matter whether one thinks of Detachment as a rule for inferring the *statement* of an intention or for inferring the *intention* itself.

inference, I may question one of the premiss in preference to inferring the conclusion. But we may suppose that in the cases of interest to us the premisses are sufficiently secure that this is not an option.

The conclusion of one application of Detachment may serve as the premiss of another.

Case 1 Suppose:

(a) I intend to clean the apartment all conditional upon your intending to clean the apartment if I intend to clean the apartment ($I_1[\varphi \mid I_2[\varphi \mid I_1\varphi]]$); and

(b) you intend to clean the apartment if I intend to clean the apartment ($I_2[\varphi \mid I_1\varphi]$).

From (a) and (b), we get by Detachment:

(c) I intend to clean the apartment ($I_1\varphi$).

From (b) and (c), we then get by a further application of Detachment:

(d) you intend to clean the apartment ($I_2\varphi$).

These various applications of Detachment (and perhaps of other rules as well) may be set out in the form of a derivation; and the resulting concept of derivation provides us with an obvious sense in which one intention may depend upon another. For we may take this to mean that there is a derivation in which the one intention serves as one of the assumptions and the other intention serves as the conclusion. Thus the derivation of intention (c) above shows that it depends upon intentions (a) and (b) and the subsequent derivation of intention (d) shows that it also depends upon intentions (a) and (b).

However, not all cases of this sort are genuine cases of dependence. Consider the following odd case:

(a) I intend to clean the apartment ($I_1\varphi$)

(a)' You intend to clean the apartment ($I_2\phi$)

(b) I intend to clean the apartment if I intend to clean the apartment all conditional upon your intending to clean the apartment ($I_1[[\phi | I_1\phi] | I_2\phi]$)

(b)' You intend to clean the apartment if you intend to clean the apartment all conditional upon my intending to clean the apartment ($(I_2[[\phi | I_2\phi] | I_1\phi])$).

From (a) and (b)', we get (by Detachment):

(c) You intend to clean the apartment if you intend to clean the apartment ($I_2[\phi | I_2\phi]$).

From (c) and (a)', we get (a)' again (but depending now on (a)). From (a)' and (b), we get:

(c)' I intend to clean the apartment if I intend to clean the apartment ($I_1[\phi | I_1\phi]$).

And from (c)' and (a), we get (a) again, but depending now on all four assumptions (a), (a)', (b) and (b)'.⁵ But clearly we do not want to say that my intention to clean the apartment depends upon that very intention or upon the other intentions

There are at least two grounds upon which one might exclude such derivations (quite apart from the oddity of the intentions involved). The first is that the derivation is uneconomical; there is no need to appeal to the other intentions in deriving my intention to clean the apartment,

⁵We may set out the derivation in tree form as follows:

$$\begin{array}{c}
 \begin{array}{c}
 I_1\phi \quad I_2[[\phi | I_2\phi] | I_1\phi] \\
 \hline
 I_2\phi \quad I_2[\phi | I_2\phi] \\
 \hline
 I_2\phi \quad I_1[[\phi | I_1\phi] | I_2\phi] \\
 \hline
 I_1[\phi | I_1\phi] \quad I_1\phi \\
 \hline
 I_1\phi
 \end{array}
 \end{array}$$

since we already have it as an assumption. The second is that the derivation is circular; the derivation of my intention to clean the apartment already rests upon the assumption that I intend to clean the apartment.

So let us agree to say that one intention (genuinely) depends upon another if there is an economical - or, alternatively, a non-circular - derivation in which the one intention serves as an assumption and the other as the conclusion. We are thereby able to give precise 'proof-theoretic' meaning to our understanding of Dependence; there should be an economical (or non-circular) derivation of each of the component intentions from the same set of underlying intentions.

The requirement of Dependence might appear to be something of a luxury. For why, apart from our desire for a common mind, should it matter that our component intentions have a common origin? And even if it does matter, then why should we insist on complete interdependence rather than on some form of partial interdependence? But there is in fact an independent motivation for insisting upon complete interdependence. Suppose, in the first place, that your intention to clean the apartment did not depend upon all of my underlying intentions. Keeping your underlying intentions fixed and eliminating some of mine, you would still be committed by the resulting set of our intentions to cleaning the apartment. But given that this is so, it is not clear why I should not simply eliminate the superfluous intentions from my own intention set. After all, you will still then be committed to cleaning the apartment; and it is possible that I would not be similarly committed (since there are fewer assumptions from which my commitment might be derived). So assuming that I would prefer to have you clean the apartment on your own to having us clean the apartment together, there is nothing to be lost by my eliminating the superfluous intentions and possibly something to be gained. We may

therefore assume, given that such superfluous intentions have in fact been eliminated, that Dependence will at least be ‘asymmetrically’ satisfied; each of our component intentions will depend upon all of the underlying intentions of the other person (though perhaps not on all of our own underlying intentions, some of them may simply serve the purpose of getting the other person to do his bit).

Suppose now that Dependence is asymmetrically satisfied and that, contrary to its being fully satisfied, my component intention to clean the apartment does not depend upon all of my own underlying intentions. Keeping your underlying intentions fixed and eliminating some of mine, I would still be committed to cleaning the apartment although you would not be so committed (given Dependence is asymmetrically satisfied). But this means that you have made yourself hostage to my intentions. For by restoring the additional intentions to my intention set, I can thereby commit you to cleaning the apartment without thereby committing myself (since I am *already* committed to clean the apartment). But this may be regarded as unsatisfactory. For it might be thought that it is only through incurring a like commitment that I should thereby be able to commit you; and so for this reason, a dependence of my component intention on only some of my own underlying intentions should also not be allowed (and similarly for you).

§3 Symmetry

A further plausible condition on joint intention is:

Symmetry My underlying intentions should be the mirror-image of yours.

Thus if it is part of the situation in which we have a joint intention that I intend to clean the apartment then it should also be part of the situation that you intend to clean the apartment, or if

it is part of the situation that I intend to clean the apartment if you do then also part that you intend to clean the apartment if I do, and similarly for any other intentions we might have.

Symmetry provides another way in which our intentions might constitute a single mind - not now by way of a common source, but by way of a common content. It is not of course possible for the content of our intentions to be exactly the same. If you intend to clean the apartment, i.e. for *you* to clean the apartment, then I will not intend for *you* to clean the apartment. At best, I can only intend for *me* to clean the apartment. But it *will* be possible for the content of our intentions to be the same once we make due allowance for the difference in who we are, what goes for me and goes for you; and Symmetry is then simply the condition that this should be so.

Symmetry is perhaps not essential to our having a joint intention and, apart from the general desire to achieve some kind of unity of mind, is perhaps of no great intrinsic value. But we are inclined to think that even if Symmetry is not *required* for us to have a joint intention, it is not something that should *stand in the way* of our having a joint intention. For if two rational agents attempt to form a joint intention, they may well end up with the same underlying intentions; and we would not want the similarity in means to prevent them from achieving a unity in purpose.

Symmetry rules out cases in which one of us 'makes a first move'. Suppose we are at a concert and the performance has just finished. For a brief moment there is a pregnant pause; and then, at an accelerating rate, the audience breaks out into thunderous applause. But how? No one wants to clap if others are not clapping (this is Vienna not New York). So perhaps someone makes a tentative first move and claps very quietly. Others now act on the intention to clap

slightly louder than the others around them (up to certain point that is, one hopes, a function of the quality of the performance). In this way, one intention builds upon and spreads throughout the auditorium until everyone is clapping at the maximal level.

But one would like it possible for us to form the joint intention to clap without any one of us making a first move. After all, there may be a differential cost associated with making a ‘first move’ (the risk of standing out, of being the only one to clap etc.); and one would like it to be possible to form a joint intention without any of us having to bear the cost alone. More generally, with a failure of Symmetry comes a difference in commitment and, with a difference in commitment, comes a possible difference in cost; and so it is only through the satisfaction of Symmetry that one can be assured that none of agents will be unfairly burdened with the cost of forming (and acting upon) a joint intention.

§4 Non-Vulnerability and Optimality

In forming a joint intention, there are certain defensive measures that each of us may wish to make. One is not to expose ourselves to the possibility of acting unilaterally.

Suppose that I have an intention to do my bit conditional upon the satisfaction of the conditions C_1, C_2, \dots, C_n (we might think of this as an intention of the form $I_1[\dots [[\phi | C_1] | C_2] \dots | C_n]$).

Then I would not want this conditional intention along with the obtaining of its conditions to commit me to doing my bit without also committing you to doing your bit. Suppose, for example, that I intend to clean the apartment conditional upon: your intending to clean the apartment if I clean the apartment. Then my conditional intention along with the obtaining of the condition (that you intend to clean the apartment if you do) will commit me to cleaning the

apartment. But it will also commit you to cleaning the apartment, given that I intend to clean the apartment and that you have the intention to clean the apartment if I do. Suppose, on the other hand, that I intend to clean the apartment conditional upon: your intending to clean the apartment conditional upon my intending to clean the apartment if you do. Then my conditional intention along with the obtaining of the condition will commit me to cleaning the apartment. But they will not, in this case, commit you to cleaning the apartment; and so I will find myself saddled with a unilateral commitment to clean the apartment.

More generally, we may require:

Non-Vulnerability Suppose some of my underlying intentions in conjunction with other intentions (mine or yours) commit me to doing my bit. Then they should also commit you to doing your bit.

In other words, I do not wish to form intentions that might end up committing me to doing my bit without also committing you to doing your bit (and similarly, of course, with you and me exchanged).

Not only may I wish to protect myself against a unilateral commitment, I may also wish to protect myself against making an unnecessary bilateral commitment. Suppose that I form the conditional intention to clean the apartment if you do and that you form the categorical intention to clean the apartment. I will thereby be committed to cleaning the apartment. But I may regret having formed the conditional intention. For I thereby convert a situation in which you have made a unilateral commitment to clean the apartment into one in which we have made a bilateral commitment; and I have thereby missed out on the chance of exploiting a situation in which you have made a unilateral commitment to clean the apartment.

This suggests the following condition:

Optimality Suppose some of my underlying intentions in conjunction with the satisfaction of their antecedent conditions commit me to doing my bit. Then the satisfaction of the antecedent conditions should commit me to doing my bit if they already commit you to doing your bit.⁶

For suppose the satisfaction of the antecedent conditions unilaterally committed you to doing your bit. My underlying intentions would then convert a unilateral commitment into a bilateral commitment and would thereby prevent me from exploiting the situation.

It has to be admitted that the last condition has a somewhat different character from the others. In many situations, especially those involving few people who are known to one another, the agents will be of a cooperative disposition. They do not want others to take advantage of them, but nor do they want to take advantage of others; they are neither saints nor bastards. Such agents would then have an interest in the satisfaction of Non-Vulnerability, since this rules out the possibility of others taking advantage of them, but they would have no interest in the satisfaction of Non-Optimality, since they would have no interest in taking advantage of others even when they had exposed themselves to the risk.

But in other situations, especially those involving many people who are not known to one another, the agents will be of a non-cooperative disposition. They do not want others to take advantage of them but are perfectly willing to take advantage of others should the occasion arise. They are potential free-riders or bastards, prepared to exploit but not willing to be exploited.

⁶It may be sufficient, for our purposes, to consider a single underlying intention under which I intend to do my bit under specified antecedent conditions. It is then automatic that the intention and its antecedent conditions will guarantee that I do my bit.

Such agents would not then be willing to form a joint intention unless they could be assured of the satisfaction of both Non-Vulnerability and Optimality.

Most writers on joint intention have largely confined their attention to the case of cooperative agents - with two or more people engaged in a cooperative enterprise in which each person is willing to do his bit as long as the others do their bit. Indeed, in many of the cases that are considered, it is not even possible for the people involved to succeed in the cooperative enterprise unless everyone does his bit (two people must move the piano, or sing a duet, or have a discussion).

But the cases of most interest to social and political philosophy are those in which the joint enterprise can succeed without everyone doing their bit and in which many people will not be willing to do their bit even though sufficiently others are. A typical case is the payment of taxes. Enough of us pay taxes for all of us to benefit. But that does not prevent many of us from not paying their taxes, even in the full knowledge that most of us do. If such people are brought into an agreement to engage in a joint enterprise then there is no reason to think that they would be willing to commit themselves to a decision that would prevent them, should the occasion present itself, from being free-riders. The very possibility of forming a joint intention or agreement in such cases will therefore depend upon the satisfaction of Optimality.

§5 Satisfaction of the Conditions

We have considered four conditions on the intentions which may underlay a case of joint intention - Dependence, Symmetry, Non-Vulnerability and Optimality. But can they all be satisfied?

In considering this question, it will be helpful to distinguish between *ordinary* and *special* intentions. We suppose that there are certain categorical intentions with an ordinary content, such as cleaning the apartment or paying one's taxes. From these categorical intentions, we can then use ordinary logical means to form other intentions. We might conditionalize, for example, to construct the intention to pay one's taxes if everyone else pays their taxes or we might conditionalize on a conditional intention to construct the intention to pay one's taxes if everyone else pays their taxes as long as every other person has the intention to pay his taxes if everyone else pays their taxes; and so on, though every increasing degrees of logical complexity. An ordinary intention is one that can be constructed in this way from the repeated application of logical operations to ordinary categorical intentions.

Suppose now that only ordinary intentions are at our disposal. Is there any way in which all four conditions might be satisfied? To get a feel for the problem, let us consider how various obvious proposals fare.

Case 1 I intend to clean the apartment and you intend to clean the apartment.

Symmetry alone is satisfied. Dependence fails, for example, since my intending to clean the apartment does not depend your intending to clean the apartment.

Case 2 I intend to clean the apartment and you intend to clean the apartment if I intend to clean the apartment.

None of the conditions are satisfied. Dependence fails, for example, for the same reason as before and Symmetry is clearly not satisfied. My categorical intention to clean the apartment makes me vulnerable; and your intention to clean the apartment if I intend to is non-optimal (since it prevents me from exploiting the situation in which you form the categorical intention to

clean the apartment).

Case 3 I intend to clean the apartment if you do and you intend to clean the apartment if I do.

Symmetry and Non-Vulnerability are satisfied but neither Dependence nor Optimality. Dependence fails since it is not possible to derive my or your intention to clean the apartment; and Optimality fails since my intention to clean the apartment if you do prevents me from standing by should you form a categorical intention to clean the apartment.

Case 4 I intend to clean the apartment if you intend to clean the apartment and you intend to clean the apartment if (I intend to clean the apartment if you intend to clean the apartment).

Dependence and Non-Vulnerability are satisfied but neither Symmetry nor Optimality. The failure of Symmetry is obvious and Optimality fails for the same reason as before.

In the light of this quick survey of some cases, it might be thought that at best only two of the four conditions can be satisfied by ordinary intentions. However, it turns out, much to my surprise, that this is not so and that any three-membered subset of the conditions can also be satisfied by ordinary intentions. The case of most interest is the one for the cooperative agent, in which all of the conditions but Optimality are to be satisfied.

To see how these conditions might be satisfied, it will be helpful to set up some terminology. Let *my intention to do my bit* be my intention to clean the apartment (and similarly for *your intention to do your bit*); let *our accord* (to clean the apartment) be the conjunction of my intention to do my bit if you intend to do your bit and your intention to do your bit if I intend to do my bit; and let *my intention to do my bit in reaching accord* be my intention to do my bit if you do your bit (and similarly for *your intention to do your bit in reaching accord*). We now

consider the following four intentions:

1. My intention to do my bit upon our reaching accord;

1'. Your intention to do your bit upon our reaching accord.

2. My intention to do my bit in reaching accord conditional upon your intention to do your bit upon our reaching accord;

2'. Your intention to do your bit in reaching accord conditional upon my intention to do my bit upon our reaching accord.⁷

Thus under 1 and 1', our accord is treated as a sufficient basis for each of us to intend to do our bit and, under 2 and 2', the previous conditional intention of the other person is treated as a sufficient basis for each of us to intend to do our bit in reaching accord.

It may then be shown, when these are the underlying intentions, that all of the conditions with the exception of Optimality, are satisfied. Let us first show how we may derive my intention to do my bit and your intention to do your bit. From 1 and 2', we get your intention to do your bit in reaching accord. Similarly, from 1' and 2, we get my intention to do my bit in reaching accord. We thereby reach accord. But then by 1, I intend to do my bit and, by 1', you intend to do your bit (and we should note that the derivation of each intentions makes use of all four intentions).

Let us now show that Optimality fails (and leave it at that). Consider my intention to do my bit in reaching accord conditional upon your intention to do your bit upon our reaching accord. Suppose the antecedents of this intention are satisfied. Then you intend to do your bit

⁷Let ψ_1 be $I_1[\varphi \mid I_2 \varphi]$ and ψ_2 be $I_2[\varphi \mid I_1 \varphi]$. Then the intention under 1 above is $I_1[[\varphi \mid \psi_2] \mid \psi_1]$ and the intention under 2 is $I_1[\psi_1 \mid I_2[[\varphi \mid \psi_1] \mid \psi_2]]$; and similarly for 1' and 2'.

(the antecedent of my intention to do my bit in reaching accord) and you intend to your bit upon our reaching accord. It follows that you intend to do your bit but not that I intend to do my bit. Thus my conditional intention converts a unilateral intention into a bilateral intention and is therefore not optimal.

So ordinary intentions can satisfy any three of the desiderata. But can they satisfy them all? I thought I had a proof that they could not but it contained a mistake. So all I am left with is:

The Impossibility Conjecture The four desiderata cannot be satisfied by ordinary intentions.

If the conjecture is indeed correct, then there is not merely a significant philosophical difference between the case of cooperative and non-cooperative agents but also a significant logical difference. For in the former case, intentions of an ordinary sort may be used to secure an agreement between the interested parties. But in the latter case, this is not possible. Either agreement is not possible at all or it must be secured by other means. And even if the conjecture is not correct, it would still be desirable to find some more straightforward way in which the intentions underlying an agreement in the cooperative or non-cooperative cases might be formulated.

But what might these other means be? What conceptual resources besides those used to formulate the ordinary categorical intentions and the corresponding logically complex intentions could there be? It is to this question that we now turn.

§6 Reciprocal Intention

There is a natural and simple way in which we might attempt to satisfy the four conditions above and which is independently plausible as a way in which we might form a joint intention. We take the underlying intentions that constitute our joint intention to clean the apartment to be as follows:

1. I intend to clean the apartment conditional upon your having a like intention
- 1' You intend to clean the apartment conditional upon my having a like intention.⁸

Here it is important that the notion of *like intention* be properly understood. Your having a like intention is not your having an intention that is like my intention to clean the apartment (in which case your having a like intention would be your having the intention to clean the apartment). Rather, your having a like intention is your having an intention that is like my intention to clean the apartment conditional upon your having a like intention. In other words, it is to be like my whole conditional intention and not its consequent part. Thus on this understanding of like intention, your having a like intention would be a matter of your intending to clean the apartment conditional upon my having a like intention.

We therefore arrive at the following formulation of the underlying intentions:

1. I intend to clean the apartment if 1'
- 1' You intend to clean the apartment if 1.

It is very plausible if these are the underlying intentions that the four conditions will be satisfied. Take each in turn:

Dependence My intention to clean the apartment can be derived by Detachment from 1

⁸A similar formulation is given by Velleman [***], who takes himself to be developing a suggestion of Margaret Gilbert [***]

and 1' but not from 1 or 1' alone.

Symmetry Obvious

Non-Vulnerability The only circumstances in which I am committed by intention 1 to clean the apartment are those in which you are committed to intention 1' and so are also committed to clean the apartment.

Optimality If antecedents of both 1 and 1' are obtain then 1 and 1' obtain (since they *are* the antecedents of 1 and 1') and so both of us are committed to doing our bit. On the other hand, if only the antecedent of 1 obtains, which is to say that 1' obtains, then neither of us is committed to doing our bit (and similarly if only the antecedent of 1' obtains). Thus in neither case are my intentions not optimal.

Let us call intentions of the form of 1 and 1' *reciprocal intentions*. It then seems clear that if there are reciprocal intentions, then our four desiderata can be satisfied. But the big question now is whether there *are* any reciprocal intentions. Is it possible for our intentions to have the reciprocal forms prescribed by 1 and 1'?

It might be thought that the answer is obvious 'no'. For if 1 and 1' are abbreviations for the sentences to their right, then the process of disabbreviating 1 and 1' will never come to an end and so there is nothing we can properly take these sentences to be.

In order to avoid difficulties of this sort, let us give a more cautious formulation. What we are looking for are sentences R and R' for which:

- (i) it is apriori that R iff you intend to clean the apartment conditional upon R';
- (ii) it is apriori that R' iff I intend to clean the apartment conditional upon R;

(iii) it is possible (not apriori false) that both R and R'.⁹

Thus there is no need for R' to *abbreviate* the sentence 'you intend to clean the apartment if R'. It will be sufficient for our purposes if the two are a priori equivalent since then one can always be substituted for the other within the scope of the intention-operator (given Rationality) and the proof of the satisfaction of the desiderata will go through much as before.

It is plausible if (i) and (ii) hold that (iii) will also hold. For suppose that it is impossible that R. Then since R' is a priori equivalent to the intention to clean the apartment conditional upon R, then R' (which has an impossible antecedent) is possible (and might even be taken to be necessary). Similarly if R' is impossible then R is possible. So either R or R' is possible. But R is possible iff R' is possible by considerations of symmetry; and given that each is possible it is then plausible that both together are possible.

Thus the problem is whether we can simultaneously solve for R and R' in the 'equations' (i) and (ii) above. If the Impossibility Conjecture is correct, then we know that R and R' cannot be formulated by ordinary means; and so we need to provide some alternative means by which the reference to 'like' intentions might be understood.

§7 The Indexical Solution

Let me briefly discuss a solution to this problem that has been independently proposed by David Velleman (for joint intention) and by Chris Peacocke (for joint *attention*). . The proposal calls for much more discussion than I can give it here, but I hope I say enough to indicate why I

⁹It may not be a priori that you and I exist and so, strictly speaking, the aprioricity should be taken to be conditional on our existence (and, for later purposes, also on our being distinct).

think it is problematic. I will stick to the case of joint intention, though similar considerations apply to the case of joint attention.

Suppose that each of us has an intention to clean the apartment if the other has a like intention; and designate these token intentions by i_0 and j_0 . Then what is the content of i_0 and j_0 ? The content of i_0 is: to clean the apartment if you have a like intention, i.e. if you have an intention j whose content is the same as the content of i_0 but with i_0 replaced with j and with the participants (you and I) interchanged. And similarly for j_0 . Its content is: to clean the apartment if I have an intention i whose content is the same as the content of j_0 but with j_0 replaced with i and with the participants interchanged.

This is not quite the proposal of Velleman or Peacocke. For one thing, the indefinite reference to j or i is replaced by definite reference to j_0 and i_0 . Thus the content of i_0 would now be: to clean the apartment if the content of j_0 is the same as the content of i_0 but with i_0 interchanged with j_0 and with me interchanged with you. They also suppose that the token intention i_0 involves *indexical* reference to the token intention j_0 ; and similarly for j_0 . But neither of these features is essential to the proposal and in fact makes it far less flexible (the first prevents just one of us from doing our bit in forming a reciprocal intention and the second prevents one from forming the *hypothesis* that you and I have formed a reciprocal intention).

What is characteristic of this proposal is that reference to the content of an intention is made in specifying the content of an intention. Now this is not in itself bad. Take the corresponding case of belief. The content of my belief might be that the content of your belief is true. Suppose that the content of your belief is that grass is green. Then both my belief and your belief seem clearly to be true. But suppose that the content of your belief is that the content of

my belief is false (this corresponds to the famous ‘envelope’ paradox). Then we do not know what to say. For it seems that my belief will be true iff your belief is true and that your belief will be true iff my belief is false. Consequently, my belief will be true iff it is false. A contradiction.

We are in the area of the paradoxes of self-reference and so the cases must be handled with some care. Now although there has been a great deal of discussion of the paradoxes there has not, as far as I am aware, been much discussion of the present kind of case in which the self reference is by way of explicit reference to content. But the most plausible view of such cases is one in which the content is only taken to be truth-evaluable when the reference to content ‘bottoms out’. Thus in the case in which the content of my belief is that the content of your belief is true and the content of your belief is that grass is green, the reference to content in my own belief will ‘bottom out’ in the content that it is raining, while in the other case in which the content of my belief is that the content of your belief is false, the reference to the content of our respective beliefs will not bottom out. There will be no saying what it is without already presupposing what it is.

Returning to the case of our having like intentions, we see that the content of our intentions will *not* bottom out; each time we attempt to say what the content is we again find ourselves making reference to what it is. It is therefore not clear that the intentions have been formed by legitimate means or that they have a truth-evaluable content.

I do not want to dismiss the present approach altogether; and it is even conceivable to me that it might somehow relate to my own approach. But as things stand, we merely have an inchoate idea, where what we need is a theory.

§8 Cool

I want now to take a detour through a theory of response dependence that I have developed. Once we see how the theory goes, we will see that it provides the resources to provide a coherent formulation of reciprocal intention.

One term may be defined by means of others. Whatever the actual meaning of the word ‘bachelor’ we might define it by means of the following definition:

x is a bachelor $=_{df}$ x is an unmarried man.

What it is (by definition) for someone to be a bachelor is for him to be an unmarried man. We have here an example of an explicit non-circular definition; and no one, it seems to me, could sensibly deny the legitimacy of such a definition.

But consider now the following question. Whatever the actual meaning of ‘cool’ (or ‘hip’), could we legitimately define it by means of the following definition:

x is cool $=_{df}$ x is taken to be cool,

so that what it is (by definition) for something to be cool is for it to be taken to be cool?

There may, of course, be some indeterminacy in what it is for something to be *taken* to be cool. Must only some of us so take it or all? And when or in what manner? But let us suppose that we somehow resolve these indeterminacies. Perhaps what we mean is that all of us right now must judge the thing to be cool. Do we then have a legitimate definition?

One obvious response to this question is that we do not. For there is a striking difference between the two definitions. The first (of ‘bachelor’) is non-circular, while the second (of ‘cool’) is circular. The very term to be defined on the left in the definiendum occurs on the right in the

definiens. But how can a circular definition succeed in assigning a meaning to the term that is to be defined? We want to know what it is for something to be cool. This is defined in terms of its being taken to be cool. But what then is the content of the taking? It is that the thing is cool which, again, is a matter of taking it to be cool. We have entered an endless circle and can provide no account of what it is to be cool in terms of concepts that we already understand.

This is the response I myself would have given a few years ago. But it suddenly struck me with great force that definitions of this sort might be legitimate after all and that, if they were accepted, then a number of seeming intractable problems concerning the nature of certain concepts would simply disappear. The matter calls for a much more thorough discussion but let me briefly indicate the reasons for my change of heart. The putative definition of ‘cool’ above has the form of an explicit definition and this makes it natural to think that we understand the term on the left by means of the expression on the right. But although it has the form of an explicit definition, it might provide us with an understanding of the term in the manner of an implicit definition. In other words, our understanding of the term to be defined might be given by the definition as a whole. So whereas we grasp the concept *bachelor* above as the concept of being an unmarried man, we grasp the concept *cool* as the concept for which to be cool is to be taken to be cool.

Moreover, grasp of concepts of this sort seems to be well within our reach. Suppose I announce to you all: look, this is what it is to be cool - to be cool is for all of you right now to take it to be cool. And suppose all of you have been convinced by my previous argument and thereby take yourselves to have grasped a concept. I now ask: is Tom cool?

In considering this question, each of you will be considering whether everyone takes

Tom to be cool. Now in considering this question, each of you may reason as follows. If I do not take Tom to be cool but everyone else does, then I will be solely responsible for Tom's not being cool. This is a risk I would prefer not to take. Moreover, everyone else may be equally disinclined to take the risk and so it looks like a safe bet to take Tom to be cool. A vote is taken and, lo and behold, everyone takes Tom to be cool and is thereby vindicated in their judgment.

Thus it very much looks as if each of us has grasped the concept *cool*. We form thoughts with its help, we reason on the basis of those thoughts, and we attempt to ascertain whether the thoughts are correct on the basis of the criterion of application for the concept.

Still, a great deal more needs to be done in defending the present position. For the concept we take ourselves to grasp might not be coherent. Suppose I define the term 'para' as follows:

x is para =_{df} x is not para.

I suppose it is conceivable that someone might take themselves to have thoughts about what is para on the basis of this definition. But there are no real thoughts to be had, since there is no coherent conception of what it is to be para in terms of which they might be given. The case of 'cool' does not look like the case of 'para'. But what is the difference? Why is the one definition legitimate but not the other?

To answer this question, we need criteria for sorting out the good circular definitions from the bad. I should like to propose two criteria by which this might be done. One I call 'Safety' and the other 'Harmony'. Let me explain each in turn.

Safety. By a purely intensional operator for a propositional attitude (such as belief, knowledge, intention or the like) I mean one whose application to a proposition is in general

independent of whether the proposition is true or false. Thus ‘belief’ is a purely intensional operator since whether one believes a proposition is in general independent of whether it is true or false. ‘Knowledge’, on the other hand, is not purely intensional since knowledge of a proposition requires that it be true.

Say that the occurrence of a term is *safe* if it is within the scope of a purely intensional operator. Thus ‘cool’ has a safe occurrence in ‘John believes that Tom is cool’ but an unsafe occurrence in ‘John knows that Tom is cool’. A definition is then said to be *safe* if each occurrence of the defined term in the definiens is safe. Thus:

$x \text{ is cool} =_{df} x \text{ is taken (i.e. believed) to be cool}$

is safe, while our previous definition:

$x \text{ is para} =_{df} x \text{ is not para}$

is not.

Harmony. There is something that counts as ‘success’ or ‘appropriateness’ for a propositional attitude. So, for example, in the case of belief it is that the belief should be true. The other criterion is that the definition of the concept should be compatible with success in our responses by which the concept is defined. In the case of the definition:

$x \text{ is cool} =_{df} x \text{ is taken (i.e. believed) to be cool}$

this means that we should be able to identify $x \text{ is taken to be cool}$ with being cool. If we make the identification then what the definition requires is that:

for every x , $x \text{ is cool iff } x \text{ is cool}$,

which is guaranteed by logic. Consider, by contrast:

$x \text{ is catch-22} =_{df} x \text{ is not taken to be catch-22}$.

Harmony then requires:

for every x , x is catch-22 iff x is catch-22

which cannot be true (unless there are no objects whatever).

What I would now like to propose is that any circular response-dependent definition can be taken to be legitimate as long as it conforms to Safety and Harmony. There is a great deal more to be said in defense of the view, but it may be shown that with these restrictions in force the definitions will not get us into trouble; they can be seen to define a single concept; and supposing the concept to exist will not lead to contradiction or lead us to accept anything that we were not already committed to before the introduction of the concepts.

§9 The Response-Dependent Account of Reciprocal Intention

I want now to apply the above theory of response dependent concepts to the earlier problem and show how there can be reciprocal intentions (and hence satisfaction of the four desiderata for joint intention) once we provide a response dependent account of what they are.

Suppose that only our joint intention to clean the apartment is in question (it is easy to generalize to other actions and to any number of agents). Let us now define what it is for one of us to have the reciprocal intention to clean the apartment:

x reciprocally intends to clean the apartment $=_{df}$ either (i) x is me and I intend to clean the apartment conditionally upon you reciprocally intending to clean the apartment or (ii) x is you and you intend to clean the apartment conditionally upon my reciprocally intending to clean the

apartment.¹⁰

This is a circular definition. The concept of reciprocally intending, defined on the left, also appears on the right. For it to be a legitimate definition according to our theory of response dependence, two conditions must be met. First, the definition should be subject to the ‘safety’ constraint; the concept to be defined should only occur within the scope of a purely intensional operator on the right. But *intention*, like *belief*, would appear to be a paradigm of a purely intensional operator. Indeed, all that strictly matters to the safety of the definition is that the intention operator be purely intensional with respect to its second argument (the antecedent condition). But this is surely so. Whether I intend to take an umbrella conditional upon its raining, for example, is independent of whether it rains.

Second, the definition should be subject to Harmony; the definition should be compatible with ‘success’ in our responses. In the case of a conditional intention, there are two ways in which it might be successful - through the non-satisfaction of its antecedent condition (so that the intention is not even ‘activated’) or through the performance of the action. For present purposes, focus on the latter and suppose that reciprocally intending to clean the apartment is just a matter of cleaning the apartment. Then the definition will tell us that I will clean the apartment just in case I am me and I clean the apartment or I am you and you clean the apartment (and similarly for you). And this is true.

So the theory tells us that this is indeed a legitimate definition. But we are now a short step away from securing the possibility of reciprocal intention. For instances of the definition

¹⁰The use of the personal pronouns ‘I’ and ‘you’ here is completely incidental to the formulation and they could just as well be replaced by names.

will yield a priori equivalences. Thus we will have the apriority of:

I reciprocally intend to clean the apartment iff either (i) I am me and I intend to clean the apartment conditionally upon you reciprocally intending to clean the apartment or (ii) I am you and you intend to clean the apartment conditionally upon my reciprocally intending to clean the apartment.

But this means that we have the apriority of:

I reciprocally intend to clean the apartment iff I intend to clean the apartment conditionally upon you reciprocally intending to clean the apartment and similarly with you in place of me - which gives us what we want.

The above resolution of our problem may appear disappointing. We wanted to establish the existence of a solution to our 'equations'. But have we not just *stipulated* the existence of a solution?

It should be pointed out, in the first place, that we have not simply stipulated a solution. The legitimacy of the proposed definition depends upon the satisfaction of certain stringent conditions (Safety and Harmony); and it is by no means a foregone conclusion that a definition meeting these conditions will also serve to solve the equations. Nor is there anything arbitrary about the adoption of the proposed definitions; they fall under a general rubric for which a general proof of legitimacy can be given.

It should be mentioned, in the second place, that the present proposal is not intended to be a case of 'business as usual'. An analogy with the extension of the number system may help to explain what I mean. There was a time, prior to the recognition of real numbers, when it was hard to say whether the equation $x^2 = 2$ had a solution. We know from Euclid's proof that it has

no rational solution. So does it have no solution at all or does it have some other kind of solution? To solve this problem, what was required was an expanded conception of what numbers there could be - one which allowed room not just for rationals, i.e. ratios of integers, but also for the limits of rationals. Once the number system was expanded in this way, there was then no difficulty in seeing that the equation had a solution. Thus the great advance lay not in solving the equation in any ordinary sense of the term but in seeing how the number system might be expanded so as to allow for a solution.

Similarly, in the present case. There is the question of whether our 'equations' have a solution. They appear to have no solution within the existing realm of concepts. So do they have no solution at all or do they have some other kind of solution? Again, what is required to solve this problem is the recognition of a broader conception of what concepts there can be - one that allows room for concepts with a certain kind of circular definition in addition to the more usual non-circular forms of definition. Once the realm of concepts is expanded in this way, there is no great difficulty in seeing that the equations have a solution. Thus the essential advance lies not in the application of the existing ontology of concepts but in its expansion to cover the new kind of case.

§10 Some Implications

The present proposal has some larger implications within social, political and legal philosophy which I would briefly like to discuss:

(1) Consider a typical prisoner dilemma-type situation: I would be better off not paying my taxes regardless of what everyone else does. So it would appear to be rational (or in my best

self-interest) to 'defect', i.e. not pay my taxes. And yet each of us would prefer an outcome in which we all 'cooperate' (pay our taxes) to one in which we all defect. .

Social and political philosophers have considered a number of ways in which this unsatisfactory outcome might be avoided. Thus we might alter the payoffs (by attaching a sanction to defection) so that it is no longer in most people's self-interest to defect or we might consider an iterated game in which cooperators will tend to do better than defectors (under an evolutionary model, it might then be the cooperators who survive).

But our work suggests another approach in which we reconfigure, not the payoffs or the personalities of the players, but the space of alternatives that they consider. We have supposed that each person is confronted with the question of whether or not to pay their taxes. But since what they intend to do will depend upon what others intend to do, it is natural to suppose that they might instead wish to consider the question of whether to make a commitment conditional upon the commitments of others.

So let us suppose that each person is considering the question of whether to pay their taxes conditional upon everyone else makes a reciprocal commitment to pay their taxes. I can either make the commitment or not. The structure of the pay-off, with these new alternatives in place, has now changed: if we all make the commitment then everyone pays their taxes and we get a positive pay-off; if we do not all make the commitment then no one pays their taxes and we get a zero pay-off. Thus by reconfiguring the space of options, we get the dominant strategy to favor cooperation over defection; and this suggests, more generally, that reciprocal commitments and the like may play an important role in overcoming the inefficiencies that might otherwise arise from our making 'straight' self-interested decisions.

(2) I have focused on one particular concept, that of joint intention. But my suspicion is that a large number of other social concepts are also response dependent. They include those for joint action, for common belief and knowledge, and for social institutions and the roles and devices within them. Let me consider the case of law as an example (my discussion will be very brief and tentative).

What is it for a putative piece of legislation to be the law? An answer to this question might take the following form. There is a certain substantive (or nonconceptual) criterion for being the law - here in the States it is a matter of being appropriately approved in both Houses and by the President. Something then is the law if it meets these criteria. That is:

x is the law $=_{df}$ there is a criterion C for being the law which x meets.

But this now pushes back the question to what it is for C to be a criterion for being the law. This is, of course, a difficult question. But one aspect of what it is for C to be such a criterion is that it be appropriately recognized or acknowledged *as* a criterion. This in its turn raises difficult questions. Who must recognize it as a criterion? Clearly, not everyone. Just the officials? But which officials? And must they not in their turn be recognized to be officials?

But I wish to focus on another difficulty. What exactly must the criterion be recognized as a criterion *for*. Now the obvious answer is that the criterion must be recognized as a criterion for being the law, i.e. it must be recognized that anything which meets the criterion is the law. Thus the definition will now take the form:

x is the law $=_{df}$ there is a criterion C satisfying certain conditions, including the condition that it is recognized that anything meeting the criterion is the law, and x meets the criterion.

But this now makes *being the law* a response-dependent concept! Part of what it is for

something to be the law is that it be recognized to be the law or, rather, that it meet a criterion which is recognized to be a criterion for being the law.

The alternative is to take the appropriate recognition of C to amount to something less than its being a criterion for being the law. There is going to be a necessary condition for being the law, call it 'being the putative law', and recognition of the criterion will amount to recognizing it as a criterion for being the putative law. As long as there is this recognition, and the other conditions are satisfied, then the putative law will thereby become the genuine law.

But the question now arises as to whether recognition of the criterion as a criterion for being the putative law could ever be enough. Imagine a society with a system of law. A group of individuals within the society then attempt to set up a rival system of law. Now can we not imagine that the rival system does just as well by 'objective' criteria and in terms of people's attitudes and beliefs in so far as they do not involve the concept of law. The only difference is that people do not regard the rival system *as* the law. And surely in this case it is not the law. But the alternative definition in terms of *the putative law* will not be able to predict this result; it will regard both systems as equally constitutive of the law.

It might be thought that in such a case there *is* a difference in attitude that does not involve the concept of law. For people will regard the one set of putative laws as binding and not those of the other set. But what is meant here by 'binding'? People may fear the consequences of breaking the 'law' in either case. They may, of course, not think of the rival 'laws' as being binding in the special way that is characteristic of the law. But it looks as if that is because they do not already regard them as the law; it is not what explains their failure to be the law. If they were suddenly to 'switch' and regard the rival set of laws *as* the law, then it looks as if their

sense of which laws were binding in this special way would go with, and be consequential upon, the switch in their view as to what is the law, and not the other way round.

To my mind, arguments such as these strongly suggest that many of our social concepts should be taken to be response dependent. The reality to which they apply is partly constituted by our conception of it as a social reality.

References

(These relate to the topics of response dependence and joint intention; and I have starred those that are especially relevant to the present paper. I have not given any references to the literature on the prisoners' dilemma or the concept of law.)

Alanen L., Heinamaa S. & Wallgreen T. (eds.) [1997] 'Commonality and Particularity in Ethics', New York: St Martin's Press.

Baier A. [1997] 'Doing Things with Others: The Mental Commons', in Alanen et al. [1997]

Balzer W and Tuomela R. [1997] 'A Fixed Point Approach to Collective Attitudes', in

Holmstrom-Hintikka, G and Tuomela R (eds) 'Contemporary Action Theory II: Social Action', Kluwer, 115-42.

Bardsley N., [2007] 'On Collective Intentions: Collective Intentions in Economics & Philosophy', Synthese 157 (2).

Belzer M. [1986] 'Intentional Social Action and We-Intentions', Analyse & Kritik 8, 86-95.

Blackburn S., [1985] 'Errors and the Phenomenology of Value', in Honderich (ed.) [1985]

- Blackburn S. [1993] 'Circles, Finks, Smells and Biconditionals', *Nous Suppl.* vol. 7.
- Bigelow J., Collins J., Pargetter R. [1990] 'Colouring in the World' *Mind* 99, vol. 394, 279-88
- Boghossian P., Velleman D. [1989] 'Color as a Secondary Property', *Mind* 98, vol. 389, 81-103.
- Boghossian P., [1989] 'The Rule-Following Considerations', *Mind* 98, 507-549
- *Bratman M. [1999] 'Faces of Intention', Cambridge, Cambridge Univ. Press
- Bratman M., [1992] 'Shared Cooperative Activity', *PR* 101, 327-41.
- Bratman M., [1993] 'Shared Intention', *Ethics* 104, 97-113
- Bratman M., [2007] 'Structures of Agency: Essays', OUP
- *Bratman M., [2007] 'Dynamics of Sociality', *Midwest Studies in Philosophy*, XXX, 'Shared Intentions and Collective Responsibility'.
- Bratman [2009] 'Modest Sociality & the Distinctiveness of Intention', *Phil. Studies* 144 (1), 149-65.
- Campbell J., 'A Simple View of Color', in Haldane & Wright [1991].
- Casati R.(ed.) [1998] 'European Review of Philosophy, Volume 3: Response-Dependence', Stanford: CSLI.
- Cohen P. R., Morgan J., Pollack M. E. [1990] 'Intentions in Communication', Cambridge: MIT Press.
- Cohen P. R., Levesque H. J., [1991] 'Teamwork', *Nous* 25.
- Copp D. [1995] 'Morality, Normativity, and Society', New York: Oxford Univ Press (esp. 15-151).
- Cuneo T. D., [2001] 'Are Moral Qualities Response-Dependent', *Nous* 35, no. 4, 569-591
- [moral]

Dancy J., [1986] 'Two Conceptions of Moral Realism', *Proc. of the Arist. Soc.*, suppl. vol. 60.

Danielson P. [1991] 'Closing the Compliance Dilemma: How It's Rational to be Moral in a Lamackian World' in 'Contractarianism and Rational Choice: Essays on David Gauthier's *Morals by Agreement*' (ed. P. Vallentyne), New York: Cambridge University Press.

Edwards J., [1992] 'Best Opinions and Intentional States', *Philosophical Quarterly* 42, 21-33.

Evans G., [1980] 'Things without the Mind' in Zak van Straaten (ed) 'Philosophical Subjects: Essays Presented to P. F. Strawson', 76-116.

*Gilbert M., [1989] 'On Social Facts', New York: Routledge, (chap. 7 on joint intention,)

*Gilbert M., [1990] 'Walking Together: A Paradigmatic Social Phenomenon', *Midwest Studies* 15, 1-14.

Gilbert, M. (2003). The structure of the social atom: Joint commitment as the foundation of human social behavior. In F. Schmitt (Ed.), *Socializing metaphysics* (pp. 39–64). Lanham, MD: Rowman and Littlefield.[for conditional commitments, p. 282?]

Gilbert M., [2006] 'Rationality in Collective Action', *Phil of Social Sciences* 36 (1): 3-17.

Gilbert M., [2008] 'Two Approaches to Shared Intention: an Essay in the Philosophy of Social Phenomenon', *Analyse & Kritik* 30, 2008.

Gilbert M., [2009] 'Shared Intentions and Personal Intentions', *Phil. Studies* 144 (1)

Gold N & Sugden R [2007] 'Collective Intentions and Team Agency', *JP* 104, 109-37.

Grosz B. J., Sidner C. L., [1990] 'Plans for Discourse', in Cohen et al. [1990]

Haldane J. & Wright C., [1991] 'Reality, Representation and Projection', Oxford: Oxford University Press.

- J. Haldane & C. Wright [1992] (eds) 'Realism and Reason', Oxford: OUP.
- Haukioja J., [2001] 'The Modal Status of Basic Equations', *Philosophical Studies* 104, no. 2, 115-122.
- Hobbes J. R. [1990] 'AI & Collective Intentionality', in Cohen et al. [1990]
- Holton R. [1992] 'Response-dependence and Infallibility', *Analysis* 52, no. 3, 180-184
- Holton R. [1993] 'Intention-Detecting', *Philosophical Quarterly* 43, 298-312.
- Honderich T., (ed.) [1985] 'Morality & Objectivity', London: Routledge
- Howard J. V. [1988] 'Cooperation in the Prisoner's Dilemma', *Theory and Decision* 24, 203-13.
- *Humberstone I. L. [1997] 'Two Types of Circularity', *Philosophy and Phenomenological Research* 57, 249-280
- Jackson F., Pettit P. [2002] 'Response Dependence without Tears', *Nous* supplement v. 12, 97-117 [concept]
- Johnston M. [1989] 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society*, supp. vol. 63, 139-174.
- Johnston M., [1998] 'Are Manifest Qualities Response Dependent?', *Monist* 81, no. 1, 3-43
- Johnston M. [1991] 'Explanation, Response-Dependence, and Judgement-Dependence', in P. Menzies (ed.) 'Response-dependent Concepts, Working Papers in Philosophy no. 1', Research School of Social Sciences, ANU.
- Johnston M. [1992a] 'Objectivity Refigured' in J. Haldane & C. Wright [1992].
- Johnston M., [1992b] 'How to Speak of the Colors', *Philosophical Studies* 68, 221-63.
- Keefe R. [2002] 'When Does Circularity Matter?', *Proceedings of the Aristotelian Society* 102.
- Kripke S., [1982] 'Wittgenstein on Rules and Private Language', Cambridge, Mass.: Harvard

University Press.

Koepsell D. and Moss L., [2003] 'John Searle's Views about Social Reality', Blackwell

Kutz C., [1996] 'Complicity: Collective Action in Ethics and Law', Ph. D. Thesis, Univ.

California, Berkeley.

Kutz C. [2000] 'Acting Together', Ph and Phen. Research 61, 1-31 [not so interesting].

LeBar M., [2005] 'Three Dogmas of Response-Dependence', Philosophical Studies 123 no. 3,

175-211

Levesque P. Cohen R., and Jose H. T. [1990] 'On Acting Together', Proc. of the National Conf.

on AI, Menlo Park: AAI Press/MIT Press, 94-99.

Levinson J., [2005] 'Aesthetic Properties', Aristotelian Society, suppl. vol. 79, 211-227

Lewis D., [1989] 'Dispositional Theories of Value', Proc. of Aristotelian Soc., suppl. vol. 63,

113-137.

McDowell J., [1978] 'Are Moral Requirements Hypothetical Imperatives', Proc. of the Arist.

Soc., suppl. vol. 52.

McDowell J., [1981] 'Non-cognitivism and Rule-Following', in Hotzman & Leich (eds.)

'Wittgenstein: to Follow a Rule', London: Routledge.

McDowell J. [1983] 'Aesthetic Value, Objectivity and the Fabric of the World', in 'Pleasure,

Preference and Value', ed. E. Schaper, Cambridge: Cambridge University Press, 1-16.

McDowell J., [1985] 'Values and Secondary Qualities', in 'Morality & Objectivity', ed. T.

Honderich, London: Routledge, 110-129, reprinted in Sayre-McCord [1988]

McDowell J., [1989] 'One Strand in the Private Language Argument', Grazer Philosophische

Studien 33-4, 285-3-3 [pain response dependent]

- McDowell J., [1992] 'Meaning and Intentionality in Wittgenstein's Later Philosophy', *Midwest Studies in Philosophy XVII*, 40-52.
- Macdonald G. & Pettit P. [1981] 'Semantics and Social Science', London: Routledge
- McGinn C., [1983] 'The Subjective View', Oxford: Oxford Univ. Press
- MacMahon C. [2005] 'Shared Agency and Rational Cooperation', *Nous* 39, 284-308
- Mantzavinos C. [ed] 'Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice', CUP
- Meggle G (eds) [2002] 'Social Facts and Collective Intentionality'
- Menzies P., Pettit P. [1993] 'Found: the Missing Explanation', *Analysis* 53, no. 2, 100-109.
- Miller A., [1989] 'An Objection to Wright's Treatment of Intention', *Analysis* 49, 169-173.
- Miller A. & Divers J., [1994] 'Best Opinion, Analytic Functionalism, and Intention-Detecting', *Philosophical Quarterly* 44, 239-415.
- Miller A., [1998] 'Rule-Following, Response-Dependence, and McDowell's Debate with Anti-Realism', in Casati (ed.) [1998]
- Miller A., McFarland D., [1998] 'Response Dependence without Reduction?', *Australasian Journal of Philosophy* 76, no. 3, 407-425.
- Miller S., [1992] 'Joint Action', *Phil Papers XXI*, 1-23.
- Miller S., [1995] 'Intentions, Ends and Joint Action', *Phil Papers* 24, 51-67.
- Miller S., [2001] 'Social Action: A Teleological Account', CUP
- Nagel T. [1979] 'Subjective and Objective' in 'Mortal Questions', Cambridge: Cambridge Univ. Press, 196-213.
- Peacocke C. [1989] 'What are Concepts?', *Midwest Studies in Philosophy* 14, 1-28.

- Peacocke C., [1995] 'A Study of Concepts', Cambridge: Cambridge University Press.
- *Peacocke C., [2005] 'Joint Attention: Its Nature, Reflexivity and Relation to Common Knowledge', in Ed Eilan et al. (eds) 'Joint Intention, Communication and Common Minds', OUP.
- Pettit P. [1991] 'The Common Mind', New York: Oxford Univ. Press.
- Pettit P. [1991] 'Realism and Response-Dependence', *Mind* 100, no. 399, 587-626
- Pettit P., [1998] 'Noumenalism and Response Dependence', *Monist* 81, no. 1, 112-132
- Pettit P., [1998] 'Terms, Things and Response-Dependence', in Casati [1998]
- Pettit P. & Schweikard d. [2006] 'Joint Actions and Group Agents', *Phil. Of The Social Sciences* 36, 18-39
- Powell M., [1998] 'Realism or Response Dependence', in Casati [1998]
- Rachel S. [2008] *Mind*, v. 117, 449-573
- Railton P., [1998] 'Red, Bitter, Good' in Casati [98]
- Rosen G. [94] 'Objectivity and Modern Idealism: What is the Question?', in M Michaelis & J. Hawthorne (eds) 'Philosophy in Mind', Dordrecht: Kluwer.
- Roth A. S., [2004] 'Shared Agency & Contralateral Commitments', *PR* 113, 359-410
- Sandu G. & Tuomela R., [1996] 'Joint Action and Group Action Made Precise', *Synthese* 105, 319-45.
- Sayre McCord [1988] 'Essays on Moral Realism', Cornell: Cornell Univ. Press.
- Searle J., [1983] 'Intentionality', Cambridge: Cambridge University Press.
- Searle J., [1990] 'Collective Intentions and Action', in Cohen [1990].
- Schelling T., [1960] 'The Strategy of Conflict', Cambridge: Harvard Univ. Press

- Schmid H. B. et al. [2008] 'Concepts of Sharedness: Essays on Collective Intentionality',
Philosophical Analysis 26.
- Sellars W. [1968] 'Science and Metaphysics', London: Routledge & Kegan Paul
(217ff. on joint intention)
- Shope R.K. [1978] 'The Conditional Fallacy in Contemporary Philosophy', Journal of
Philosophy 75, 397-413.
- Smith H., [?] 'Deriving Morality from Rationality', 240-2, esp. fn 18.
- Smith M., [1986a] 'Peacocke on Red and Red', Synthese 60, 559-76.
- Smith M., [1989] 'Dispositional Theories of Value', Proc. of Aristotelian Soc., suppl. vol. 63,
89-111.
- Smith M., [1995b] 'Internalism's Wheel', Ratio, 277-302.
- Smith P. & McCulloch G. [1987] 'Subjectivity and Color Vision', Proc. of the Arist. Soc., suppl.
vol. LX1
- Smith M. [1991] 'Objectivity and Moral Realism' in Haldane & Wright [1991]
- Smith M., [1998] 'Response-Dependence without Reduction', in Casati [1998]
- Stoutland F. [1997] 'Why are Philosophers of Action so Anti-Social' in Alanen et al. [1997].
- Sullivan P. M. [1994] 'Problems for a Construction of Meaning and Intention', Mind 103, no.
410, 147-168.
- Tollefson D., [???] 'Joint Action and Joint Attention', Phil of the Soc. Sciences
- Tsohatzidis S. (ed) [2007] 'Intentional Acts & Institutional Facts', Springer
- Tuomela R.
[1984] 'A Theory of Social Action'

- [1989] *Actions by Collectives*, *Phil Perspectives* 3, 471-96
- [1995] 'The Importance of Us: A Philosophical Study of Basic Social Notions', *Stanford Series in Philosophy*
- [2002] *The Philosophy of Social Practices* CUP
- [2003] 'The We-mode and the I-mode' in F. Schmitt 'Socializing Metaphysics'
- [2005a] 'We-Intentions Revisited', *Phil Studies* 125, 327-369.
- [2005b] 'The Philosophy of Sociality' OUP
- Tuomela R., Miller K., [1988] 'We-Intentions', *Philosophical Studies* 52, 367-89.
- Tuomela R. [1990] 'What Goals are Joint Goals', *Theory and Decision* 21, 1-20.
- Tuomela R., [1990] 'What are Goals and Joint Goals', *Theory and Decision* 28, 1-20.
- Tuomela R., [1991] 'We Will Do It: An Analysis of Group-Intentions', *Philosophy & Phenomenological Research* 51, 249-77.
- Tuomela R., [???] 'Joint Intention, We-Mode and I-Mode', *Midwest Studies*
- Velleman D., [1989] 'Epistemic Freedom', *Pacific Philosophical Quarterly*, 73-97.
- *Velleman D., [1997] 'How to Share an Intention', *Phil. & Phenomenological Research* 57, 29-50.
- Vermazen B., [1993] 'Objects of Intention', *Philosophical Studies* 70, 85-128 (defends intending that)
- Wedgwood R., [1998] 'The Essence of Response-Dependence', in Casati [1998]
- Wiggins D., [1987a] 'Needs, Values and Truth', Oxford: Blackwell.
- Wiggins D., [1987b] 'A Sensible Subjectivism' in Wiggins [1987a] (esp. section 5)
- Wiggins D., [1987c] 'Truth, Invention and Meaning' in Wiggins [1987a] and in Sayre McCord

[1988] (esp. section 6)

Wright D., [1987d] 'Further Reflections on the Sorites Paradox', *Philosophical Topics* 15, no. 1, esp. 274 -279.

Wright C. [1987e] 'On Making up One's Mind: Wittgenstein on Intention', in Weingartner and Schutz (eds.) 'Logic, Science and Epistemology', Vienna: Holder-Pichler-Tempsky, 391-404.

Wright C. [1988] 'Moral Values, Projection and Secondary Qualities', *Aristotelian Society suppl.*, vol. 62, 1-26.

Wright C., [1989a] 'Critical Notice of Colin McGinn 'Wittgenstein on Meaning'', *Mind* 98, 289-305.

Wright C., [1989b] 'Wittgenstein's Rule-following Considerations and the Central Project of Theoretical Linguistics', in A. George ed. 'Reflections on Chomsky', Oxford: Blackwell, 233-64

Wright C., [1991] 'Anti-realism: The Contemporary Debate - W(h)ither Now?' in Haldane & Wright [1991]

Wright C. [1992] 'Order of Determination, Response-Dependence, and the Euthyphro Contrast', in Haldane & Wright [1992]

Wright C., [1992] 'Truth and Objectivity', Cambridge, Mass: Harvard University Press.