The divided self: rationality and irrationality in political theory

Grahame Lock

There are good reasons for supposing that the mind – and therefore the person or the self – are in some substantial and important sense internally divided. In this article I shall discuss a number of the philosophical arguments put forward in recent years for the division of the mind, person or self, and attempt to show, with the aid of examples of different kinds, why this phenomenon, in some of its varied forms, is relevant to certain recent work in the field of social and political theory.

1. Arguments for the divided self

David Pears writes (Pears 1984: 67) that the different kinds and degrees of irrationality displayed by human beings seem to require us to adopt the apparently drastic hypothesis that a person in fact consists of (at least) two separate 'systems'. What kinds of irrationality would require us to draw this conclusion? Self-deception is one such phenomenon; akrasia (weakness of the will) is another. I shall here concentrate on the former, though saying a few words about the latter. When self-deception occurs, something is presumably deceiving something else. For how otherwise could a person P 'deceive himself'? Take the case, for instance, where P is a member of a religious sect to which he has devoted many years of militant activity. It is then discovered that the leader of this sect, to whom its members are personally devoted in some mystical sense, is an opportunistic fraud. The evidence to this effect is overwhelming. P is made aware of this evidence. Some 'part' of him (P_1) is rational enough to be convinced by it; another 'part' of him (P_2) resists. P_2 now adopts some tactic whose result is that P_1, against the overwhelming evidence, refuses to believe in his leader's evil-doing. P_1 is now deceived with respect to the truth. This means that P_1 must be in some substantive sense divided from P_2; for if it was not, then it could not be deceived. P_2, after all, itself knows the truth –

indeed its knowledge of this (dangerous) truth is what lies behind its successful attempt to deceive P_1. Curiously, however, we now have a construction in which that 'part' of the person who is better informed, or better in touch with reality (P_2), is precisely not this person's centre of consciousness and of initiation of action. Indeed, it is just because P_1 is this centre that *it* has to be deceived.

This is a familiar difficulty in the philosophical study of self-deception. I shall not pursue it further in this article, which has a different aim. The point here is simply to draw attention to the apparent need, in the light of such a phenomenon as 'self-deception', to admit a division of the mind, person or self.

A similar conclusion may be derived from an examination of akratic behaviour. A typical example of akratic behaviour is that of the man who, having decided that, all things considered he will be better off if he saves part of his salary, nevertheless spends it all when pay-day comes (perhaps telling himself each time that he will begin saving next month). Here again we must assume that some 'part' of him (P_{ii}) is indeed convinced that saving is the best policy, and would – if in command – put aside a reasonable sum. But another 'part' (P_i), which is in fact in command, gets the better of P_{ii}. Thus the person acts 'against his better judgement'. One of the curious aspects of this situation is that while we are obliged in such cases to assume that it is P_i which is 'in command' within the person made up of P_i and P_{ii} (for otherwise no akratic weakness would manifest itself), it is just this fact which leads us to conclude that the person in question is not 'in command of himself' – this indeed being one of the defining characteristics of akrasia. Again, I shall not here follow up the philosophical puzzle. I simply want to draw attention to the fact that some kinds of akratic behaviour also appear to require us to hypothesize a division of the mind etc.

A difficulty of a rather different kind lies in the identification of a person or self *through time*. Many people behave in an apparently irrational manner in the sense that they (for example) postpone an immediate pain – say a surgical operation – even when they have good reason to believe that this postponement will result in substantially greater pain at a – much – later date (either in severe illness, or in the necessity of a more serious operation). One explanation of this inclination, which would render it less irrational, is that people do not entirely identify with themselves at other periods in time than the present. If a person P at time t_1 (Pt_1) considers himself to be only remotely related to the person who he will become in thirty years' time (Pt_2), then Pt_1 seems to have good reason for not being very worried about a pain which Pt_2 will suffer. Analogously, Pt_2, when his ti-

me comes, ought to be able to understand Pt_1's original decision not to undergo an operation. This does not however necessarily mean that Pt_2 ought to approve of Pt_1's decision, for approval would require the addition of a moral premise endorsing egoism. After all, of two completely different people P and Q we should not naturally say that the mere fact that they are different people absolves the one from concern for the welfare of the other. And Pt_1 and Pt_2 are certainly not completely different people.

That many people do indeed reject complete identification with their own selves at periods other than the proximate present is also illustrated by a common attitude to the past. A person Pt_1 may well feel that he has very little in common with the person (Pt_0) who he was, say, thirty years previously. He may for example feel that he not longer ought to be held responsible for misdeeds of that earlier period. The mature conservative (like Richard Wagner) would surely resist any discrimination against himself on account of his earlier radical views, even though he – Wagner t_1 – is only too ready to condemn radicals of his own time. Of course he may be prepared to criticize his previous self, Wagner t_0. But that is precisely because he (Wagner t_1) *disidentifies* with Wagner t_0. Criminals who have been 'going straight' for some considerable time often especially resent being arrested and prosecuted for a crime carried out several years previously, and for similar reasons – that they are no longer criminals, that they are no longer 'the same persons' who committed the crimes in question.

Derek Parfit has argued (Parfit 1984) that there is indeed something valid in such arguments, though only in their metaphysical aspect. Our personal identity over time, he claims, involves no more than psychological continuity. And this in turn consists in no more than overlapping chains of psychological connectedness (that is to say, of experience-memories and suchlike). Now the relevant point here is that this kind of psychological connectedness is itself not a transitive relation. A person will typically remember today what he did yesterday in some reliable detail, just as yesterday he remembered what he did the day before, and so on. But his memory today of what happened to him on some arbitrary date ten, twenty, thirty or fifty years ago will not usually be either detailed or accurate – it may indeed be entirely absent. What links the person of today with himself at a much earlier date is thus not the strong relation of psychological connectedness itself, but the above-mentioned overlapping chains of that relation. Yet, claims Parfit, what matters to any one of us who cares about himself is not so much these overlapping chains as the strong relation of connectedness. But identity *is* a transitive relation. (The baby baptised Vladimir Ulyanov in 1870 is the same person as the graduate of the Simbirsk gymnasium who entered Kazan University in 1887; and this student is the same person as the man who, under the name of Lenin, became First People's Commissar in 1917; *therefore* the baby is identical with Lenin the commissar. And so on.) Psychological connectedness on the other hand is not transitive. (A person today has a memory of himself yesterday; yesterday he had a memory of himself the day before; and so on backwards for many years. But it does not follow that this person today has a memory of himself many years ago.) So psychological connectedness cannot yield the relation of personal identity.

Why does Parfit claim that what matters is therefore not personal identity but rather the strong relation of psychological connectedness? One of his arguments is drawn from a thought experiment. Suppose, he says, that technology had so advanced that human bodies could be scanned, blueprinted in the minutest detail and reconstructed, in the sense that an organic replica was made of the original body out of new matter. Suppose further that, whenever such a replica was made, the original (still living) body was destroyed at the very moment when the replica came into existence. Now the replica, whose brain would be qualitatively identical with that of the original, would 'remember' whatever the original remembered just prior to the moment of its destruction – that is to say, the relation of psychological connectedness (or a relation qualitatively indistuinguishable from it) would be preserved. But the deeper relation of identity would of course be destroyed. How much does this matter? Hardly at all, says Parfit. For it is not at all obvious that, as long as the relation of psychological connectedness is preserved, the original and the replica are *different* people. But how can they not be different if they are not quantitatively identical? The answer, I think, is not only that the relevant criteria are not the same in the two cases, but that which criterium one chooses to adopt depends on what one cares about. The replica, for instance, will have the same knowledge, abilities, motives and intentions as its original. If I am now destroyed, while writing this article, and replaced by a replica, then my replica will continue to write where I left off. Perhaps this has just happened. When I (now the replica) refer back to my previous arguments, this latter really is in a strong sense *my* argument. When I finish the article, the whole article will in a strong sense be *my* article. And I shall properly be praised or blamed for it. Why then should I worry? I may of course be aware that I am not identical, in the sense of quantitative identity, with the baby which became the philosopher of which I am a replica. But neither am I identical, in this sense, with that philosopher. I am simply qualitatively identical with him. Who could ask for anything more. Not the original, for he no longer exists. And not me either, for obvious reasons.

If this argument holds, then the notion of some kind of substratum or spiritual substance which would underlie the changes that any person undergoes in his life, and thus guarantee his identity through and despite those changes would have to be dropped. Just as (says Parfit) a nation is not something which exists separately, apart from its citizens and its territory, so a person is not something which exists separately, apart from the occurrence of a series of interrelated physical and mental events. Thus, he concludes, persons do of course exist, but only in the sense in which nations exist. Persons or selves are therefore 'not fundamental'.

I remarked earlier that Parfit would go along to some extent with the view, which many people hold, that they cannot properly be held responsible for actions which they committed many years previously, on the ground that they have since become a 'different person'; and similarly that they need not worry too much about their fate in the distant future, for by then they will have become a 'different person' too. (In Parfit's terminology, this is because, between Pt_0, Pt_1 and Pt_2, the relation of psychological connectedness is absent or very weak.) But I added that his sympathy with this point of view is restricted to its metaphysical aspect. For there is another feature to it, namely the *moral* aspect. And it is this which is of some importance to political theory.

If Parfit is right, what we call a person is in fact a set of successive selves, related one to another as a number of overlapping chains of psychological connectedness. Any such self – a person at time t_1 (Pt_1) – will care (we naturally suppose) about itself. But its duration, in any case in a strong sense of durable identity, is limited. Sooner or later it will (more or less) disappear, to be replaced by another, indirectly related but substantially different self – a self in turn doomed to disappear, and so on ad infinitum (or until death). Why then should any given self care about its distant ancestors or prospective heirs?

The answer to this question cannot of course (in the light of the above) be that these are all one and the same self. This answer must, according to Parfit, lie in the moral sphere. A self should care about its own distant ancestors and heirs for the same reason that it should care about other selves, about selves belonging to other persons. But this in turn means that the affective unity – that is, the identity through time – of an individual person is conditional on the existence of a temporal succession of *altruistic* selves. Why any self should take an altruistic rather than an egoistic standpoint is a question of a different order. Given however that it does so, there seems to be no good reason, on the theory in question, why its altruism should be directed towards its own distant ancestors and heirs rather than towards other selves to which it is related in other ways (in ordinary language: to those of other people now).

What does this rather abstract argument issue in? What it would do, it seems to me, is to undermine a central philosophical presupposition of both methodological and normative individualism – including the presently fashionable neo-liberal variety. For these doctrines do presuppose, what Parfit denies, that 'individual persons are fundamental'. Note by the way that a quite different ground is thus prepared for the critique of individualism than that which pits against it a so-called 'holistic' standpoint. What is here being suggested is not that something larger or broader than the individual 'is fundamental', but something more limited and essentially impermanent – such that to attempt to apply a self-interest theory (or theory of egoism) to it would border on the absurd.

But if we are thus obliged to assume some kind of altruism, why should we limit its range of application to the intra-personal, or even assign the intra-personal any privilege in this respect over the inter-personal?

Various more or less widespread assumptions in social theory are rendered vulnerable by these questions. Take for instance Jon Elster's heuristic principle of explanation, which commands: 'First assume that behaviour is both rational and self-interested...' (Elster 1985: 359). But on what grounds should we make such an asumption? If – as Parfit suggests[1] – intra-personal altruism on the part of the self is a condition of a coherent personhood, then what reason do we have to assume that altruism disappears as soon as inter-personal relations come into the picture?

2. Self-deception and the voter's illusion

There are other arguments for the division of the mind, person or self, and which have potential implications for social and political theory. I shall not attempt to discuss them all here, but will treat the application of two representative examples – one a weak theory of division, the other the strong, Freudian theory.

The weak theory is discussed by Quattrone and Tversky in their paper 'Self-deception and the voter's illusion' (Quattrone and Tversky 1986). They describe a particular form of self-deception which consists in the confusion of causal and diagnostic contingencies. Their characterization of the concept of self-deception is borrowed from Gur and Sackheim (Gur and Sackheim 1979), who adduce the following criteria of the phenomenon: first, that the individual in question simultaneously holds two contradictory beliefs; second, that he is unaware of the fact that he holds one of the two beliefs; and third, that this lack of awareness is motivated. These three criteria do seem to suggest some kind of division of the mind into

compartments, systems or whatever. For the belief (B_1) of which 'he' is unaware must at the same time be a belief of which some part of him is, or at least at some time was aware. It is just this fact – that something in him was aware not only of this belief, B_1, but also of the contradiction between B_1 and B_2 – which led to the suppression of B_1. These two beliefs need to be 'kept apart' in the mind, which means that they have to be kept in two different places, such that communication – in one direction – is impossible. It is this need which prompts us to think in terms of compartments and the like.

But we are not obliged, at least not in order to explain such simple cases of self-deception, to adopt a full-fledged Freudian or similar theory of the division of the mind. We might rather adopt a weaker theory, such as that proposed by Donald Davidson, according to whom the mind contains a number of 'semi-independent structures' or 'provinces' (Davidson 1982). These may, he says, be 'overlapping territories'. They must however posses at least some degree of autonomy. The reason for this is that forms of motivated irrationality like self-deception presuppose that there is causal action by one sphere of the mind on another. Thus mental causes that are not reasons for the mental states (here: beliefs) which they cause will account for irrational belief formation. In this way desire can cause a belief to be formed even though the belief which it causes is in contradiction with another belief for which the person in question has good grounds. (Under these circumstances wishful thinking becomes self-deception.) We shall see in fact that not only desires, but also beliefs can play such an (irrational) causal role.

The cases which interest us here are, as already announced, cases which involve a confusion between the causal and the diagnostic, where this confusion itself has a mental cause. Quattrone and Tversky first discuss the fascinating problem confronted by strict Calvinists, who believe in predestination. Thus they believe that the class of human beings can be divided – indeed, has already been divided by God – into those who are 'chosen' and therefore saved, and those who are 'not chosen' and therefore damned.[2] Good works have no causal influence on the matter, for the relevant decision has been made in advance (that is, in all eternity). But from this fact the conditional proposition does not follow that if a person performs no good works (or too few) he can still be saved. On the contrary, there is a universally true conditional law to the effect that, for any person, he/she will be saved only if (though not if and only if) he/she has performed good works.[3] It follows that if he/she has not performed good works, he/she will not be saved. Is the truth of these propositions not a good reason to act virtuously (even supposing that, in purely secular terms, to act virtuously

or to perform good works brings less pleasure, satisfaction and happiness than to sin)? Not at all. The reason is that the performance of good works stands in a diagnostic, not in a causal relation to salvation. One might compare the above laws with a medical law like: for any person, if he or she has small blisters on the chest and back, then he or she will develop a fever. The blisters are not the cause of the fever, but diagnostic of it. Both are symptoms of chickenpox.

In such a medical case it is not usually difficult to distinguish causal from diagnostic relations. But in other kinds of cases it may be. The situation of the devout Calvinist seems to be a case in point. Good works are (according to the doctrine) diagnostic of salvation. They do not guarantee salvation; but there is no salvation without them. The special difficulty which arises here is that it is *actions* and not a state of affairs (a blistered chest or whatever) which are diagnostic. And we normally consider actions to be voluntary, the product of free will. If the performance of good works is a necessary but not a sufficient condition of salvation, and if it is determined in advance whether any given individual is saved or damned, then it would seem to follow that free will can operate *at most* among the damned. For they can choose to sin or to be virtuous (though this latter choice will not help them). Those destined for salvation, on the other hand, will act virtuously, for they must do so. But this does not necessarily mean that they experience their actions as unfree. There may be nothing in the 'subjective experience' of the members of the two groups which allows them to determine whether their choice of virtue is or is not free. Thus even on the supposition that this distinction makes sense, it might still not be possible to make use of it in order to discover one's status. So neither the damned nor the redeemed know for sure what their fate will be.

Let us now suppose that some Calvinist finds himself in a situation where he is tempted to sin. He knows that if he commits even one such sinful deed, he is destined for hell. But he knows on the other hand that whether or not he commits this deed makes no causal difference to whether he ends up in heaven or hell. If to sin is more fun, he might therefore just as well sin. Yet to sin is to provide himself with conclusive evidence that he is damned. One reason for which he might refrain from sinning and choose virtue is thus to *avoid discovering* whether he is damned. This is not in itself irrational. But it is in psychological fact unlikely that motives for action in such cases do avoid a degree of irrationality, in the form of a kind of self-deception.

Quattrone and Tversky carried out experiments, whose aim was to determine whether there exists a tendency to confuse diagnostic and causal relations in cases where there is something of importance to the individual

at stake – for example, his health or – more directly relevant to the present theme – the election to office of a particular political party. Their results appear to show that, in a medical examination whose results were purely diagnostic of life expectancy, subjects tended to 'cheat' so as to produce outcomes indicating a long life. This cheating involves for obvious reasons a measure of self-deceptive, motivated irrational belief-formation.

The authors pose the question as to whether any comparable phenomenon can be observed among voters. There is a well-known paradox of voting to the effect that, since a single vote among millions either has an insignificant causal effect or an insignificant chance of having any causal effect at all (the one or the other depending on the electoral system), it is hardly worth the trouble for any individual to go and vote. But this is true of all individuals. If all drew this conclusion, and none (or very few) voted, this would have serious and probably undesired consequences. But however serious the consequence, they are never a good reason for any individual to vote, because his vote has, as we noted, insignificant causal weight.

Now there is an argument, often resorted to by political scientists, that individuals take the trouble to register their vote not because of its causal weight, but because of a sense of 'civic duty' or something of the kind. This is however, in the context, not a very persuasive argument. Indeed, as Elster remarks – following Barry (1979)[4] – political science seems to display in its recourse to such an explanation a kind of 'theoretical schizophrenia'. For to explain *that* individuals vote, appeal is made to their supposed civic spirit. This explanation having been established, *how* they vote is then frequently accounted for in terms of self-interest. Of course, the suspicion of theoretical schizophrenia on the part of the political scientist can be removed by projecting the division onto the voter himself, and supposing that the latter is in fact a split personality, split between a private and a civic self. But suppose that we are unwilling to make this attribution – that is, to believe that the civic self drives the voter to the polling booth, where the private self takes over then we may return to the Quattrone-Tversky hypothesis, namely that there is a confusion in the voter's mind between the status of his vote as diagnosis or cause. The authors argue that the situation may be the following. There exist (say) two political parties, A and B. The numbers of citizens supporting each of these two parties are roughly equal. Thus the outcome of an election will be determined, or at least strongly influenced, by the numbers of supporters of each party who actually turn up and vote on election day. The typical citizen whose thought process we shall analyze is a supporter of (say) party A. One of the characteristics of the typical citizen is that he regards himself as

typical – not of citizens in general, but of supporters of this party. So he reasons that if he goes to vote on election day, millions of other supporters of party A will do the same, because they think just like him. If on the other hand he fails to vote, then so will they, and for the same reason. Let us assume that he is right. Then there exists a true conditional proposition to the effect that 'if I (the citizen in question) register my vote, then the chance that party A will win the election is enormously greater'.

We may now suppose that this citizen 'forgets' – perhaps suppresses – the reasoning which led him to formulate this true proposition, remembering only the proposition itself. He then interprets this proposition in a psychologically 'natural' manner (that is, he makes a 'natural' mistake), taking it to express a *causal* relation. Of course, his process of reasoning may not be so clear or explicit. But this only makes it more comprehensible that the mistake occurs. In more realistic terms, therefore, we might guess that when election day comes, he does go to vote, with in the back of his mind the idea that, since it seems to be true that if he votes, then the chance of his party winning the election is much greater, his vote must in some sense be terribly important. Again, Quattrone and Tversky carried out experiments which seemed to show that many people do indeed tend to reason in this way and to act accordingly.

This would imply a degree of irrationality on the part of the typical voter – and probably of motivated irrationality, in fact of some kind of self-deception. And this in turn would raise all the questions concerning the divided mind or self which have been sketched out above – though, as already mentioned, this case would only require a weak theory of division.

3. Newcomb's problem

The reader who is inclined to think that he or she (in contrast to the typical voter) is perfectly capable of distinguishing between diagnostic and causal relations, and is immune from any confusion between them, might by the way like to test his immunity by a consideration of 'Newcomb's problem'.[5] This can be briefly presented as follows. There exists a being (human or non-human – its further characteristics do not matter) with prodigious and uncanny powers of prediction. This being hosts a television quiz show. In the show each contestant is confronted with two boxes, B1 and B2. Box B1 always contains a guaranteed $ 1,000. Box B2 contains either nothing or $ 1 million. A contestant may open either both boxes or B2 alone. Normally any contestant would choose to open both, since by doing so he can (or so it seems) only gain and cannot lose. But – and it

turns out to be a big but – contestants know the following to be the case. It is the host of the show who places the money in the boxes. This he does, in each case, *before* the contestant makes his choice. What he does is based on his private prediction as to whether a given contestant will open both boxes or only B2. In the first few broadcast shows, most contestants (for the reason mentioned above) choose to open both boxes. In every case where a contestant chooses to open both boxes, B2 is empty. A few contestants (for any reason) choose to open only B2. In every such case, B2 contains $ 1 million. In other words, the host's prediction is always accurate. In the next series of shows, some contestants reason as follows: 'Somehow the host can predict what any contestant will do, and acts accordingly, leaving B2 empty or putting in $ 1 million. There is excellent evidence of his ability. So he will also have predicted what I shall do. If I open both boxes, he will have predicted that I shall open both boxes. Therefore he will have left B2 empty, and I shall win only the contents of B1, namely $ 1,000. But if I open only B2, he will have predicted this choice, and will have put $ 1 million in the box. So I should open only B2'.

But other contestants ridicule this line of reasoning. They say: 'Whatever prediction the host has made in my case, at the moment when I have to make my choice he has already put the $ 1 million in B2 or left it empty. The box is sealed and guarded. Nothing I do can affect whether or not B2 contains the money. (My choice can *at most* be diagnostic of the presence of the money in the box.) So I can better choose to open both boxes, because in either case I then get $ 1,000 more than if I open only B 2'.

If you think that this second reaction is obviously the right one, then you will be strengthened in your conviction that you have no trouble in distinguishing between the diagnostic and the causal. But the question is whether you really would be prepared to act in line with your opinion. You may for instance be the ten thousand and first candidate to take part in the show. of the preceding ten thousand, five thousand chose to open both boxes; every one of these five thousand found box B2 empty. The other five thousand chose to open only box B2; every one of them found $ 1 million in it. You now stand before the boxes. You have of course your philosophical convictions. You are in particular convinced that there can be no good reason not to open both boxes – except of course the past evidence just referred to. But you are aware that this evidence (for all the reasons rehearsed above) is irrelevant to your decision. Yet you might – or at least, some people like you might – nevertheless be tempted to open only B 2. And you might even think up some justification for this choice – for example the meta-theoretical consideration that there is a chance that your philosophical convictions are ill-founded (that, contrary to your otherwise firm belief, there exists backwards causality[6], or something like that); that this chance has to be weighed up against the evidence provided by the first ten thousand candidates; and that the $ 1 million prize ought to play some role in the calculation. Generally speaking, however, it is unlikely that most of the candidates who choose, under the influence of the evidence, to open only box B2, will succeed in elaborating (or even attempt) such a philosophical defence of their decision. More likely, I think, is that their behaviour would have to be assessed in terms similar to those employed by Quattrone and Tversky, who suggest that, when action and outcome are consequences of a common antecedent cause, and even when they are in some sense known to be such, many people do tend to believe (or to act as if they believed) that by choosing to perform that action they have increased the probability of the outcome. And this given the knowledge involved – does, as I already noted in the more restricted context, seem to require them to engage in self-deception.

Now it might be thought that Newcomb's problem starts from such unlikely premises that, however interesting in its own right, it cannot have any application to everyday life, and in particular to political life. But this is not the case. For, as David Lewis (1986) has shown, the well-known phenomenon of the Prisoner's Dilemma is itself nothing but a (double) Newcomb Problem. And the Prisoner's Dilemma is of considerable importance to political analysis, for it concerns the answer which a rational agent must give, under specified circumstances, to the question 'Should I rat on my partner?', where the partner may be an accomplice, a colleague, a fellow-citizen and so on.[7]

Why is the Prisoner's Dilemma (PD) a double Newcomb Problem? We can, says Lewis, look at the question in the following way. We first 'translate' the PD by making the pay-offs rewards instead of punishments. Thus the worst punishment becomes a zero reward, and the least punishment the highest reward. Now a 'prisoner' who rats gets in any case the equivalent of opening the box with $ 1,000 in it. If prisoner no. 1 (P1) rats and prisoner no. 2 (P2) does not, then P1 gets an extra $ 1 million, thus $ 1,001,000 in total; but P2 gets nothing. The situation is symmetrical for P2. If both rat, each gets $ 1,000. If neither rats, each gets $ 1 million.

So it is always in P1's interest to rat, and always in P2's interest too. For (seen from P1's point of view), if P2 rats, then P1 is better off if he also rats ($ 1,000 > nothing). But if P2 does not rat, P1 is still better off by ratting ($ 1,001,000 >) $ 1 million). This holds symmetrically for P2. But this state of affairs is equivalent in the relevant respects to that described by Newcomb's problem. For the latter can be summed up as follows.

Each contestant can so choose as to get a guaranteed $ 1,000 – this he

does by opening both boxes. Whether or not he gets $1 million depends however on the prediction of his choice made by the host of the show. But this prediction has already been made at the moment when the contestant makes his choice. So whether he gets the $ 1 million in fact depends only on his choice plus a state of affairs (whether or not the $ 1 million has been placed in B2), on which state of affairs his choice has no causal effect. Similarly in the Prisoner's Dilemma, in our version, each 'prisoner' can so choose as to get a guaranteed $ 1,000; this he does by ratting. Whether or not he gets the $ 1 million depends however on the choice made by the other prisoner. On this factor, the choice made by the other, his own choice has no causal effect. The Newcomb Problem and the Prisoner's Dilemma are therefore in the relevant respects equivalent.

Newcomb's problem is, more essentially, characterized by the apparent fact that I will get the $ 1 million if and only if a certain (potentially) predictive process yields an outcome which would warrant the prediction that I will not take the guaranteed $ 1,000.[8] Now, suggests Lewis, such a (potentially) predictive process does not need to be made by some strange being who hosts a quiz show. We can for example make the predictions ourselves. One way is by simulation. I so to speak create a replica of (the relevant parts or aspects of) myself, put this replica in my anticipated predicament, and look and see whether the replica takes the guaranteed $ 1,000. Thus I can reason that 'I will get my $ 1 million if and only if my replica does not take the guaranteed $ 1,000'. (This is not a division of the self, only a projection.) What we now have is of course the diagnostic conditional. And this conditional does not warrant, on the premise that my replica does indeed not take the $ 1,000, the conclusion that I ought not to go for this $ 1,000. For my choice to go or not to go for the $ 1,000 (to open both boxes or to open only B2) can have no causal effect on whether B2 contains $ 1 million. Thus I always do better to go for the $ 1,000. Similarly, in the Prisoner's Dilemma, I always do better to rat.

Social scientists do not of course apply the 'replica approach' to the Prisoner's Dilemma. The reason is that the rational choice of each – to rat – is in this problem already given for all cases by the fact that no enforceable agreement can be made between myself and the other prisoner; my equilibrium strategy must then be to rat (the result being the equilibrium point at which both of us rat). Some non-egoistic social scientists, however, wishing to avoid this result, suggest that one might indeed try to apply the 'replica approach'.[9] My fellow-prisoner, they reason, will in certain circumstances predictably resemble me in his dispositions to act. If I were in a Prisoner's Dilemma situation, I would choose solidarity. But my fellow-prisoner is just like me, that is, he is my 'replica'. So he will choose

the same. Therefore it can be rational for me not to rat.

The counter-arguments are, first, that if I have good reason to believe that P2 will not rat, all the better for me, P1: I then rat, and profit. But should I not hesitate on account of his being my replica (and therefore I being his)? Not at all. For this relation is only diagnostic. Whether or not I change my decision at the last moment, and decide to rat after all, can have no causal effect on whether he is loyal. Besides, if he really is my replica, then the fact that I changed my mind at the last moment is only reason to believe that he did the same, and ratted too. In that case I would be a fool if I chose loyalty. Second, the argument surely cannot in fact be stated as it was stated above, namely in the formula that 'if I were in a Prisoner's Dilemma situation, I would choose solidarity'. For this would be to hold a position of unconditional solidarity, and this position is near to being incoherent. How can P1 act in solidarity with a P2 who himself rats? Solidarity, like marriage, requires a minimum of co-operation between two people. What is meant might rather be that 'I would choose solidarity if my fellow-prisoner did so too'; and that, since he is my replica, he will do so. Therefore it is rational for me to do so too. But this change in the formulation of the conditional substantially modifies the argument in a pertinent way. For my refusal to rat now follows not from an unconditional devotion to solidarity, but from a conditional preference for solidarity plus an empirical estimate of the *probability* that my fellow-prisoner will hold the same conditional preference *and* that he will also make the same empirical estimate of my probable choice. In other words, it will depend on the extent to which we are replicas of one another. And the extent to which this is true will depend on the extent to which we believe it to be true.

Could we each be justified in choosing this strategy, assuming conditional solidarity? Yes, if we have good enough reason to believe that we are indeed replicas of one another in the sense that we resemble each other *enough*. Note then that the resemblance need not be exact. In other words, ascription of probability need not be as high as 1 or 'certain'. This is fairly obvious in respect to the present case. For it is the basis on which we do in fact transform Prisoner's Dilemmas into non-Prisoner's Dilemma situations: by taking account of reasonably reliable probabilistic information as to the choice which to other will make, via information concerning his psychological dispositions.

Now the same point concerning the exactness or probabilistic reliability of the available information is applicable to the Newcomb Problem in its original formulation.[10] The two conflicting views as to the rationality of opening both boxes or only B2 may arise, if they arise at all, even when

the reliability of the predictions concerned has been quite poor. In the original formulation the pay-offs are $ 1,000 and $ 1 million. Thus the ratio of the two is 0.001. Abstracting from various marginally complicating factors, we may say that the conflicting views in question can arise whenever the expected value of taking the guaranteed $ 1,000 is less than that of turning it down (of opening only box B2); and that this is the case whenever the (average) estimated reliability is greater than (1 + 0.001)/2, that is, greater than 0.5005. This, as Lewis points out, is not a very demanding standard. It is often met in everyday-life cases.

Steven Brams (1976: 197-200) has argued that Newcomb's problem, in its converted, symmetrical form, is indeed applicable to the analysis of real-life political problems. Thus for instance, given a number of American politicians who must decide whether to enter the race for their party's nomination for the presidency, the actions of one or more of these actors – for example Robert Kennedy in 1968 – may be capable of analysis in terms comparable, *mutatis mutandis*, with the behaviour of Newcomb's Being:

> Kennedy ... agonized for several weeks over whether or not to enter the primaries in 1968. Not only did he face the question of whether Humphrey would enter the primaries after Johnson took himself out of the race just prior to the Wisconsin primary, but he also had to assess how he would do against McCarthy and possible stand-ins (i.e. favorite sons) for the Johnson administration if Humphrey did not run in the primaries. In this sense, Kennedy was forced to make predictions not only about Humphrey's actions but also about those of his surrogates who might run; given these predictions, Humphrey (and possible stand-ins) themselves faced choices, based in part on the intelligence (or clairvoyance) they attributed to Kennedy in being able to predict their own behavior.

So the Newcomb Problem is both relevant to ordinary situations and, in particular, to real-life political problems. Is it also relevant to the theme the divided self? We should recall in this connexion Quattrone and Tversky's claim that the confusion to which they draw attention between diagnostic and causal relations, of which Newcomb's problem is a striking example, is not just a confusion, but often involves a measure of self-deception. They argue that 'when people select actions to infer an auspicious antecedent cause, then [in order] to accept the inference as valid, they often render themselves unaware of the fact that they selected the action just in order to infer the cause' – that is, that they selected an action pruposefully chosen to produce a favourable diagnosis which would subsequently be easy to 'confuse' with a causal relation (Quattrone and Tversky 1986: 41). This supposition seems to require, as we saw, at least some rudimentary division of the self, mind, person or whatever.

4. Game theory and Freudian theory

But we can start out from another direction. Instead of arguing to such a rudimentary division, we can look at the consequences of presupposing a strong – that is, theoretically filled-in and elaborated – theory of the divided self, such as Freud's theory of the mind. We can then attempt to show, for instance again with regard to a representative example of what kind of consequences the application of such a theory might have.

This representative example is again the theory of games, as originally developed by John von Neumann and Oskar Morgenstern (1944), and since widely applied in all brances of the social and human sciences, including political science. Game theory studies the practical logic of interdependent decisions. It is in the first instance prescriptive: it aims to tell you what choice of tactic, for example, is the rational choice under given constraints (or that there is no single rational choice, and so on).

Note that game theory does not assert that people do in fact behave rationally, either always or mostly or even often. But its applicability requires the assumption of at least some substantial degree of rationality. Nor does it require the assumption of egoism. Such an assumption does avoid great complications in calculation, but it is not essential: one can try to quantify pay-offs for each player, taking his altruistic concern for others into account. The difficulty for game theory is not therefore that it cannot in principle take account of the existence and causal efficacy of either irrational or altruistic conduct. What would however be a problem for the theory is a certain problematization of the concepts of rationality/irrationality and of egoism/altruism.

It is well-known that Freudian theory problematizes the former pair of concepts by dividing the mind up into (sub-)systems with distinct characteristics and/or functions: Unconscious, Preconscious and Conscious on the one hand; id, ego and super-ego on the other. It is true that Freud's theory is not entirely clear on the question of how these (sub-)systems are related to one another. Pears suggests that, in a certain sense, the (sub-)systems (or some of them) can be understood by analogy with persons, that is to say, as rational agents (Pears 1982). If for example the psyche of some individual contains two (seriously) contradictory beliefs, together with the realization that these beliefs are contradictory, then either the irrationality must simply be eliminated, or – in the kinds of cases which interested Freud – one or another of the beliefs will be banished from the main system (for instance, from Consciousness) into a sub-system (for instance, the Unconscious).[11] If the individual in question has strong cognitive and/or affective reasons for holding on to both of these beliefs, then

this simultaneous process of banishment and conservation has obvious prima facie advantages. But who or what is the rational agent in this process?[12]

Les us take a case in point: Pears' story of the girl who, having good reason to believe that her lover was unfaithful, but 'self-deceived' to the effect – and therefore also believing – that he was not, avoided going to the café where she had strong evidence to believe that she would find him with another woman. And all this without being 'aware' that she was avoiding this particular café (or if so, why). In such a case we should, I suppose, have to say that although *she* (= her 'main system') was indeed self-deceived as to her lover's unfaithfulness, her sub-system – which kept her away from the café in question – was not. On the contrary. And it was this sub-system which was controlling the whole pattern of her behaviour in this matter. Yet, *being* her main system rather than her sub-system, (if she were her sub-system too, she would simply know that her lover was unfaithful) *she* can hardly be held responsible for her actions![13] This is a curious state of affairs. What for instance would be its consequences if generalized and applied to the relevant kinds of cases in the field of legal responsibility?

It is not clear that the game theorist, who works with an unproblematized notion of the rational or maximizing individual, can deal with the complications arising from the much more sophisticated Freudian notions of rationality and irrationality, in which the irrationality of one system or sub-system may be a causal consequence of a rational strategy set in motion by another, all within the same individual.

I cannot here go into all the complexities of the Freudian story. The main point is that Freud tackles problems like that of self-deception by distinguishing, within the same mind, a subject and an object of deception – roughly , the unconscious and the conscious mind respectively, both of which are agents. For instance, the unconscious mind, acting on the goal of preventing the conscious mind from believing something very disagreeable, causes it to misperceive reality – say, to misread a newspaper headline. Freud gives an example in his *Psychopathology of everyday life*. He had a son who, during the First World War, was drafted into the army and whose regiment was engaged in fighting in the region of Görz (Gorizia, a town in what is now western Yugoslavia). Picking up an evening paper one day, he read – or thought he read – the headline 'Der Friede von Görz' ('The peace of Gorizia'). But what in fact stood there was just the opposite: 'Die Feinde vor Görz' ('The enemy at the gates of Gorizia') (Freud 1901:113). It looks as if his unconscious mind had here adopted the tactic of causing him (that is, his conscious mind) to misread the printed words by substituting a typographically sufficiently similar and emotionally welcome variant.

But in such cases (and they are legion, of many types, and often occur in political contexts) *what* or *who* is it that can be said to be responsible for the action or omission in question? Ex hypothesi, not the conscious mind, but something like the unconscious mind. Yet if this is the case, our actions and choices seem often to be determined not by our conscious selves (with which however we usually identify ourselves) but by something – the Unconscious – which, again ex hypothesi, stands outside of our conscious control. But how, if this is the case, can we hold to the premise or assumption of rational choice theory in general, and game theory in particular, to the effect that the human individual can be unproblematically regarded as represented by a 'single, complete transitive preference ordering' or something of the kind?[14]

Michel Plon (1976) suggests that what a player would need to suppose, in order to take game theory seriously, is that he is indeed (identical with) his conscious mind; nor can he accept that his conscious mind is divided up into scheming and conflicting systems and sub-systems. That is to say (as Plon puts it): a person would have to refuse to recognize the existence of his own Unconscious. This refusal – of course it might well be an unconscious refusal – is what then makes it possible for him to aspire to *mastery* over his opponent in the game. For if he did possess a Freudian Unconscious he could not hope to master even that, let alone his antagonist! Freud notoriously remarked that 'the ego is not master [even] in its own house', just because of its place in the topographical structure of the mind, and because of its often conflictual relation with the other (sub-)systems.

What does all this mean in practice? Does it mean that the assumptions of game theory are so far removed from the reality of the human psyche that this theory has no useful application? This is Plon's view. But it is perhaps not a complete view. What, after all, explains the apparent success of game theory in the field of social and politicial theory, if it is in the above sense an artifical theory? I shall hazard a brief answer to this question. Its success in application is, I suspect, in part real, and derives from *the artificiality of certain aspects of the social and political process*.

Game theory is thus applicable in the analysis of a socio-political situation to the extent that this situation can be described in *formal* terms – that is to say, when (only) official motives count, and where the players act in a role or function rather than in a personal capacity. This is often so. But one should not too quickly assume, in any given case, that such a situation holds. For instance: workers in their role of trade unionists always (we may suppose) want higher wages (other things being equal); consumers always spend their money in accordance with the appropriate 'equilibrium combination'; policemen always try to preserve law and order. But, in

all three cases (and of course in many others which might be mentioned) they often seem to behave in fact in a manner which fails to fit with the expectations following from these roles. This is of course not surprising. But it does show that the applicability of game theory may be even more restricted than might at second sight – that is, according to my suggestion above – appear to be the case.

5. Dworkin and Rawls: the private and the public self

The distinction which this suggestion makes use of may by the way have something to do, *mutatis mutandis*, with another distinction, that drawn by Ronald Dworkin between the private and the public self – the point in this case being that political philosophy may be justified in attributing particular characteristics to human individuals (say, that they posess free will, and can therefore choose freely between available courses of political action; that they act from certain *publicly* comprehensible motives, and so on) all this *even if* a 'deeper' investigation of a psychological or philosophical kind reveals that these assumptions are in fact false. The legitimization of this approach might in turn be – as John Rawls suggests – that what we attribute to people when we are talking politics are characteristics the possibility of whose possession and display is given by the political system itself. For example, the definition of individual freedom relevant to the description of the possibilities of individual political behaviour is not a metaphysical one, concerning the reality or unreality of free will, and so on, but a legal-constitutional (and perhaps sociological) account of the existence or non-existence of an effective right to freedom of thought, speech, assembly etc. Similarity, a certain kind of philosophical investigation of motive would be largely irrelevant (so the argument goes) to the study of the kinds of motives which lead individual voters to choose this or that political party.

The suggestion of Dworkin is therefore that, in a sense, political action is in itself essentially 'artificial' (though he does not use this term). But, whether or not this idea really does lie behind his arguments, I do not think that it is *generally* true. This means that it does not take us far in solving the main problem at hand, that of the applicability of game theory. So, returning to that problem, we must hold to the principal point, namely that many of the situations of intersubjective co-operation and conflict which game theory would like to treat cannot be described in the necessary exclusively formal or public terms; and most cannot be *wholly* described in these terms. If that is so, then – on the supposition of the strong

theory of the division of the self which psycho-analytic theory implies – game theory must often fail. For this reason too, the thesis of the divided self is of importance to political theory.

Notes

1. This is how I understand the implication of some of his central claims in Parfit (1984).

2. St. Paul in Romans 8:30: 'Moreover whom He did predestinate, He also called'.

3. Thus I am supposing that Calvinist doctrine holds that, for the class of human beings, $(x)(Sx \rightarrow Vx)$, but not $(x)(Sx \leftrightarrow Vx)$, because not $(x)(Vx \rightarrow Sx)$, where 'S' stands for 'being saved' and 'V' for 'being virtuous'.

4. Cited in Elster (1986), p. 26.

5. See R. Nozick (1969).

6. Cf. J.L. Mackie (1985).

7. On Prisoner's Dilemmas and their place in the general scheme of decision theory, see the article by Huib Pellikaan in this number of *Acta Politica*.

8. See Lewis (1986), p. 301.

9. This suggestion has for instance been made to me by my colleague Mark Bovens. His remark prompted the remarks which follow.

10. For details, see Lewis (1986), p. 302.

11. Cf. Pears (1984), chapter IV.

12. For a truly wondrously complicated story of unconscious strategy, see Freud's account of the 'Rat Man', in Freud (1909).

13. For a more detailed presentation of this and other connected problems see Lock (1987).

14. Cf. Steedman and Krause (1986). But their critique of the assumptions of decision theory is less radical, though better elaborated, than what is here suggested. See also von Neumann and Morgenstern (1944), who qualify the above-mentioned principle. But their qualifications do not bear on the point at issue.

Bibliography

Barry, Brian, *Economists, sociologists and democracy*. Chicago University Press, Chicago 1979.

Brams, Steven J., *Paradoxes in politics*. The Free Press, New York 1976.

Davidson, Donald, Paradoxes of Irrationality. In: Richard Wollheim and James Hopkins (ed.), *Philosophical essays on Freud*. Cambridge University Press, Cambridge 1982.

Elster, Jon, *Making sense of Marx*. Cambridge University Press, Cambridge 1985.

Elster, Jon, (ed.), *The multiple self*. Cambridge Univ. Press, Cambridge 1986.

Freud, Sigmund, 1901, *The psychopathology of everyday life*. New translation in Sigmund Freud, *Standard edition*, vol. VI. Hogarth Press, London 1960.
Freud, Sigmund, 1909, *Notes upon a case of obsessional neurosis*. Translation in Sigmund Freud, *Standard edition*, vol. X. Hogarth Press, London 1955.
Gur, R.C., and H.A. Sackheim , Self-deception: a concept in search of a phenomenon. *Journal of Personality and Social Psychology*, 37, 1979.
Lewis, David, *Philosophical papers*, vol. II. Oxford University Press, New York 1986.
Lock, Grahame, Filosofie en psychoanalyse: de verdeelde geest. *Thauma* (Leiden), 3, 1987, 11, March 1987.
Mackie, J.L., Newcomb's paradox and the direction of causality. In: J.L. Mackie, *Logic and knowledge*. Clarendon Press, Oxford 1985.
Neumann, John von, and Oskar Morgenstern, *Theory of games and economic behavior*. John Wiley, New York 1944.
Nozick, Robert, Newcomb's problem and two principles of choice. In: N. Rescher, (ed.), *Essays in honor of Carl G. Hempel*. Reidel Dordrecht 1969.
Parfit, Derek, *Reasons and persons*. Clarendon Press, Oxford 1984.
Pears, David, Motivated irrationality, Freudian theory and cognitive dissonance. In: Richard Wollheim and James Hopkins (ed.), *Philosophical essays on Freud*. Cambridge University Press, Cambridge 1982.
Pears, David, *Motivated irrationality*. Clarendon Press, Oxford 1984.
Plon, Michel, *La théorie des jeux: une politique imaginaire*. François Maspero, Paris 1984.
Quattrone, George A., and Amos Tversky, Self-deception and the Voter's Illusion. In: Jon Elster (ed.), *The multiple self*. Cambridge University Press, Cambridge 1986.
Steedman, Ian, and Ulrich Krause, Goethe's *Faust*. Arrow's possibility theorem and the individual decision-taker. In: Jon Elster (ed.), *The multiple self*. Cambridge University Press, Cambridge 1986.

Boekbesprekingen

W.J. Dercksen, **Industrialisatiepolitiek rondom de jaren vijftig. Een sociologisch-economische beleidsstudie**, Van Gorcum, Assen-Maastricht 1986.

De Nederlandse regering stond na de periode van het herstelbeleid (1945-1949) voor de opgave de economische ontwikkeling krachtig te stimuleren. De Marshall-hulp bood daartoe een gunstige uitvalsbasis, maar de economie moest op eigen kracht verder. De vraag was op welke doelstellingen het macro-economisch beleid gericht zou worden. Zou gekozen worden voor een actief, planmatig en sturend overheidsbeleid, zoals met name de sociaal-democraten in de periode vlak vóór de Tweede Wereldoorlog met het planisme beoogden, of zou het sociaal-economisch beleid voorwaarden moeten scheppen om een zo gunstig mogelijke positie voor de vrije markteconomie te bereiken?

Na enkele jaren van wederopbouw werd vanaf 1949 via de eerste industrialisatienota een industrialisatiebeleid in gang gezet dat de sociaal-economische ontwikkeling van ons land tot in de jaren zestig zou toonzetten. Achteraf bleek dit, gezien het bereikte welvaartsniveau, een succesformule te zijn. De voorwaarden werden geschapen waaronder particuliere ondernemingen in staat waren te produceren en te exporteren. Het doel van dit beleid was door industrialisatie de beroepsbevolking voldoende werkgelegenheid te bieden, de uitstoot van arbeidskrachten uit de landbouw op te vangen en via export de handels- en betalingsbalans te verbeteren.

In dit proefschrift worden de verschillende facetten van dit beleid bestudeerd: de loonpolitiek, het begeleidend prijsbeleid, de fiscale politiek (vervroegde afschrijving, investeringsaftrek), de financiering van industriële investeringen, de bevordering van de arbeidsproduktiviteit, het technologiebeleid, de stimulering van het arbeidsaanbod via het technisch onderwijs en ten slotte het regionale industrialisatiebeleid. Dercksen stelt dat de industrialisatiepolitiek destijds een gecoördineerd project van overheidsbeleid 'avant la lettre' was. De genoemde aspecten van dit gecoördineerde project behandelt hij systematisch door de gehanteerde beleidsinstrumenten te beschrijven en de resultaten (in zoverre die waarneembaar en rechtstreeks aan het gevoerde beleid zijn toe te schrijven) te inventariseren. De directe koppeling tussen beoogde beleidsdoelstellingen en beleidsuitkomsten geeft deze