Plato, Hare and Davidson on Akrasia

C. C. W. Taylor

Mind, New Series, Vol. 89, No. 356 (Oct., 1980), 499-518.

Stable URL:
http://links.jstor.org/sici?sici=0026-4423%28198010%292%3A89%3A356%3C499%3APHADOA%3E2.0.CO%3B2-%23

Mind is currently published by Oxford University Press.

Mind (1980) Vol. LXXXIX, 499–518

Plato, Hare and Davidson on Akrasia[1]

C. C. W. TAYLOR

My starting point is Donald Davidson's discussion of the problem of akrasia in 'How is Weakness of the Will Possible?' (in J. Feinberg (ed.), *Moral Concepts*, Oxford, 1969). In Davidson's formulation, the problem consists in the fact that three propositions, which taken individually are each persuasive (indeed he goes so far as to say that they seem self-evident, p. 95), appear to form an inconsistent triad. These propositions are (using Davidson's formulation and numbering):

P1. If an agent wants to do x more than he wants to do y and he believes himself free to do either x or y, then he will intentionally do x if he does either x or y intentionally;

P2. If an agent judges that it would be better to do x than to do y, then he wants to do x more than he wants to do y;

P3. There are incontinent actions.

Davidson's solution of the problem is to argue that these three propositions are not in fact inconsistent. Specifically, the incontinent agent (i.e. the man who, given a choice between doing x intentionally and doing y intentionally, does y though his judgement is in favour of doing x) does not judge that it would be better to do x than to do y. Rather, he judges that, all things considered, it would be better to do x than to do y, which, Davidson argues, is compatible with its not being the case that he judges that it would be better to do x than to do y. Thus while from P1 and P2 we may derive by syllogism the following proposition (true in Davidson's view),

P4. If an agent judges that it would be better to do x than to do y and he believes himself free to do either x or y then he will intentionally do x if he does either x or y intentionally,

P4 does not contradict P3. For given Davidson's analysis of what it is that the incontinent man judges, P3 should be formulated as

1 I am grateful to Jennifer Hornsby for her helpful comments.

P3′. Some agent judges that, all things considered, it would be better to do x than to do y, and believes himself free to do either x or y, and does y intentionally.

And given that there is no incompatibility between 'A judges that, all things considered, it would be better to do x than to do y' and 'It is not the case that A judges that it would be better to do x than to do y', P4 and P3′ can both be true.

In this paper I shall not directly discuss Davidson's proposed solution. I merely remark that, even if his argument is correct, it does not seem to succeed in solving the problem. For if we have adequate grounds for accepting P2 as true, we surely have equally good grounds for accepting the following:

P2′. If an agent judges that, all things considered, it would be better to do x than to do y, then he wants to do x more than he wants to do y.

And given Davidson's analysis of the incontinent man's judgement, as expressed in P3′, we now have the triad P1, P2′, P3′, which is inconsistent.

Davidson does not give reasons for believing P2; in fact his view of it is somewhat obscure, since he says both that P1–3 seem self-evident and that it is easy to doubt whether they are true in their form as stated (particularly P1 and P2: paraphrasing pp. 95–96). I interpret him as meaning that, while there may be questions as to the precisely correct wording, the general principles expressed in P1–P3 are certainly true (and would perhaps if precisely formulated be self-evident). Against this unargued acceptance (in principle) of P2, it is a sufficient *ad hominem* argument to assert that P2′ has an equal claim to unargued acceptance. More seriously, those who incline towards accepting P2 do so on the general ground that there is some necessary connection between what an agent judges it better to do and what he wants more to do. Whatever the nature of that connection, it must hold as well for what an agent judges it better to do, all things considered, as for what he judges it better to do, *sans phrase*. For if practical judgements give expression to desires, then the judgement 'All things considered, it is better to do x than y' is a paradigm instance of a judgement fitted for that role. Or, on the other hand, if practical judgements are conceived of as generating the desire for what is recognised in the judgement,

how could 'All things considered, it is better to do x than to do y' fail to exercise that mysterious force? After all, Davidson regards that proposition as expressing the agent's judgement in the light of all available considerations (pp. 110–111); and if that kind of judgement lacks the capacity to generate desire, then how could any judgement have that capacity?

But Davidson's approach to the problem is not merely (a) obscure and (b) likely to be unsuccessful even if capable of clarification. It is also misconceived. The solution to the problem lies in the approach rejected by Davidson at the beginning of his paper, viz. that of giving up one or more of the principles P1–P3. In the remainder of this paper I shall (i) argue that P2 is false, (ii) show why some philosophers have accepted P2 or some similar principle, (iii) attempt to show what was wrong with their reasons for accepting that principle, and (iv) enunciate and defend a principle which is sufficiently close to P2 to be confused with it, but which is (a) distinct from P2, (b) true and (c) consistent with P1 and P3.

I

Since P2 is a universal proposition, it will be shown to be false by a counter-example. Any typical instance of action against one's better judgement provides such a counter-example. For example, consider a man who is tempted to sleep with someone else's wife, has the opportunity to do so, but judges that all things considered, it would be better not to sleep with her than to do so. Then, giving in to temptation, and acting against his better judgement, he sleeps with her. It seems to me plainly true of this man both that he judges it better not to sleep with the woman than to sleep with her, and that it is not the case that he wants not to sleep with her more than he wants to do so. (In fact, though this is irrelevant to the status of the case as a counter-example, he wants to sleep with her more than he wants not to do so.) We have thus satisfied the requirement of producing a counter-example which shows P2 to be false.

But, it will be objected, this is to go too fast. No one who, like Davidson, accepts both P2 and P3 will agree that this case is a counter-example to P2, since this sort of case is just the kind of case which P3 admits. Such a philosopher, then, maintains that the case of the akratic man described above is not a case of an

agent who both judges that doing x is better than doing y and who does not want to do x more than he wants to do y. That is to say, he maintains either that the akratic man does not judge that doing x is better than doing y, or that he wants to do x more than he wants to do y. Let us consider these positions in turn.

The first position is a highly paradoxical one for someone to hold who also believes that there are akratic actions, for the phenomenon of akrasia is nothing other than that of action *against one's better judgement*. Davidson's actual position is, indeed, a version of this. The akratic man does not judge that doing x is better than doing y: rather he judges that, all things considered, doing x is better than doing y. But, as I have already pointed out, this provides no defence against the counter-example, since Davidson is committed to P2',

If an agent judges that, all things considered, it would be better to do x than to do y, then he wants to do x more than he wants to do y,

and the instance of akrasia described above will serve as a counter-example to this principle too. If Davidson wishes, as he must, to resist the claim that this is a counter-example, he cannot now maintain that the agent does not judge that, all things considered, it would be better to do x than to do y, since on his own hypothesis that is just what the akratic agent does judge. Hence he must maintain that he wants to do x more than he wants to do y. I.e. this attempted defence of the first position (the only one with any appearance of plausibility), has turned out to depend on the second position, which I shall now consider.

According to this position, the man in my example, who succumbs to the temptation of adultery while judging it better not to do so, wants not to sleep with the woman more than he want to sleep with her. It is important to keep in mind first of all that the desires in question are desires relative to one particular action, viz., the desire to sleep with (and the desire to refrain from sleeping with) *this* woman on *this* occasion. For it may well be the case that the man wants (dispositionally, and as a long-term aim) to be the sort of man who avoids adultery more than he wants (dispositionally, and as a long-term aim) to be a Don Juan, this being manifested by his general pursuit of the former aim and his general eschewing of the latter (allowing for occasional lapses). But this is simply irrelevant to whether he wants not to sleep

with *her now* more than he wants to sleep with *her now*. Again, to avoid any possible irrelevance, it is not in dispute that the akratic man does want both to commit adultery with her now and to avoid it: characteristically, the akratic situation is one of conflicting desires. The question, to repeat at tedious length, is which of these particular things he wants more than the other.

What then decides the question whether *A* wants to do a particular thing *x* more than he wants to do a particular thing *y*? One way of answering this question is by observing which of the two *A* actually does: there is one use of the expression '. . . wants . . . more than . . .' according to which, given the choice between doing *x* and doing *y* but not the possibility of doing both, *A* wants to do *x* more than he wants to do *y*, iff *A* does *x* and does not do *y*. In this use '. . . wants . . . more than . . .' is equivalent to '. . . prefers . . . to . . .'; the question of what *A* prefers to do is *settled* by consideration of what he does, irrespective of whether he acts eagerly or reluctantly. The only requirements are that *A* should act freely and intentionally. Davidson's P1 gives a weaker version of this criterion. Given either version, if *A* intentionally did *y* in preference to *x* it follows that he did *not* want to do *x* more than he did *y*.

Since Davidson presents P1 as a principle which is true generally, and not merely given a particular use of the expression '. . . wants . . . more than . . .' he leaves himself defenceless against the counter-example I have given. But more detailed consideration of the phenomena of wanting shows that Davidson has a defence, albeit not one which can ultimately save him. By contrast with the use mentioned above, there is another use of the expression '. . . wants . . . more than . . .' where the agent's choice of *x* in preference to *y* is neither necessary nor sufficient for the truth of '*A* wants *x* more than *y*'. Given this use, what is decisive is the agent's attitudes to *x* and *y*, e.g. whether he is eager to do one, reluctant to do the other, whether the thought of one or the other fills him with enthusiasm etc. With this conception of wanting in mind (which we might call wanting as inclination to distinguish it from wanting as preferential choice) there is no contradiction in such a statement as 'Of course I wanted to go fishing much more than I wanted to go to the meeting, but all the same I chose to go to the meeting, much against my inclinations'. Davidson's P1 is false given this conception of wanting and may therefore be taken as restricted in its application to wanting as preferential choice. This would then allow Davidson to say that from the fact that *A* intentionally

did *y* in preference to *x* it does not follow that he wanted to do *y* more than he wanted to do *x*, provided that 'wanted to do *y* more than he wanted to do *x*' is here understood as 'had a stronger inclination to do *y* than he had to do *x*'. Hence Davidson could still hang on to his claim that, contrary to my proposed counter-example, *A* in fact wanted to do *x* more than he wanted to do *y*.

But this defence leaves Davidson in a worse position than ever, for his claim that the agent in the example wants to avoid adultery with her now more than he wants to commit it has now to be understood as the claim that he has a stronger inclination to avoid adultery with her now than to commit it. Yet a situation of conflict between desire for a long-term good and desire for immediate pleasure is a paradigm instance of conflict between reasoned desire on the one hand and inclination on the other. The fact that the agent found the idea of adultery so intensely attractive as to have great difficulty in adhering to his judgement about what it was best to do is a sufficient condition of the truth of the proposition that his inclination to commit adultery was stronger than any inclination he may have had to refrain. Had he none the less resisted this inclination, it would not have ceased to be true that his inclination toward adultery was stronger. We should then have had a situation where the question 'Did *A* want to commit adultery more than he wanted to refrain from it?' has no single answer. Given the conception of wanting as preferential choice, he wanted to avoid it more than he wanted to commit it, but given the conception of wanting as inclination he wanted to commit it more than he wanted to avoid it. But in the situation as described, the question has the same answer on either conception of wanting; whether we are asking about the agent's preferential choice or his inclinations, in either case he wanted to commit adultery more than he wanted to avoid it.

On either conception of wanting, then, the counter-example to P2 stands. A defender of P2 has now only one recourse, viz. to claim that there is some further conception of wanting such that an agent who both feels a stronger inclination to do *y* than to do *x* and actually does *y* in preference to *x* may still truly be said to want to do *x* more than he wants to do *y*. This would be a counsel of desperation. I at least am clear that I have no such conception, and I doubt whether anyone has. I can understand the proposition '*A* wants *x* more than he wants *y*' in terms of *A*'s preferential choices, or in terms of his inclinations, but otherwise I have no

idea what it might mean. But suppose that this is merely conceptual inadequacy on my part. Even if we were to grant that there is some as yet unelucidated conception of wanting which allows us to say that our akratic man wanted to avoid adultery more than he wanted to commit it, that would not save the Davidsonian position, defined as the position that P1–P3 are all true. For it is now admitted that, on this new (supposed) conception of wanting, *A* may want to do *x* more than he wants to do *y*, and believe himself free to do either *x* or *y* and do *y* intentionally and not do *x*. I.e. the defender of the Davidsonian position can save P2 only at the cost, not only of introducing an as yet unexplained conception of wanting, but also of giving up P1. He would do better to give up P2 and have done with it.

I have now shown by means of a counter-example that on an ordinary understanding of what it is to want *x* more than *y* P2 is false. It has also emerged incidentally that P1 holds only where *A* wants *x* more than *y* in the sense that *A* chooses *x* in preference to *y*, and that it too is false where *A* wants *x* more than *y* in the sense that *A* has a stronger inclination for *x* than he has for *y*. In the next sections I shall consider the reasons why two other philosophers, Plato and R. M. Hare, have accepted P2 or a related principle.

II

For evidence that Plato accepted a version of P2 we may turn to *Prot.* 358 b–e. Here Plato says

(i) If pleasure is the good, then if anyone judges that *x* is better than *y*, and he is able to do either *x* or *y*, he does *x*;

(ii) If *A* judges that *x* is good and *y* bad, he is not willing to go for *y* in preference to *x*;

(iii) If *A* judges that (a) *x* is bad, (b) *y* is bad and (c) *y* is worse than *x*, and if *A* must choose either *x* or *y*, he will choose *x*.

Plato's explicit words fall short of the full generality of P2, since (i), which asserts that an agent will always choose what he takes to be the better of two alternatives, is dependent on the assumption that pleasure is the good, whereas (ii) and (iii), which are apparently independent of that assumption, make the lesser claims that an agent will always prefer anything good to anything

bad, and that he will always prefer the lesser of two evils. In fact, I doubt whether the restriction is of any significance. For, firstly, in the *Protagoras* at any rate, Plato argues that pleasure is in fact the good,[1] which, together with the conditional expressed in (i) above, gives the generalisation that if anyone judges x better than y, and believes himself able to do x or y, he will do x, which amounts to P2 (given the conception of wanting as preferential choice). Secondly, the discussion of the mechanics of choice at 356 a–c shows that Plato sees the same principles applying to the choice between two good alternatives as to those between two bad and a bad and good, which are the cases explicitly dealt with by (ii) and (iii). I therefore interpret Plato as holding that it is as foreign to human nature to fail to choose the better of two alternatives as it is to choose a bad alternative when a good is available, or to choose the greater evil when only bad alternatives are available.

While this thesis is asserted rather than argued for in the *Protagoras* it is fairly clear that it rests ultimately on the two assumptions (i) that for an agent to judge one alternative better than another is for him to judge that alternative more in his interest than the other, (ii) every agent always does what he thinks will best promote his interest. (i) The ultimately self-interested ground of choice of alternatives as better or worse appears clearly e.g. in the argument with Polus in the *Gorgias*. Polus maintains that while it is more shameful or disgraceful to treat someone else unjustly than to be treated unjustly it is worse to be treated unjustly (474c), Socrates maintaining, on the contrary, that acting unjustly is both more shameful and worse. The immediately preceding discussion has made it clear that what is at issue is whether acting unjustly tends to bring it about that the agent has a completely satisfactory life, as Polus maintains by appeal to the example of successful tyrants, or whether it tends to make the agent wretched and unfortunate, i.e. to give him an unsatisfactory life. In order to provide the necessary proof that the unjust man will have an unsatisfactory life Socrates establishes the following principle (474d–475b) about the opposition *kalon* (beautiful, creditable, praiseworthy)—*aischron* (ugly, shameful, discreditable):

> (x) (y) (If x is more *kalon* than y then x is pleasanter than y or x is more useful than y and if x is more *aischron*

1 For a defence of this controversial view see C. C. W. Taylor, *Plato, Protagoras*, Oxford, 1976, pp. 208–209.

than *y* then *x* is more unpleasant than *y* or *x* is *worse* than *y*).

It is clear that the intention of this principle is to establish creditableness/discreditableness as a function of the two factors, pleasure/pain and usefulness/harmfulness. The positive side of the latter opposition is designated by the term translated 'usefulness' (*ōphelia*), the negative by the term 'bad' (*kakon*), which is universally the opposite of 'good' (*agathon*). It is clear that in this argument the pairs 'good-bad' and 'useful-harmful' function as interchangeable. Socrates uses the above principle to argue that since acting unjustly is admitted to be more disgraceful than being treated unjustly, and since it is clearly not more unpleasant, it must be worse (i.e. more harmful), and immediately (475 d–e) forces Polus to admit that neither he nor anyone else would prefer what is worse to what is less bad. It is plain that what is meant is that no one will prefer what is worse for himself, since Plato is well aware that many people actively pursue what is worse *for others*. Similar moves are made in the next stage of the argument, where Socrates tries to show that if one has done wrong it is better for one to be punished than to go unpunished. The man who is justly punished is treated in a creditable way. But by the previous principle whatever is creditable is either pleasant or beneficial. Punishment is not pleasant for the person who undergoes it. Therefore the man who undergoes just punishment 'undergoes good things' (has good things happen to him). Therefore, he is benefited (476e–477a). Similarly, if vice is the most shameful thing it must be the most painful or the most harmful; but it is not painful, therefore 'since it exceeds in the greatest harm it is the greatest evil (*kakon*) of all things' (477e). It is clear, then, that throughout this argument the question of whether one alternative action is better or worse, as opposed to more or less creditable or honourable, than another is the question of whether that action tends more to promote or hinder the interest of the agent, which is identified as his achievement of a fully satisfactory life (*eudaimonia*).

Assumptions (i) and (ii) are both apparent in one of the few passages where Socrates gives an argument to support any version of P2. The passage in question is *Meno* 77b–78b, where Socrates is examining Meno's suggestion that excellence (*aretē*) is to be defined as 'Desiring fine things and being able to get them' (77b).

Socrates first of all secures Meno's immediate agreement that to desire fine things is to desire good things, and then argues that the inclusion of 'desiring good things' in Meno's specification of *aretē* is redundant since it is impossible for anyone to desire anything except what he takes to be good. Hence his proposed definition can be reduced to 'the ability to acquire good things' (78c). The first point to note is that the distinction between what is fine or honourable (*kalon*) and what is good (*agathon*), which was essential to Polus's position, is completely ignored in this argument, in that Socrates and Meno move immediately from 'desiring fine things' to 'desiring good things' and thereafter conduct the discussion wholly in terms of good and bad, finally substituting 'the ability to acquire *good* things' for Meno's 'being able to get fine things'. This might be taken to suggest that the concept of goodness employed here is not, as in the previous argument, identical with the concept of what promotes the interest of the agent, but the development of the argument shows that this inference would be mistaken. Meno starts by maintaining that some people want good things but others want bad things. The latter class of persons consists of two sub-classes, those who want bad things in the mistaken belief that the things which they want are good things, and those who want bad things in the knowledge that the things they want are bad (77b–c). Socrates asks (d 1–3) whether those who want bad things (of either sub-class) think that the bad things will benefit them, or whether they recognise (*gignōskōn*) that the bad things (sc. which they want) harm whoever gets them; Meno replies that some (sc. those in the first sub-class) think that the bad things will benefit them, while others (sc. the members of the second sub-class) recognise that they will harm them. The assertion that every bad thing harms the person who has or gets it shows that in this argument too things are judged good or bad iff they promote or hinder the interest of the agent, since it would be commonplace that no one could be harmed otherwise than by a bad thing, or benefited otherwise than by a good thing.

So far then, this argument is seen after all to embody assumption (i), viz. that for an agent to judge one alternative better than another is for him to judge that alternative more in his interest than the other. The conclusion of the argument gives assumption (ii), at least in the weaker form that no one so acts as to do what he thinks will damage his interests. The argument continues as

follows. Those who want bad things in the mistaken belief that they are good things are not strictly to be described as wanting bad things, but rather as wanting good things, though the things they want are in fact bad. (It would be clearer to say that the description under which they specify what they want is not 'something bad' but 'something good' or, adopting Santas's terminology,[1] that the intended object of their desire is not something bad but something good.) The members of the other sub-class, however, are assumed to want bad things in the knowledge that they will harm them. But they also believe that anyone who is harmed is made wretched to the extent to which he is harmed. But no one wants to be wretched and unfortunate. Hence no one wants to be harmed and hence no one wants bad things.

One might object to this argument on a number of different counts. Firstly, one might challenge the truth of either of the premisses of the concluding stage, viz. that to be harmed is to be made wretched and unfortunate to the extent to which one is harmed or that no one wants to be wretched and unfortunate. One objection to the first premiss would clearly be irrelevant: this would be the objection that, since one can be harmed without being aware of it, being harmed doesn't imply that one is thereby made wretched, in the sense of subjectively miserable. This misses the mark because the Greek terms *athlios* and *kakodaimōn* imply that the state of the person to whom they apply is a deplorable one, but not necessarily that that person is aware that his state is such. More to the point is the objection that some sorts of harm don't seem significant enough to count towards making the person harmed wretched and unfortunate, descriptions which imply that his overall state is a deplorable one. Thus if I kick someone in the shin, causing pain and a bruise, I may perhaps be said to have harmed him, but it is doubtful if I have really contributed to making him wretched and unfortunate. One line of reply to that objection is to concede that it succeeds in differentiating trivial from serious harm, but to maintain that the kinds of conduct which Socrates is trying to explain away (drinking to excess, dissipating one's fortune, etc.) fall under the

1 'The Socratic Paradoxes', *Philosophical Review*, lxxiii (1964), pp. 147–164, reprinted in A. Sesonske and N. Fleming (eds.), *Plato's Meno: Text and Criticism*, Belmont, California, 1965.

latter description. Another is to take the notion of 'being unfortunate' as 'having one's interests adversely affected': this gives an unexceptionable first premiss, that anyone who is harmed to any extent has his interests adversely affected to that extent, and requires that the second premiss be reformulated as 'No one wants to have his interests adversely affected'. The truth of this premiss might be attacked whether it is expressed in this new formulation or in the original one; surely some people do want their interests to be adversely affected, or even (in an extreme case) want to be wretched and unfortunate. For example, some people want to suffer for their sins, not necessarily with a view to the improvement of their subsequent lot, but just because they have been so deplorably wicked that it is right that they should suffer extremes of misfortune. Others again might want their interests to be sacrificed for the attainment of some ideal. It is indeed possible to defend this premiss against this objection, though the defence lacks plausibility, in my view. I shall not, however, pursue that topic, for even if the truth of both premisses is granted, the major objection still stands that those premisses do not entail the conclusion. Even if it is true that to be harmed is to be made, to some extent, wretched and unfortunate, it does not follow that no one wants things he knows to be harmful. This argument requires the principle that if A wants to do x, and knows that doing x will bring about the existence of state S, then he wants the existence of S; the argument consists in fact of the application of *modus tollendo tollens* to this principle. But the principle is false; for A may want to do x despite the fact that he knows that doing x will bring about the existence of S, a state which he wants not to occur; or he may be quite indifferent to the fact that doing x will bring about S, and hence it will not be the case that he wants S to occur. If he actively wants S not to occur, then this may induce him not to *do* x in fact, but he may still *want* to do x, which of itself falsifies the principle. And if he does x, thereby bringing about S, the principle is still false. For if he does x, in total indifference to the fact that doing x will have the foreseen consequence of bringing about S, or bitterly regretting the fact that it will have that inevitable but deplored consequence, in neither case does he want to bring about S. That is to say, there is nothing which he wants under the description 'object which will bring about S'. The concept of wanting an object under a description figures in a true principle about the relation of desired

objects and their foreseen consequences, viz. 'If *A* wants to do *x* under the description 'action which will bring about the existence of *S*', then *A* wants the existence of *S*', or alternatively 'If *A* wants to do *x* *because* he believes that doing *x* will bring about the existence of *S*, then *A* wants the existence of *S*'. But we cannot save Plato's argument by substituting this principle for the false one on which he actually relies. For given the true principle and Plato's two premisses all that follows is that no one wants anything *because* he believes that that thing will harm him, whereas Socrates' argument requires the conclusion that no one wants anything *which* he believes will harm him. Obviously, most of the cases which Socrates is trying to exclude are cases of wanting something *although* one believes it will harm one, not of wanting something *because* one believes it will harm one.

I thus conclude that the only argument which Plato gives in support of any version of P2 is inadequate. Nor do I believe that any argument can be adequate which endeavours to support P2 by appeal to self-interest as the basis of evaluation. For if P2 is formulated explicitly in terms of the agent's interest, i.e.

If an agent judges that it would be more in his interest to do *x* than to do *y*, then he wants to do *x* more than he wants to do *y*,

counter-examples can readily be supplied, as Davidson himself shows by his demonstration (pp. 101–102), that one may as readily want to act, and actually act, against one's better self-interested judgement as against one's better judgement when supported by other considerations. Why might one incline to accept the self-interested version of P2? Perhaps because one holds the plausible belief that every agent wants the promotion of his own interest more than he wants anything else. That belief would license the further re-formulation of P2 as

If an agent judges that doing *x* would promote what he most wants more than doing *y* would, then he wants to do *x* more than he wants to do *y*.

If that principle is read throughout as a principle of preferential choice, then it becomes

If an agent judges that doing *x* would promote the attainment of what he chooses in preference to anything else more than doing *y* would, he chooses to do *x* in preference to doing *y*.

That is to say, every agent adopts what he takes to be the best means of attaining his ultimate goal. That is plainly false: given the choice of better or worse means of attaining his goal an agent might choose the worse, against his better judgement, for all sorts of reasons, such as squeamishness or laziness. Read as a principle about inclinations, or a mixed principle relating inclination to preferential choice, it is just as plainly false. For example consider the following:

If an agent judges that doing x would tend to promote the attainment of what he chooses to do in preference to anything else more than the doing of y would, he has a stronger inclination to do x than to do y.

This is false, since an agent may find that the taking of the means necessary to his preferred end runs strongly counter to his inclinations; consider the case of a resistance fighter whose preference for the freeing of his country from occupying forces obliges him to conquer his natural revulsion and kill an enemy soldier. Moreover, the belief which leads to this re-formulation (viz. that every agent wants the promotion of his own interest more than he wants anything else), though plausible, is itself false, as is shown by the counter-examples on the one hand of self-sacrifice and on the other of action against one's better judgement. In the first sort of case the agent wants to do something else (e.g. sacrifice his life for an ideal) more than he wants to promote his interest because he judges it better to do so, in the latter he wants to do something else (typically pursue some short-term pleasure) more despite the fact that he judges it better to pursue his long-term interest.

That concludes what I have to say about Plato. I hope that I have shown (a) that Plato accepted a version of P2 because he believed that every agent's ground of evaluation was self-interested, and (b) that that belief is both false and, even if true, insufficient for the truth of P2. I turn now to some brief consideration of Hare.

III

Hare's commitment to P1 and P2 arises from his theory of the prescriptive nature of evaluative judgement and from the close connection which he sees between evaluative judgements and desires. I shall take it that the former is sufficiently familiar to

make exposition otiose, but the latter requires some discussion. He maintains that if *A* wants to do *x* more than he wants to do *y* he assents to the prescriptive judgement 'Let me do *x* in preference to *y*'[1] and in terms of his general theory sincere assent to that judgement requires that he act on it. Hence Hare accepts P1. I am not clear whether he maintains that if *A* wants to do *x* more than he wants to do *y*, *A* is committed also to the judgement 'It is better for me to do *x* than to do *y*'. The latter, of course, would also commit him to accepting 'Let me do *x* in preference to *y*', but from the fact that he is already committed to that we can't of course infer that he is committed to what entails it. But it is clear that the converse implication from evaluative judgement to desire does hold, in Hare's view:

> if we use the word 'desire' in a wide sense, we can say that any evaluation, just because it is prescriptive, incorporates the desire to have or do something rather than something else. The wide sense in which we are here using 'desire' is that in which any felt disposition to action counts as a desire . . .[2]

It isn't altogether clear whether Hare means that any evaluation *whatever* incorporates some desire, or whether the point is restricted to evaluations by an agent of courses of action etc. to be undertaken by himself. If the former, then his thought is that an evaluative judgement such as 'Fischer ought to move his bishop', since it entails the judgements 'Let Fischer move his bishop' and 'Let anyone, in circumstances similar to those in which Fischer is now, move his bishop', includes the desire that those prescriptions be fulfilled. The sense in which assent to such prescriptions incorporates a desire is presumably explained by the statement quoted above that 'any felt disposition to action counts as a desire'. The assent to such prescriptions is seen as something like the acceptance of a commitment that the specified state of affairs should occur, which, at least in the case of the universalised prescription, is in turn explicable as a commitment to *action* of the appropriate sort, should the relevant circumstances ever apply to the prescriber. This commitment may in turn be described without undue distortion as the recognition or perhaps

1 *Freedom and Reason* (Oxford, 1963), pp. 70–71, 91.
2 Ibid., p. 170.

formation of a disposition to act. Since it is quite unclear how anyone could have such an attitude to someone else's action (e.g. no meaning attaches to the statement that I am committed to Fischer's moving his bishop) it is fairest to Hare to assume him to mean that desires are incorporated primarily in first person evaluative judgements, and that they are derivatively incorporated in second and third person judgements in so far as the latter, being universalisable, entail universal prescriptions which in turn entail first person prescriptions. Thus my judgement 'Fischer should move his bishop' does not incorporate the desire on my part that Fischer should move his bishop, but does incorporate the desire on my part that were I in a position relevantly similar to that in which Fischer now is, I should move my bishop.[1] This restriction does not affect the main issue, since it is clearly Hare's view that in making an evaluative judgement on a course of action to be undertaken by himself, the agent incorporates a desire to do whatever it is that he evaluates more favourably, i.e. Hare accepts P2:

If an agent judges that it would be better to do x than to do y, then he wants to do x more than he wants to do y.

As I have already argued that P2 is false, I shall confine myself to consideration of Hare's reasons for accepting it. Fundamental to his reasoning is his thesis that evaluative judgements entail prescriptions, e.g. the judgements 'It would be better for me to do x than y', 'I ought to do x rather than y' and 'It would be all right for me to do x but not all right for me to do y', all entail the prescription 'Let me do x rather than y'. Hare's language in the passage quoted above suggests that we have in such a prescriptive judgement an explanatory concept which will explain how it is that evaluative judgements have action-guiding force. Evaluative judgements incorporate desire, i.e. felt dispositions to action, *because* they are prescriptive. But on the contrary, the sense of such a prescription as 'Let me in situation S

1 But in a later paper Hare implicitly maintains that the judgement 'Let it be the case that $A\phi$s' expresses the desire of the utterer, not merely that he should ϕ were he in a case similar to that in which A now is, but that A should ϕ in the actual case. The paper is 'Wanting: Some Pitfalls' in R. Binkley *et al.* (eds.), *Agent, Action and Reason* (Oxford, 1971), reprinted in Hare, *Practical Inferences* (London and Basingstoke, 1971); the crucial passage occurs on p. 88 of the former volume and on p. 50 of the latter.

do x' has itself to be explained as follows, that if an agent judges that in situation S he ought to do x or that it would be permissible for him to do x, then he will be inconsistent if he fails to do x should S obtain. That is to say, the entailment of the prescriptive judgement has to be explained in terms of the commitment to action on the part of the person making the evaluation, and cannot therefore itself explain that commitment. Hare has therefore to give independent grounds for positing that commitment to action in the first place. It clearly will not do to say that the commitment arises because evaluative judgements express desires, and desires carry a commitment to action. For firstly it is just false that, on any ordinary understanding of the word 'desire', all evaluative judgements express some desire of the person making them. And secondly, in trying to elucidate his notion of the incorporation of desires in evaluative judgements, Hare is obliged to explain his 'wide sense' of 'desire' as 'any felt disposition to action', and we in turn have been obliged to elucidate 'felt disposition' as commitment. So once again, the explanans (i.e. that evaluative judgements express desires) turns out to contain the explanandum (i.e. that evaluative judgements carry a commitment to action).

IV

In my concluding section I propose an alternative principle to P2. I am happy to leave P1 to stand, provided that it is understood as a principle of preferential choice. I have already pointed out that it is false when applied to inclinations.

A number of alternatives to P2 might be proposed. Thus the following has at least the merit of truth:

> If an agent judges that it would be better to do x than to do y he believes that he has better reasons for doing x than for doing y.

But it could reasonably be objected that, while true, this is uninformative, since it merely links one belief or judgement to another, and moreover to a judgement of the same basic sort, since the judgement that one action is better than another is linked to the judgement that one set of reasons is better than another. For these reasons I propose instead the following:

P2″. If an agent judges that it would be better to do x than to do y he ranks the doing of x by him on this occasion higher than he ranks the doing of y by him on this occasion.[1]

This has the merit of bringing out the fact, which those who defend P2 in its original form have grasped, that judging it better to do x than to do y is not purely a matter of recognising that the doing of x possesses some property to a higher degree than the doing of y, but also has some necessary connections with the agent's motivation and behaviour. But it avoids the error of P2, which was to focus those connections on the agent's desires, when in fact they are scattered over a wider range of the agent's attitudes and behaviour. For those connections are mediated by the truth of 'A ranks the doing of x by him on this occasion higher than he ranks the doing of y by him on this occasion'. And the truth of that proposition is loosely linked, in the way characteristic of cluster-concepts, to the satisfaction of an open disjunction of conditions of which the following are typical: A does x spontaneously and unhesitatingly in preference to y; A feels pleased that he has done x in preference to y; A feels remorse that he has not done x in preference to y; A regards this as a typical case of choice between doing x and doing y, and admires people who in such cases do x in preference to doing y. No single condition is either necessary or sufficient, since any condition may be absent, provided that some other is satisfied, or may be present but be overridden by some contrary condition. Yet the satisfaction of some disjunction of conjunctions of those conditions is sufficient, and that of the disjunction of conditions necessary, for the truth of 'A ranks the doing of x by him on this occasion higher than he ranks the doing of y by him on this occasion'. Analogous conditions may be specified for the truth of 'A ranks the doing of actions of type E higher than he ranks the doing of actions of type F'.

1 Cf. Gary Watson, 'Skepticism about Weakness of Will', *Philosophical Review*, lxxxvi (1977), pp. 316–339. Watson accepts (p. 321) the principle 'If one wants to do x more than one want to do y, one *prefers* x to y, or ranks x higher than y on some scale of values, or "desirability matrix"', and claims that that supports the thesis that P2 is 'true if understood in the language of evaluation' (ibid.). Thus understood, P2 is (apparently) expressed as 'If a person judges x to be better than y, then he or she values x more than y'. But that is a version of P2″, not P2. P2 and P2″ are not equivalent, since P2″ is true and P2 false. P2″ does indeed follow from P2, when the latter is taken together with Watson's principle, but to derive P2 from P2″ one requires not that principle but its converse, which, as I argue below (p. 517), is false.

The force of P2″ is then to specify that the conditions necessary for the truth of '*A* ranks the doing of *x* by him on this occasion higher than he ranks the doing of *y* by him on this occasion' are in turn necessary for the truth of '*A* judges that it would be better to do *x* than to do *y*'. It follows that the truth of '*A* judges that it would be better to do *x* than to do *y*' is sufficient for those conditions, i.e. for the satisfaction of the disjunction of the open set of conditions which I have indicated, though not for the satisfaction of any single condition. Now it is true that some of these conditions imply certain propositions about the agent's desires; thus if *A* is ashamed of himself for having done *y* rather than *x*, *A* wishes that he had not done *y*, but had done *x* instead. Again, if *A* is glad that people whom he likes do *x* rather than *y*, *A* in general wants it to be the case that people whom he likes should do *x* rather than *y*. But it is a gross oversimplification to allow the inference from (i) '*A* ranks the doing of *x* by him on this occasion above the doing of *y* by him on this occasion' to (ii) '*A* wants to do *x* on this occasion more than he wants to do *y* on this occasion'.

For instance, (ii) may be false, since *A*'s inclination was to do *y* and he acted on it, but yet it may be true that after he had done *y* he wished that he had done *x*. Or again, (ii) may be false, but it may be true that *A* in general wants it to be the case that people whom he likes should do *x* rather than *y*, and also true that *A* regards this as a typical case of choice between *x* and *y*, which may be sufficient for the truth of (i).

This account of P2″ seems to me to do justice to the truth of Davidson's remark (p. 99) that 'if someone really (sincerely) believes he ought, then his belief must show itself in his behaviour', while at the same time showing the error in the parenthesis which concludes that sentence '(and hence of course in his inclination to act, or his desire)'. For I hope that I have shown that the connections between the judgement that *x* is better than *y* and behaviour are too subtle to be adequately expressed in any single statement of connection between judgement and desire, such as that given by the original P2. Of course, we might take P2 just as a schematic formulation of the complicated web of connections between judgement and desire which I have so roughly sketched. But given that re-reading P2 will fail to satisfy another requirement on which Davidson rightly insists (ibid.) as essential to his account of the problem, viz. that 'want' or any expression which is substituted for it should be univocal between

P1 and P2. For if 'wants to do x more than he wants to do y' in P2 is read as 'ranks the doing of x higher than the doing of y', and if that is elucidated as I have attempted, then it must have the same reading and elucidation in P1. And it will, I hope, be clear without further discussion that, given that elucidation of ranking the doing of x higher than the doing of y, P1 will be false.

P2″ then, is (a) true and (b) distinct from the original P2. It is also consistent with P1 (interpreted as a principle of preferential choice) and P3. And it is the fact that these principles are consistent which constitutes, in my view, the solution to the problem of akrasia.

For cases of akrasia are cases where an agent judges that it is better to do x than to do y, but does y, and this is possible because the agent's judging it better to do x than to do y does not imply either that he does x in preference to doing y or that he wants to do x more than he wants to do y, but rather that he ranks the doing of x by him on this occasion higher than he ranks the doing of y by him on this occason. That ranking does indeed have certain implications concerning what he does and what he wants, but those are not the simple implications specified by P1 and P2. The mistakes of the eminent philosophers whom I have been criticising seem to me to arise from a common source, namely failure to appreciate the complexities of the concept of desire and of its relations with judgement on the one hand and action on the other.

CORPUS CHRISTI COLLEGE, OXFORD