PHILIP PETTIT

DECISION THEORY, POLITICAL THEORY AND THE HATS HYPOTHESIS

In a typically suggestive aside, John Watkins writes:

> Instead of the unified preference-map of normative decision theory, most of us operate with one or other of a number of preference-systems, according to the 'hat' we are currently wearing. When we come home from the office or go out to a party, we may switch easily and unnoticingly from one system to another. (A holiday-maker lazes on the beach in hot sun and hedonic mood. Then he jumps up: a swimmer is in difficulties. His hedonic calculus switches off, moral concern switches on.) But sometimes, of course, the other preference-system stays switched on. If they indicate conflicting decisions the agent will have to make some sort of meta-decision about which system shall now predominate before he can decide what to do.[1]

This paragraph offers a statement of what we may call 'The Hats Hypothesis'. The hypothesis is that human beings are predisposed to act in such a way that we can usually represent their choices as a function of two variables: on the one side, the social context, or at least the perceived social context, of the action; on the other, a utility function, in particular a set of preferences, which that context selects as suitable.[2]

I believe that the hats hypothesis is of great interest and I hope that in this paper I may do something to show why. In the first section I consider what the hypothesis might mean and how it might be defended. In the second I consider the negative significance of the hypothesis for political theory, arguing that it casts doubt on a central assumption of the public choice critique of non-market institutions. And finally in the third section I raise the question of how far it may have positive significance in political theory; I suggest that it gives us a reason for putting faith in an institution that I describe as the intangible hand.

1. THE DEFENCE OF THE HATS HYPOTHESIS

Watkins introduces the hats hypothesis in the context of decision theory. According to that theory, every agent has a more or less complete set of preferences over certain states of affairs, in particular over the outcomes or potential outcomes of choice; this set is expected to satisfy other constraints

Fred D'Agostino & I. C. Jarvie, Freedom and Rationality: Essays in Honor of John Watkins, 23–34.

too but here we may ignore those. Decision theory formulates a requirement of consistency on the states of affairs – in particular, the states of affairs associated with his options – which a rational agent can consistently come to prefer. The different variants of Bayesian theory, for example, require that if an agent's degrees of preferences over relevant outcomes mean, given his probability function, that one option has higher expected utility than alternatives, then in consistency the rational agent must prefer that option.

On one interpretation, what the hats hypothesis says is that every agent has a number of preference-sets over states of affairs, one for each sort of social context, and that as he moves between those contexts he assumes, so to speak, different decision-theoretic identities. That situation would vindicate the claim that we can usually represent an agent's choices in a two-dimensional manner as a function of social context and the contextually relevant preference-set; the agent's probability function will also make a contribution but, assuming it is part of the unchanging background, I do not mention it explicitly.

This interpretation however does not make the hypothesis very interesting. The reason is that the two-dimensional sort of representation envisaged here would seem to ensure the availability, at least in principle, of a one-dimensional counterpart. If I can represent your choices over options A and B as a function of context C^1 and utility function U^1, and your choices over options X and Y as a function of C^2 and U^2, then I can surely construct a single utility function out of the other two, defined over options A-in-C^1, B-in-C^1, X-in-C^2, Y-in-C^2, and I can represent your choices as a function of this alone. Thus what the hats hypothesis would be is simply the claim that for purposes of decision theory options should be more finely individuated than is common; they should be characterised in a way that takes account of the context in which they present themselves.

But this observation need not depress us, for there is a different and more appealing construal of the hypothesis available. In order to motivate this I need to introduce a distinction between preferences over states of affairs – the preferences relevant in decision theory – and preferences over properties that those states of affairs may have.

The best introduction to the distinction may be to consider first that there are two readings possible of many sentences of the form 'I desire that p'. The ascription can mean that for the state of affairs expressed by 'p', I desire it; I prefer it to the relevant alternatives. Alternatively it can mean that while I may not actually desire that state of affairs – my preferences may not yet be formed – I have a tendency to desire it so far as it promises to make it true that p; the ascription implies that as between any two possible states of affairs between which I am otherwise indifferent, the fact that only one makes it true that p will lead me to prefer it. If I say that I desire to see the sun, that may mean that as between two given alternative states of affairs, say a skiing holiday and one in the tropics, I desire the second. Alternatively however it may just mean that the property of enjoying sunshine is one for which I am liable to prefer one state of affairs to another. Again if I say that I desire to have a job teaching, that may signal a preference for one state of affairs among a fixed set of alternatives or it may simply register that the property of being a teaching job is one that adds in my calculations to the value of any potential career.[3]

This ambiguity may seem to force a primitive distinction on us, as many philosophers have thought, between the desire *simpliciter* for a state of affairs and the *prima facie* desire for that state of affairs. But it is more naturally taken as pointing us to a distinction between the different objects which desire may take, in the one and only sense of the term 'desire'.[4]

I have desires for particular states of affairs. And I have desires for the properties which particular states of affairs may have. I have at the moment the desire to go on writing and the desire to have lunch with a certain colleague; these are desires for particular states of affairs.[5] And I have at the moment, as I have reliably, a variety of other desires that are desires for the realisation of different properties in whatever states of affairs eventuate. I have a desire for the property which a state of affairs might have of bringing me honour or giving me pleasure, of making the world a more just place or of reducing the amount of suffering there. With any such property the fact that it is displayed by a state of affairs disposes me to desire that state of affairs; it will mean that I prefer the state of affairs to a possible object – it will hardly be a real alternative – whose only relevant difference is that it does not have the property.

It is the merest common sense that side by side with our desires for particular states of affairs, we also have such property-desires. The common sense picture is that when I desire a particular state of affairs I always do so on account of some property or properties that it displays. In scholastic terms, the state of affairs is the material object of the desire, the property its formal object. *Quidquid appetitur sub specie boni appetitur*: whatever is desired is desired under the aspect of good.[6]

If we read the ambiguity as pointing us towards this distinction among the objects of desire, we can still draw a divide between desire *simpliciter* and *prima facie* desire. The divide has a derived status, not a primitive one. Under

the picture adopted here I come to form a desire for every alternative only so far as I identify it as the bearer of certain attractive properties. I come to desire that *p*, period, only so far as I desire that *p, qua F*. And this is just to say, we may take it, that I come to desire that *p simpliciter* only so far as I first have a *prima facie* desire that *p*.

Decision theory inclines us to neglect these observations, since its concerns have to do wholly with how rationality constrains the material objects which someone can consistently desire. Bayesian decision theory, as we have mentioned, lays it down that if a rational agent has preferences among the possible outcomes of two options – and options and outcomes are states of affairs – such that the first option has higher expected utility than the first, then he will choose the first. But it would be a mistake to let decision theory suggest that material objects are the only objects of desire and that the only requirement of consistency that is imposed by rationality is consistency among such objects. Every material desire has a formal as well as a material object and it is just as much a requirement of rationality that an agent be consistent in the sorts of formal objects that move him.

Furnished with the distinction between preferences over states of affairs and preferences over properties, we may return to the hats hypothesis. The reading of the hypothesis which I propose is that it bears on preferences over properties and that what it says is that as we move between different social contexts, different properties become appropriate for determining how we form preferences over states of affairs, in particular over the options that confront us.[7] In a family context the fact that an option promises to help my children's career may be appropriate, giving me a reason for example to spend our spare cash on a home computer. In a political context in which I have to decide on where to award a government contract, it would not be suitable; even if my children work for one of the competing companies, the fact that awarding the contract there would further their careers is not something that it is appropriate for me to consider.

Under the interpretation suggested, the hats hypothesis is a hypothesis about the norms of social life, norms that are relevant to the preferences we come to form. It is a proposition that belongs to the sociology of morals. Whether the hypothesis is borne out or not is an empirical matter but it is surely an intuitive and compelling proposition. Norms are regularities to which most of us conform, and to which most of us approve of anyone else conforming, so that we are naturally authorities on the question of what our norms are.[8] The hats hypothesis, under our interpretation, is intuitively plausible and this in itself suggests that it is probably sound.

Does the interpretation make the hats hypothesis interesting? I think so and the remaining sections of the paper should bear out the claim. But it may be useful here to say why the construal does not fall prey to the objection raised against the earlier interpretation. The objection to that interpretation was that it reduces the hats hypothesis to the observation that in a decision-theoretic representation of an agent, options ought generally to be characterised in a context-bound way. A similar objection may seem to carry against our construal. For can't the claim that the properties appropriate for determining preferences vary from context to context be recast in context-free form? Can't the claim that F is an appropriate property in context C^1 be put as the claim that the property of being F-in-C^1 is appropriate regardless of context, and so on for other cases?

It can; but here this is no objection. The problem in the case of the earlier interpretation was that the context-free reconstrual showed that the hats hypothesis added little to decision-theoretic lore. That problem does not come up with the present interpretation. Let the context-free reconstrual go through, if you wish. It doesn't matter, because the hypothesis still has a substantial point to make over and beyond the points acknowledged in decision theory. It means that our preferences over states of affairs, which decision theory discusses, are formed on the basis of preferences ignored in decision theory: preferences over properties whose appropriateness is contextually related.

2. THE NEGATIVE SIGNIFICANCE OF THE HATS HYPOTHESIS

If the hats hypothesis holds, under our interpretation of it, then important consequences follow for political theory. In this section I will look at a negative consequence: viz., that the critique of non-market institutions associated with the public choice school is put in question.[9]

The public choice critique of non-market institutions, coarsely formulated, comes in a sequence of claims.

1. Whether for empirical or methodological reasons, we should agree with economists that individuals are mostly rank egoists in their behaviour within the market.
2. If we are prepared to make this assumption of individuals in the market, then we should be consistent and apply it also to agents in the polity, the bureaucracy, and every sphere of social life.[10]
3. But if we do apply the assumption of rank egoism, then we ought to be

complacent about non-market institutions only so far as they, like the market, involve an invisible hand whereby private interests are exploited to produce public benefit.

The punch comes in the last proposition. A feature of the competitive market, at least as idealised in neo-classical economics, is that it is meant to deliver the relevant public good – competitive prices – despite the fact, as it is assumed, that individual agents pursue their personal advantage at every point. Private 'vice' is public 'virtue'. This feature does not look likely to hold with most political and bureaucratic structures. Such institutions are not designed to snatch a relevant public benefit from the jaws of private greed. On the contrary, they look like institutions where individuals will put that public benefit at risk if they always pursue their own personal advantage. The politician will pander excessively to the electorally most threatening groups, the bureaucrat will try to expand the size of his personal empire, and so on.

The upshot of the critique is a deep scepticism about political and bureaucratic structures. The following quotation is a good example of the line.

> The market is usually characterised as the private sector, with government agencies and officials occupying the public sector. But what can this possibly mean? It surely doesn't mean that consumers and the managers of business firms pursue private interests, whereas everyone who works for government pursues the public interest. The senator who claims that 'the public interest' guides all his decisions is in fact guided by a personal interpretation of the public interest, filtered through all sorts of private interests: re-election, influence with colleagues, relations with the press, popular image, and a place in the history books. Senators may be less interested than business executives in maximising their private monetary income, but they're probably more interested on average in acquiring prestige and power. The same kind of analysis applies to any employee of a government agency, whether it is a high appointed official on a regulatory commission or someone just starting a job at the lowest civil-service rank. However lofty, noble, or impartial the stated objectives of a government agency its day-to-day activities will be the consequence of decisions made by ordinary mortals, subject to the pull and push of incentives remarkably similar to those that operate in the private sector.[11]

We are now in a position to say what the negative significance of the hats hypothesis is. The hypothesis raises a serious question about the public choice critique of non-market institutions. It means that the fact that we assume rank egoism of agents in the market does not entail that we ought equally to assume such rank egoism of agents in, for example, political and bureaucratic contexts.[12]

To assume rank egoism in an agent, plausibly, is to assume that the properties by reference to which he forms his preferences over states of affairs, in particular over the options he faces, are all self-regarding features. He forms his material preferences by reference only to formal objects such as his own pleasure or success or honour; he is insensitive to any effects on others or on the world more generally. The hats hypothesis means that we may assume rank egoism among agents in economic contexts without thinking that consistency requires us to impose the assumption elsewhere. Economic, political and bureaucratic contexts may make different properties appropriate for determining preferences over states of affairs; they may encourage the wearing of different hats.

The hats hypothesis shows that the consistency claim put forward in Proposition 2 above is overstated. That we assume that market agents are rank egoists does not necessarily entail that we must regard agents in other institutional contexts as rank egoists. It may be that we assume that market agents are rank egoists, for example, so far as we assume that all agents behave according to local institutional norms and that the norms of the market are egoistic. Thus the assumption that market agents are rank egoists may be entirely consistent with different assumptions about how people behave in other contexts. The transition from Proposition 1 to Proposition 2, as the latter Proposition stands, is a *non sequitur*.

3. THE POSITIVE SIGNIFICANCE OF THE HATS HYPOTHESIS

The hats hypothesis reveals a lacuna then in the usual case for the public choice critique. That is some consolation for those who put their faith in non-market institutions. But it is not much since, for all that has been said so far, non-market institutions may still be subject to the sort of malaise alleged by public choice theorists.

The hats hypothesis holds that different properties are appropriate in different contexts for determining agents' preferences over states of affairs. Let it be granted, in particular, that the properties appropriate in political and bureaucratic contexts are indeed different from the self-interested considerations alleged to be suitable in economic settings. Even then the hypothesis offers little consolation for those who look with optimism to non-market institutions. For so far we have been given no reason to think that people will form their preferences on the basis only of contextually appropriate

properties. The hypothesis itself does not rule out the possibility, for example, that in awarding a contract in a competition between rival firms, a minister in government may think like a father and award the contract to a firm which employs his son or may think like an entrepreneur and award the contract to the firm that offers the largest bribe.

But after the bad news comes the good. It turns out that if we add a further plausible assumption to the hats hypothesis, all of this changes and the hypothesis assumes positive as well as negative significance in political theory; it becomes significant other than for the negative feature of raising a doubt about the public choice critique of non-market institutions. The extra assumption required is that people generally desire honour. They want others not to think badly of them and, if possible, they want others to think well. In particular they want others to see them as forming their preferences over states of affairs on the basis of contextually suitable properties and only contextually suitable properties.

This assumption is intuitively acceptable, fitting for example into that long tradition of European thought in which the love of esteem is hailed as one of the great human passions.[13] Adam Smith gives forceful expression to it as follows.

> Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren. She taught him to feel pleasure in their favourable, and pain in their unfavourable regard. She rendered their approbation most flattering and most agreeable to him for its own sake; and their disapprobation most mortifying and most offensive.[14]

The assumption should also recommend itself nowadays, for it is built into at least two major schools of contemporary social theory. Clearly it should be acceptable to those in the sociological tradition of theory within which the desire for status ranks with the desire for wealth and for power as one of the basic human desires. And it should be congenial even to those who work in the more economic tradition of rational choice. The motivational postulate of that theory is almost canonically formulated by John Harsanyi. "People's behaviour can be largely explained in terms of two dominant interests: economic gain and social acceptance".[15] The desire for social acceptance is precisely the sort of thing postulated in our assumption.

If we add this assumption to the hats hypothesis, then we begin to find reason for thinking that people in political and bureaucratic contexts will form their preferences on the basis of contextually appropriate properties – say, properties relevant to the public interest – rather than in the fashion of

rank egoism. So far as such agents run a risk of exposure if they form their preferences on any other basis, their desire for esteem will encourage them to form their preferences on this. Any significant chance of exposure, even a low one, may be all that is needed, since even a single disclosure – a single scandal – may destroy the agent's reputation with the relevant audience.

But here's an objection. What the desire for esteem will encourage, so it may be said, is not sensitivity to the contextually appropriate properties, only sensitivity to the property of seeming to be sensitive to those properties. The observation however is not damaging. First, even if true this wouldn't matter from the point of view of how people behave – and that is the final evaluative viewpoint – so long as there are few opportunities where it is rational to reject the dictate of the contextually appropriate properties in the hope of not being detected. And secondly, since it will be costly and difficult actually to form one's preferences with a view to honour, while seeming to do so with a view only to contextually appropriate properties, the rational strategy may well be to look normally only to those properties, safe in the knowledge that this is probably the best way to promote one's honour and other relevant goods. The rational strategy may be to go restrictive, making one's decisions by criteria other than reference to the goods – in this case honour – with which one is ultimately most concerned.[16]

If we follow the line of thought suggested, ignoring the objection, then we begin to connect with a tradition of thinking which has been neglected by recent liberals, particularly by those of the public choice bent. That is the tradition of republicanism which, derived from ancient Roman sources, came to dominate progressive political thinking from the Renaissance, in particular from Machiavelli, down to the end of the 18th century.[17] One of the main assumptions in that tradition is that if we cannot rely on nature to produce contextually appropriate deliberation and action, we can at least rely on pride.[18] One of the distinctive contributions of the tradition was to emphasise that the way to exploit pride for the collective good of society is to subject all holders of public office to the glare or threat of publicity.[19]

The institutions which exploit pride and publicity in this way, many of them the product of the republican tradition, are various. Consider for example the institution of the jury. Jurors are encouraged to meet suitable standards of deliberation and action – ultimately to form their preferences on the basis of appropriate properties – by a number of stratagems. First, those with such a special interest in the outcome that the threat of publicity might not inhibit them are filtered out in jury selection. Secondly, jurors are protected by confidentiality from any threats to their welfare that might be

made from outside. And thirdly, jurors are exposed to internal pressure to behave virtuously by being put in a situation where they have to answer to other members of the jury for how they vote. The institution of the jury exemplifies a general pattern. It is replicated for example in the appointment committee, as that is usually established in the bureaucracy and related areas, including universities. Members of such committees are vetted for a special interest, are protected from external pressure, and are subjected to the internal pressure of having to answer for how they vote.

Moving to higher levels of public office, judges, parliamentarians and members of government are also subjected to publicity constraints which, when they work well, ought to produce behaviour that is consonant with the context. Such officers are forced to answer for how they deliberate and act, in particular to show that they form their preferences over options in a contextually suitable manner: by reference to the appropriate properties. Judges are at least informally required to make statements in defense of their rulings and politicians are expected to justify themselves to parliament and to the public. And apart from such direct pressures, judges and politicians are also faced with the prospect of appeals and inquiries which are designed to inhibit any tendency to deviate from the standards we would wish to obtain.

These are brisk observations but I hope they may be sufficient to show that the hats hypothesis with which we began leads in directions that are of great positive significance for political theory. In conclusion I would like to introduce a nice metaphor which serves to catch the lesson of our observations.

Public choice enthusiasts, and liberals more generally, rely on one sort of mechanism in particular – the invisible hand – to achieve the social alchemy we all desire: that is, to ensure that public benefits will be produced by agents who are not mainly concerned with those benefits and who may even be rank egoists. The invisible hand does its work by suitably aggregating whatever actions are forthcoming from the individuals involved. Thus, to take the case of the idealised competitive market, it puts the producers of different goods and services in predicaments such that by each rationally pursuing their personal advantage – undercutting one another and undermining potential cartels – they ensure that prices fall to the competitive level.[20]

What we have identified here is a different sort of mechanism for achieving a similar magic: a mechanism that I describe, in parallel metaphor, as the intangible hand. The intangible hand produces the public benefit sought, not by suitably aggregating whatever actions are forthcoming from individuals, but by influencing individuals in such a way that they act as if they were

predominantly concerned with the benefit in question. It is a formative rather than an aggregative mechanism.

There is much more to be said about the nature of these two institutional mechanisms and about their distinctive advantages and disadvantages. But it must wait for another place. I have been concerned in this paper only to make a connection between acceptance of the hats hypothesis and recognition of the intangible hand. I hope enough has been said to forge that link.[21]

Australian National University

NOTES

[1] John Watkins, 'Imperfect Rationality', in Robert Borger and Frank Cioffi, Eds., *Explanation in the Behavioural Sciences*, Cambridge University Press, 1970, p. 207.

[2] Watkins would be impatient of the idealisation involved in ascribing a utility function, even a contextually specific utility function, to an agent; but the idealisation is convenient for my purposes and nothing of substance is affected by it.

[3] On this ambiguity see Frank Jackson, 'Internal Conflicts in Desires and Morals', *American Philosophical Quarterly*, Vol. 22 (1985) and 'Davidson on Moral Conflicts' in E. LePore and B. McLaughlin, (Eds.), *Actions and Events*, Blackwells, Oxford, 1985.

[4] On this point see Jackson, 'Internal Conflicts in Desires and Morals'.

[5] That such a state of affairs is conceived by me as a particular does not mean of course that it may not turn out in different ways, as different possible worlds are realised; it is a more abstract particular, an expectation of such particulars.

[6] The picture is well entrenched in the Aristotelian tradition. See Joseph Raz (Ed.), *Practical Reasoning*, Oxford University Press, 1978. See too David Milligan, *Reasoning and the Explanation of Actions*, Harvester Press, Brighton, 1980, Chapter 3. It is invoked in Jackson's 'Internal Conflicts in Desires and Morals'. The Gorman Lancaster translation of commodities into characteristics represents a parallel in economics. See Amartya Sen, *Choice, Welfare, and Measurement*, Blackwells, Oxford, 1982, p. 30.

[7] See Philip Pettit, 'The Life-World and Role-Theory', in Edo Picevic (Ed.), *Phenomenology and Philosophical Understanding*, Cambridge University Press, 1975.

[8] See my paper 'The Inevitability of Norms', *Ethics*, forthcoming.

[9] For an overview of the tradition see Dennis Mueller, *Public Choice*, Cambridge University Press, 1979 and Iain McLean, *Public Choice: An Introduction*, Blackwells, Oxford, 1987.

[10] On this consistency requirement see Geoffrey Brennan and James Buchanan, 'The Normative Purpose of Economic Science: Rediscovery of an Eighteenth Century Method', *International Review of Law & Economics*, Vol. 1, (1981), and 'Predictive Power and the Choice Among Regimes', *The Economic Journal*, Vol. 93, (1983). Brennan and Buchanan are themselves sophisticated enough in their account of

consistency not to let it force them to endorse Proposition 2. But the same does not hold of others.

[11] Paul Heyne, *The Economic Way of Thinking*, 4th Edition, Science Research Associates, Chicago, 1983, pp. 271–72.

[12] See in this connection my paper 'Towards a Social Democratic Theory of the State', *Political Studies*, Vol. 35 (1987), 42–55.

[13] See for example Arthur O. Lovejoy, *Reflections on Human Nature*, Johns Hopkins Press, Baltimore, 1961, Lecture V.

[14] *The Theory of Moral Sentiments*, D.D. Raphael and A.L. Macfie (Eds.), Liberty Classics, Indianapolis, 1982, p. 116.

[15] 'Rational Choice Models of Behavior versus Functionalist and Conformist Theories', *World Politics*, Vol. 22, (1969), 513–38. The postulate is quoted with approval in Michael Taylor, 'Rationality and Revolutionary Collective Action', in Taylor (Ed.), *Rationality and Revolution*, Cambridge University Press, 1967, p. 66.

[16] See Philip Pettit and Geoffrey Brennan, 'Restrictive Consequentialism', *Australasian Journal of Philosophy*, Vol. 64, (1986), 438–55.

[17] See J.G.A. Pocock, *The Machiavellian Moment: Florentine Political Thought and the Atlantic Republic Tradition*, Princeton University Press, 1975.

[18] See Arthur O. Lovejoy, *Reflections on Human Nature*, Johns Hopkins Press, Baltimore, 1961, Lecture V.

[19] For a fuller development of this theme see Philip Pettit, 'The Freedom of the City: A Republican Ideal', in Alan Hamlin and Philip Pettit (Eds.), *The Good Polity: Essays in Normative Analysis of the State*, Blackwells, Oxford, 1988. That paper draws in particular on two articles by Quentin Skinner: 'Machiavelli on the Maintenance of Liberty', *Politics*, Vol. 18, (1983); and 'The Idea of Negative Liberty: Philosophical and Historical Perspectives', in Richard Rorty, J.B. Schneewind and Quentin Skinner (Eds.), *Philosophy in History*, Cambridge University Press, 1984. See also his more recent paper 'The Paradoxes of Political Liberty,' in S.M. McMurrin (Ed.), *The Tanner Lectures on Human Values*, Vol. 7, Cambridge University Press, 1986.

[20] The predicaments are all prisoner's dilemmas. See Philip Pettit 'Free Riding and Foul Dealing' *Journal of Philosophy*, Vol. 83 (1986); reprinted in *The Philosopher's Annual*, Vol. 9, (1987).

[21] I am much in the debt of John Braithwaite and Geoffrey Brennan, both close colleagues for discussions over a long period on matters related to the distinction between the invisible hand and the intangible hand. The distinction is central to a book which I hope to publish with Braithwaite entitled *Not Just Deserts: A Republican Theory of Criminal Justice*. It is the topic of a paper that I hope to co-author with Brennan. I am also indebted to Fred D'Agostino and Barry Hindess for comments on an earlier draft.

J.J.C. SMART

NEGATIVE UTILITARIANISM

1

One of John Watkins's many notable contributions to philosophy is his paper 'Negative Utilitarianism', which is the second part of a symposium of that title, the other symposiast being H.B. Action.[1] Both symposiasts consider and reject a form of utilitarianism that had been extracted by my brother Ninian Smart from some remarks in Popper's *Open Society and Its Enemies*.[2] However the interpretative principle of charity should prevent us from interpreting Popper as a negative utilitarian in my brother's sense: as my brother indeed remarks, if Popper did have the negative utilitarian principle as a fundamental axiom of his ethics, he also had at least two other principles as well. Certainly if Popper had been a negative utilitarian at all he would have been a singularly muddle headed one.

Indeed in an addendum to the 5th edition, p. 296, Popper makes it clear that he did not regard himself as a utilitarian of any sort. He says that he did not put up the negative utilitarian principle as a *criterion* of right action, and that he would object to the idea of such a criterion just as much as he would object to the idea of a criterion of truth. I am not sure that this by itself makes him a non-utilitarian, since the utilitarian should not put up his or her principle as a *criterion* of right action, if this is meant as definition of right action. Certainly as a utilitarian myself I put up the principle as the expression of an attitude, and not as definitional. (If it were definitional it would be ethically trivial.) As a matter of fact I regard the word 'criterion' as it has been used in philosophy as a bit of a weasel word: I believe that in the hands of Wittgenstein at least it has led to obscurity and fence sitting. Be that as it may, Popper makes it clear that minimization of suffering is only one of several moral principles that he accepts. Moreover in the addendum Popper asserts that he put up the principle of minimization as a matter for public policy rather than for private morality.

The negative utilitarian principle is that we should minimize the amount of suffering and unhappiness in the world. Classical utilitarianism is concerned with the maximization of happiness. In classical utilitarianism unhappiness is conceived as negative happiness, and if we considered happiness as negative

Fred D'Agostino & I. C. Jarvie, Freedom and Rationality: Essays in Honor of John Watkins, 35–46.