command system, such as the rationing of goods during a war or a famine. But even here it will often be better, from the point of view of social welfare, to allow individuals, should they wish, to engage in trade with their rations. And the reason here is the same as the one explored in section 3 – the fact that not all private information is publicly known.

All this is not to say that the claims of distributive justice cannot conflict with individual rights. They can, and an enormous literature, both in political philosophy and economics, bears witness to this. But not all rights are equally compelling. In any case, I have not attempted to discuss precisely which rights are instrumental in promoting distributive justice in an economy with dispersed information. They will clearly vary from case to case. My aim has been to argue that a pure command system, narrowly defined, is not the optimum mode of organisation even from the point of view of distributive justice.

11 Sour grapes – utilitarianism and the genesis of wants[1]

JON ELSTER

I want to discuss a problem that is thrown up by all varieties of utilitarianism: act and rule utilitarianism, average and aggregate, cardinal and ordinal.[2] It is this: why should individual want satisfaction be the criterion of justice and social choice when individual wants themselves may be shaped by a process that preempts the choice? And, in particular, why should the choice between feasible options only take account of individual preferences if people tend to adjust their aspirations to their possibilities? For the utilitarian, there would be no welfare loss if the fox were excluded from consumption of the grapes, since he thought them sour anyway. But of course the cause of his holding the grapes to be sour was his conviction that he would be excluded from consumption of them, and then it is difficult to justify the allocation by reference to his preferences.

I shall refer to the phenomenon of sour grapes as *adaptive preference formation* (or adaptive preference change, as the case may be). Preferences shaped by this process I shall call adaptive preferences.[3] The analysis of this mechanism and of its relevance for ethics will proceed in three steps. Section I is an attempt to circumscribe the phenomenon from the outside, by comparing it with some other mechanisms to which it is closely related and with which it is easily confused. Section II is an analysis of the fine

[1] Earlier drafts of this paper were read at the universities of Oslo, Oxford and East Anglia, resulting in major improvement and changes. I am also grateful for valuable and invaluable comments by G. A. Cohen, Robert Goodin, Martin Hollis, John Roemer, Amartya Sen, Arthur Stinchcombe and Bernard Williams.

[2] In fact, the problem is relevant for all want-regarding theories of ethics and justice. John Rawls' theory might seem to escape the difficulty, because it relies on primary goods rather than on utility or preferences. But in fact even his theory needs preference in order to compare undominated bundles of primary goods, and then the problem of sour grapes could easily arise.

[3] The term 'adaptive utility' is used by Cyert and DeGroot (1975), but in a sense more related to what I here call endogenous preference change due to learning. These authors also use the term to refer to what should rather be called 'strategic utility', which is the phenomenon that expected future changes in utility due to learning can be incorporated in, and make a difference for, present decisions. I do not know of any discussions in the economic literature of adaptive preferences in the sense of the term used here, but some insight can be drawn from the economic analysis of Buddhist character planning in Kolm 1979.

grain of adaptive preferences, and proposes some criteria by which they may be distinguished from other preferences. And section III is a discussion of the substantive and methodological implications of adaptive preference formation for utilitarianism, ethics and justice.

I

I shall compare adaptive preference formation to one mechanism that in a sense is its direct opposite; and then to five mechanisms that either have similar causes or bring about similar effects. The purpose of this conceptual mapping is to prepare for the discussion in section III of the exact relevance of adaptive preferences for ethics.

The opposite phenomenon of sour grapes is clearly that of 'forbidden fruit is sweet', which I shall call counteradaptive preference formation.[4] If when I live in Paris I prefer living in London over living in Paris, but prefer Paris over London when in London, then my wants are shaped by my feasible set, as in adaptive preference formation, but in exactly the opposite way. The question then is whether, in the theory of social choice, we should discount wants that have been shaped by counteradaptive preference formation. If someone wants to taste the forbidden fruit simply because it is forbidden, should we count it as a welfare loss that he is excluded from it? And would it be a welfare gain to give him access, if this would make him lose his taste for it? An ordinal-utilitarian theory of social choice offers no answers to these questions. This indeterminacy in itself points to an inadequacy in that theory, although we shall see in section III that counteradaptive preferences are less troublesome for ethics than adaptive ones, because they do not generate any conflict between autonomy and welfare.

Adaptive preference formation is now to be distinguished, firstly, from preference change through learning and experience. Consider the example of job preferences. Imperfect regional mobility sometimes leads to dual labour markets, e.g. to income in agriculture being systematically lower than in industry. Such income gaps may reflect the agricultural labourer's preference for being his own master, or for certain commodities that are cheaper in the countryside than in the city. The labourer may prefer to stay in the countryside rather than move to the city, even if the demand for agricultural goods is too small to enable him to earn the same monetary income as a factory worker. What are the welfare implications of this state of affairs? The standard answer is that a transfer of the labourer to the city

[4] For the record, it may well be adaptive in some larger sense to have counteradaptive preferences, because of the incentive effects created by a moving target.

implies a loss in welfare for him and, *ceteris paribus*, for society. Consider, however, an argument proposed by Amartya Sen:

> Preferences about one's way of life and location are typically the result of one's past experience and an initial reluctance to move does not imply a perpetual dislike. The distinction has some bearing on the welfare aspects of employment policy, since the importance that one wishes to attach to the wage gap as a reflection of the labourer's preferences would tend to depend on the extent to which tastes are expected to vary as a consequence of the movement itself.[5]

On a natural reading of this passage, it seems to sanction the transfer if the *ex post* evaluation of city life makes it preferable to the countryside life that was more highly valued *ex ante*. We then need to ask, however, about the exact nature of the induced change in preferences. Two possibilities come to mind. One is that the transfer would imply learning and experience, another that it is due to habituation and resignation (adaptive preference change). On the first explanation the process is irreversible, or at least it cannot be reversed simply by a reverse transfer to the countryside. (It may, of course, be reversed by learning even more about the alternatives.) The second explanation does, however, permit a reversal of the preference change. I do not imply that irreversibility is a sufficient reason for concluding that preference change is due to learning more about the alternatives: preference change due to addiction also is irreversible in some cases. Nor is it exactly a necessary condition, for it is easy to think of ways in which preference change due to learning may be reversed, and not only through more learning. But, in the present context, irreversibility is the salient feature that permits us to distinguish between these two mechanisms of induced preference change: the reversal to the initial situation does not by itself bring about a reversal of the preferences.

Explanations in terms of learning can be fitted into an extended utilitarian framework, in which situations are evaluated according to *informed* preferences rather than just the given preferences. One should attach more weight to the preferences of someone who knows both sides of the question than to someone who has only experienced one of the alternatives. These informed preferences are, of course, those of the individuals concerned, not of some superior body. They are informed in the sense of being grounded in experience, not in the sense (briefly mentioned in section III) of being grounded in meta-preferences. They differ from given preferences at most in their stability and irreversibility. Informed preferences could be implemented in social choice by a systematic policy of experimentation that gave individuals an opportunity to learn about new alternatives without definite commitment. This no doubt would leave the persons

[5] Sen 1975, pp. 43–54.

involved with more information, but also with less character.[6] If individuals were reared every second year in the countryside, their eventual choice would be better informed, but they would have less substance as persons.

Be this as it may, it is clear that explanations in terms of habituation and resignation cannot even be fitted into this extended utilitarianism. If preferences are reversibly linked *to* situations, then preferences *over* pairs of situations appear in a very different light. If an initial preference for city life could be reversed by extended exposure to the countryside and *vice versa*, then Sen's argument (in my reading of it) implies that we do not have to bother with preferences at all. And this is not an extension of utilitarianism, but its breakdown. At least this holds for ordinal utilitarianism.[7] Cardinal utilitarianism, in its classical version, is perfectly capable of handling the problem, by comparing the total want satisfaction of countryside life with countryside preferences to city life with city preferences. But, as further argued in section III, cardinal utilitarianism then has to face other and even more serious problems.

Adaptive preference formation can be distinguished, secondly, from precommitment, by which I mean the deliberate restriction of the feasible set.[8] If my preferred alternative in the feasible set coincides with my preferred alternative in a larger set of possible alternatives, this may indeed be due to adaptive preference change, but it may also happen because I have deliberately shaped the feasible set so as to exclude certain possible choices. Some people marry for this reason, i.e. they want to create a barrier to prevent them from leaving each other for whimsical reasons. Other people abstain from marriage because they want to be certain that their love for each other is not due to adaptive preference formation. It does not seem possible to ensure both that people stay together for the right reasons, and that they do not leave each other for the wrong reasons. If one deliberately restricts the feasible set, one also runs the risk that the preferences that initially were the reason for the restriction ultimately come to be shaped by it, in the sense that they would have changed had they not been so restricted.

Another example that shows the need for this distinction is the desire for submission to authority. As brilliantly argued by Paul Veyne[9] in his study of authority relations in Classical Antiquity, the mechanism of sour grapes may easily lead the subjects to glorify their rulers, but this is then an

[6] This observation owes much to Williams 1976a.
[7] I am grateful to G. A. Cohen for pointing out to me the crucial difference between ordinal and cardinal utilitarianism in this respect.
[8] Elster 1979, Ch. II has an extended analysis of this notion, with many examples.
[9] Veyne 1976. For an exposition and interpretation of Veyne's view, see Elster 1980.

ideology induced by and posterior to the actual submission, not a masochistic desire that generates and justifies it. As in the preceding example, we need to distinguish between preferences being the cause of a restricted feasible set, and their being an effect of the set. The oppressed may spontaneously invent an ideology justifying their oppression, but this is not to say that they have invented the oppression itself.

Adaptive preferences, thirdly, differ from the deliberate manipulation of wants by other people. If one only wants what little one can get, one's preferences are perhaps induced by other people in whose interests it is to keep one content with little:

> *A* may exercise power over *B* by getting him to do what he does not want to do, but he also exercises power over him by influencing, shaping or determining his very wants. Indeed, is it not the supreme exercise of power to get another or others to have the desires you want them to have – that is, to ensure their compliance by controlling their thoughts and desires? One does not have to go to the lengths of talking about *Brave New World*, or the world of B. F. Skinner, to see this: thought control takes many less total and more mundane forms, through the control of information, through the mass media and through the processes of socialisation.[10]

There is an ambiguity in this passage, for does it propose a purposive or a functional explanation of wants? Do the rulers really have the power to induce deliberately certain beliefs and desires in their subjects? Or does the passage only mean that certain desires and beliefs have consequences that are good for the rulers? And if the latter, do these consequences explain their causes? As argued by Veyne, the purposive explanation is implausible.[11] The rulers no doubt by their behaviour are able to induce in their subjects certain beliefs and values that serve the rulers' interest, but only on the condition that they do not deliberately try to achieve this goal. From the rulers' point of view, the inner states of the subjects belong to the category of *states that are essentially byproducts*.[12] The functional explanation hinted at in the reference to 'processes of socialisation' is no more plausible. True, adaptive preference formation may have consequences that are beneficial to the rulers, but these do not explain how the preferences came to be held. On the contrary, the very idea of adaptation points to a different explanation. It is good for the rulers that the subjects be content with little, but what explains it is that it is good for the subjects. Frustration with the actual state of affairs would be dangerous for the

[10] Lukes 1974, p. 23.
[11] Veyne 1976, passim.
[12] Farber 1976 has a brief discussion of a similar notion, 'willing what cannot be willed'. He restricts the idea, however, to the inducement of certain states (belief, sleep, happiness) in oneself, whereas it can also be applied to paradoxical attempts to induce by command certain states (love, spontaneity, disobedience) in others. For the latter, see the works of the Palo Alto psychiatrists, e.g. Watzlawick 1978.

rulers, but also psychologically intolerable to the ruled, and the latter fact is what explains the adaptive preferences. How it explains them is brought out by the next distinction.

Adaptive preference formation, fourthly, differs from deliberate character planning. It is a causal process taking place 'behind my back', not the intentional shaping of desires advocated by the Stoic, buddhist or spinozistic philosophies, by psychological theories of self-control or the economic theory of 'egonomics'.[13] The psychological state of wanting to do a great many things that you cannot possibly achieve is very hard to live with. If the escape from this tension takes place by some causal mechanism, such as Festinger's 'reduction of cognitive dissonance',[14] we may speak of adaptive preference change. The process then is regulated by something like a drive, not by a conscious want or desire. If, by contrast, I perceive that I am frustrated and understand why, I may deliberately set out to change my wants so as to be able to fulfil a larger part of them. I then act on a second-order desire, not on a drive. To bring home the reality of the distinction between drives and second-order wants, consider counter-adaptive preferences. No one could choose to have such preferences, and so they can only be explained by some kind of perverse drive of which it can be said, metaphorically speaking, that it has the person rather than the other way around.

The difference between adaptive preference formation and deliberate character planning may show up not only in the process, but in the end result as well. One difference is that I may, in principle at least, intentionally shape my wants so as to coincide exactly with (or differ optimally from) my possibilities, whereas adaptive preference formation tends to overshoot, resulting in excessive rather than in proper meekness.[15] Another is that adaptive preference change usually takes the form of downgrading the inaccessible options ('sour grapes'), whereas deliberate character planning has the goal of upgrading the accessible ones.[16] In a less than perfect marriage, I may adapt either by stressing the defects of the wise and beautiful women who rejected me, or by cultivating the good points of the one who finally accepted me. But in the general case adaptive preferences and character planning can be distinguished only by looking into the actual process of want formation.

Lastly, adaptive preference formation should be distinguished from wishful thinking and rationalisation, which are mechanisms that reduce frustration and dissonance by shaping the perception of the situation

[13] Schelling 1978.
[14] Festinger 1957; 1964.
[15] Veyne 1976, pp. 312–13.
[16] Kolm 1979.

rather than the evaluation of it. The two may sometimes be hard to tell from each other. In the French version of the fable of the sour grapes, the fox is deluded in his perception of the grapes: they are too green. (And similarly for counteradaptive preferences, as in 'The grass is always greener on the other side of the fence'.) But in many cases the phenomena are clearly distinct. If I do not get the promotion I have coveted, then I may rationalise defeat either by saying that 'my superiors fear my ability' (misperception of the situation) or 'the top job is not worth having anyway' (misformation of preferences). Or again I may change my life style so as to benefit from the leisure permitted by the less prestigious job (character planning).

Just as one cannot tell from the preferences alone whether they have been shaped by adaptation, so one cannot always tell from the beliefs alone whether they arise from wishful thinking. A belief may stem from wishful thinking, and yet be not only coherent, but true and even well-founded, if the good reason I have for holding it is not what makes me hold it. I may believe myself about to be promoted, and have good reasons for that belief, and yet the belief may stem from wishful thinking so that I would have held it even had I not had those reasons. This shows that wishful thinking, like adaptive preference formation, is a causal rather than an intentional phenomenon. Self-deception, if there is such a thing, has an intentional component in that I know the truth of what I am trying to hide from sight. But if what I believe out of wishful thinking is also what I have reason to believe, there can be no such duality. Wishful thinking, it seems to me, is best defined as a drive towards what I want to believe, not as a flight from what I do not want to believe.[17]

In the short run the result of wishful thinking and of adaptive preference change is the same, viz. reduction of dissonance and frustration. In the long run, however, the two mechanisms may work in opposite directions, as in the following important case. This is the classical finding from *The American Soldier* that there was a positive correlation between possibilities of promotion and level of frustration over the promotion system.[18] In the services in which the promotion chances were good, there was also more frustration over promotion chances. In Robert Merton's words, this paradoxical finding had its explanation in that a 'generally high rate of mobility induced excessive hopes and expectations among members of the group so that each is more likely to experience a sense of frustration in his present position and disaffection with the chances for promotion'.[19] Other explanations have also been proposed that make the frustration depend on

[17] I elaborate on these slightly cryptic remarks in Elster (forthcoming).
[18] Stouffer 1949.
[19] Merton 1957, Ch. VIII.

rational rather than excessive expectations.[20] We might also envisage, however, a quite different explanation in terms of sour grapes: frustration occurs when promotion becomes sufficiently frequent, and is decided on sufficiently universalistic grounds, that there occurs what we may call a release from adaptive preferences. On either hypothesis, increased objective possibilities for well-being bring about decreased subjective well-being, be it through the creation of excessive expectations or by the inducement of a new level of wants. The relevant difference between the two mechanisms for ethics is the following. Giving the utilitarian the best possible case, one may argue that frustration due to wishful thinking should be dismissed as irrational and irrelevant. But on the standard utilitarian argument, it is hardly possible to dismiss in the same manner frustration due to more ambitious wants. If we are to do so, we must somehow be able to evaluate wants, but this brings us outside the standard theory.

To recapitulate, then, adaptive preference formation has five distinctive features that enable us to locate it on the map of the mind. It differs from learning in that it is reversible; from precommitment in that it is an effect and not a cause of a restricted feasible set; from manipulation in that it is endogenous; from character planning in that it is causal; and from wishful thinking in that it concerns the evaluation rather than the perception of the situation. These phenomena are all related to adaptive preference formation, through their causes (reduction of dissonance) or their effects (adjustment of wants to possibilities). They also differ importantly from adaptive preferences, notably in their relevance for ethics. Some of these differences have been briefly noted in the course of the discussion; they form a main topic of section III below.

II

From the external characterisation of adaptive preferences, I now turn to the internal structure of that phenomenon. I shall take an oblique route to the goal, beginning with a discussion of the relation between adaptive preference formation and freedom. In fact, both welfare and freedom, as well as power, have been defined in terms of getting or doing what one most prefers. It is well known, but not particularly relevant in the present context, that the attempt to define power in terms of getting what you want comes up against the problem of adaptive preferences.[21] It is equally

[20] Boudon 1977, Ch. V.

[21] Goldman 1972, following Robert Dahl, calls this the problem of the *chameleon*. Observe that adaptive preferences do not detract from power, as they do from welfare and freedom. If you have the power to bring about what you want, it is irrelevant whether your wants are shaped by the anticipation of what would have been brought about anyway. There is nothing shadowy or insubstantial about preemptive power.

well known, and more to the point, that adaptive preferences also create problems for the attempt to define freedom as the freedom to do what you want.

We need to assume that we have acquired some notion of what it means to be *free to do* something. This is not a simple question. It raises problems about the relation between formal freedom and real ability; between the distributive and the collective senses of mass freedom; between internal and external, positive and negative, man-made and natural, deliberate and accidental obstacles to freedom. I cannot even begin to discuss these issues here, and so I shall have to take for granted a rough notion of what freedom to act in a certain way means. But not all freedom is freedom to do something; there is also freedom *tout court*, being a free man. Freedom in this sense clearly in some way turns upon the things one is free to do – but how?

We may distinguish two extreme answers to this question. One is that freedom consists in being free to do what one wants to do. This view is sometimes imputed to the Stoics and to Spinoza, with dubious justification. In a well-known passage Isaiah Berlin argues against this notion of freedom: 'If degrees of freedom were a function of the satisfaction of desires, I could increase freedom as effectively by eliminating desires as by satisfying them; I could render men (including myself) free by conditioning them into losing the original desire which I have decided not to satisfy.'[22] And this, in his view, is unacceptable. Berlin is not led by this consideration into the opposite extreme, which is that freedom is simply a function of the number and importance of the things one is free to do, but his view is fairly close to this extreme.[23] The possibility of adaptive preferences leads him into downgrading the importance of actual wants, and to stress the freedom to do things that I might come to want even if I do not actually desire them now.

There is, however, an ambiguity in Berlin's argument. 'Conditioning men' into losing the desires that cannot be satisfied is a form of manipulation, which means that the ensuing want structure is not a fully *autonomous* one. And I completely agree that full (or optimal) satisfaction of a non-autonomous set of wants is not a good criterion of freedom. And the same holds for the adjustment of aspirations to possibilities that takes place behind my back, through adaptive preference formation. But there is a third possibility, that of autonomous character formation. If I consciously shape myself so as only to want what I can get, I can attain full satisfaction of an autonomous want structure, and this can with more justification be called freedom, in the Stoic or spinozistic sense. Being a free

[22] Berlin 1969, pp. xxxviii–xl.

[23] See Berlin 1969, p. 130 n for an exposition of his view.

man is to be free to do all the things that one autonomously wants to do. This definition is less restrictive than Berlin's (and certainly less restrictive than the extreme view to which he is closest), but more restrictive than the extreme Berlin is attacking, that being free is to be free to do the things one wants, regardless of the genesis of the wants.

If this definition of freedom is to be of real value, we need a definition or a criterion for autonomous wants. This I cannot provide. I can enumerate a large number of mechanisms that shape our wants in a non-autonomus way, but I cannot say with any confidence whatsoever that the wants that are not shaped in any of these ways are *ipso facto* autonomous. And so it seems that for practical purposes we must fall back on a definition similar to Berlin's. But I think we can do better than this. We can exclude operationally at least one kind of non-autonomous wants, viz. adaptive preferences, by requiring freedom to do otherwise. If I want to do x, and am free to do x, and free not to do x, then my want cannot be shaped by necessity. (At least this holds for the sense of 'being free to do x' in which it implies 'knowing that one is free to do x'. If this implication is rejected, knowledge of the freedom must be added as an extra premiss.) The want may be shaped by all other kinds of disreputable psychic mechanisms, but at least it is not the result of adaptive preference formation. And so we may conclude that, other things being equal, one's freedom is a function of the number and the importance of the things that one (i) wants to do, (ii) is free to do and (iii) is free not to do.

An alternative proof that my want to do x is not shaped by the lack of alternatives would be that I am not free to do x. It would be absurd to say that my freedom increases with the number of things that I want to do, but am not free to do, but there is a core of truth in this paradoxical statement. If there are many things that I want to do, but am unfree to do, then this indicates that my want structure is not in general shaped by adaptive preference formation, and this would also include the things that I want to do and am free to do, but not free not to do. And this in turn implies that the things I want to do and am free to do, but not free not to do, should after all count in my total freedom, since there is a reason for believing the want to be an autonomous or at least non-adaptive one. The reason is weaker than the one provided by the freedom to do otherwise, but it still is a reason of a sort. Given two persons with exactly the same things which they both want to do and are free to do, then (*ceteris paribus*) the one is freer (or more likely to be free) who is free not to do them; also (*ceteris paribus*) the one is freer (or more likely to be free) who wants to do more things that he is not free to do.

These two criteria do not immediately carry over from freedom to welfare. The objects of welfare differ from the objects of freedom in that,

for some of them at least, it makes little sense to speak of not being free to abstain from them. It makes good sense to say that freedom of worship is enhanced by the freedom not to worship, but hardly to say that the welfare derived from a certain consumption bundle is enhanced by the option of not consuming that bundle, since one always has that option. Nevertheless it remains true that (i) the larger the feasible set and (ii) the more your wants go beyond it, the smaller the probability that your wants are shaped by it. Or to put it the other way around: a small feasible set more easily leads to adaptive preferences, and even with a large feasible set one may suspect adaptive preferences if the best element in the feasible set is also the globally best element.

On the other hand, even if the best element in the feasible set is also globally best, preferences may be autonomous, viz. if they are shaped by deliberate character formation. The question then becomes whether we can have evidence about this beyond the (usually unavailable) direct evidence about the actual process of want formation. Quite tentatively, I suggest the following *condition of autonomy for preferences*:

> If S_1 and S_2 are two feasible sets, with induced preference structures R_1 and R_2, then for no x or y (in the global set) should it be the case that xP_1y and yP_2x.

This condition allows preferences to collapse into indifference, and indifference to expand into preference, but excludes a complete reversal of preferences. Graphically, when the fox turns away from the grapes, his preference for raspberry over strawberry should not be reversed. The condition permits changes both in intra-set and inter-set rankings. Assume x,y in S_1 and u, v in S_2. Then xP_1u and xI_2u could be explained as a deliberate upgrading of the elements in the new feasible set. Similarly xP_1y and xI_1y could be explained by the fact that there is no need to make fine distinctions among the alternatives that are now inaccessible. And uI_1v and uP_2v could be explained by the need to make such distinctions among the elements that now have become available. By contrast, xP_1u and uP_1x would indicate an upgrading of the new elements (or a down-grading of the old) beyond what is called for. (Recall here the observation that adaptive preferences tend to overshoot.) Similarly xP_1y and yP_2x (or uP_1v and vP_2u) are blatantly irrational phenomena, for there is no reason why adjustment to the new set should reverse the internal ranking in the old.

For a conjectural example of preference change violating this autonomy condition, I might prefer (in my state as a free civilian) to be a free civilian rather than a concentration camp prisoner, and to be a camp prisoner rather than a camp guard. Once inside the camp, however, I might come to prefer being a guard over being a free civilian, with life as a prisoner ranked bottom. In other words, when the feasible set is (x,y,z), I prefer x

over y and y over z, but when the feasible set is (y,z) I prefer z over x and x over y. In both cases the best element in the feasible set is also globally best, not in itself a sign of non-autonomy. But in addition the restriction of the feasible set brings about a reversal of strong preferences, violating the condition. If the restricted set had induced indifference between x and y, both being preferred to z, this would have been evidence of a truly Stoic mastery of self. For another example, consider the labourer who after a transfer to the city comes to reverse his ranking of the various modes of farming, preferring now the more mechanised forms that he previously ranked bottom. Thirdly, observe that modernisation does not merely imply that new occupations are interpolated at various places in the prestige hierarchy, but that a permutation of the old occupations takes place as well.

When a person with adaptive preferences experiences a change in the feasible set, one of two things may happen: readaptation to the new set, or release from adaptation altogether. Proof of the latter would be if the globally best element were no longer found in the feasible set. And even if the feasible best remained the global best, release from adaptation might be conjectured if no reversal of strong preferences took place. Readaptation was illustrated in the city–countryside example, whereas release from adaptation is exemplified below in the example of the Industrial Revolution. In this example the release is diagnosed through the first criterion, that the global best is outside the feasible set. The second criterion (no strong reversal of preferences) presumably would not find widespread application, because of the relative rarity of conscious character planning.

A final remark may be in order. It is perhaps more common, or more natural, to think of preferences as induced by the actual state than by the feasible set. I believe, however, that the distinction is only a conceptual one. Consider again the city–countryside example. To live in the city may be considered globally as a state which (when in the city) I prefer over the countryside, considered as another global state. With a more fine-grained description of the states, however, it is clear that there are many modes of farming, all accessible to me when in the countryside, and many modes of city life that I can choose when I live in the city. Adaptive preferences then imply that according to my city preferences my globally best alternative is some variety of city life, but there may well be some varieties of countryside life that I prefer to some city lives. But in a useful shorthand we may disregard this and simply speak of states as inducing preferences, as will be done in the example developed below.

III

To discuss the relevance of adaptive preferences for utilitarian theory, I shall take up the question whether the Industrial Revolution in Britain was a good or a bad thing. In the debate among historians over this question,[24] two issues have been raised and sometimes confused. First, what happened to the welfare level of the British population between 1750 and 1850? Secondly, could industrialisation have taken place in a less harsh way than it actually did? Focussing here on the first issue, what kind of evidence would be relevant? Clearly the historians are justified in singling out the real wage, mortality, morbidity and employment as main variables: their average values, dispersion across the population and fluctuations over time. But if we are really concerned with the question of welfare, then we should also ask about the level of wants and aspirations. If the Industrial Revolution made wants rise faster than the capacity for satisfying them, should we then say that the Pessimist interpretation was correct and that there was a fall in the standard of living? Or, following the non-Pessimist[25] interpretation, should we say that an increased capacity for want satisfaction implies a rise in the standard of living? Or, following Engels,[26] should we say that, even if there was a fall in the material standard of living, the Industrial Revolution should be welcomed because it brought the masses out of their apathetic vegetation and so raised their dignity?

The problem is analogous to the one of *The American Soldier*, and as in that example there is also the possibility that frustration (if such there was) stemmed from excessive expectations and not from rising aspirations. If that proved to be the case, the utilitarian might not want to condemn the Industrial Revolution. He could say, perhaps, that insatisfaction derived from irrational beliefs should not count when we add up the sum total of utility. If we require preferences to be informed, then surely it is reasonable also to require beliefs to be well-grounded? But I do not think the utilitarian could say the same about frustration derived from more ambitious wants, and if this proved to be the main source of insatisfaction he could be led into a wholesale rejection of the Industrial Revolution. I assume in the immediate sequel that there was indeed some frustration due to a new level of wants, and try to spell out what this implies for utilitarianism. Later on I return to the problem of excessive expectations.

Imagine that we are initially in pre-industrial state x, with induced

[24] Elster 1978a, pp. 196 ff. has further references to this debate.

[25] As argued in Elster 1978a, the terms 'optimism' vs. 'pessimism' are misleading. The issue of pessimism vs non-pessimism is the factual one discussed here, and the question of optimism vs non-optimism the counterfactual one of alternative and better ways of industrialisation.

[26] Engels 1975, pp. 308–9.

utility functions $u_1 \ldots u_n$. We may think of these as either ordinal and non-comparable (i.e. as shorthand for continuous preferences) or as fully comparable in the classical cardinal sense. I shall refer to the two cases as the ordinal and the cardinal ones, but the reader should keep in mind that the crucial difference is that the latter permit one, as the former do not, to speak unambiguously of the sum total of utility. Assume now that industrialisation takes place, so that we move to state y, with induced utility functions $v_1 \ldots v_n$. In addition there is a possible state z, representing a society in which more people enjoy the benefits of industrialisation, or all people enjoy more benefits. Given the utility functions, we assume some kind of utilitarian device for arriving at the social choice. In the ordinal case, this must be some kind of social choice function; in the cardinal case we say that one should choose that state which realises the greatest sum total of utility. We then make the following assumptions about the utility functions $u_1 \ldots u_n$:

Ordinal case: According to the pre-industrial utility functions, x should be the social choice in (x,y,z)
Cardinal case: According to the pre-industrial utility functions, the sum total of utility is larger in x than it would be in either y or z.

We then stipulate the following for the utility functions $v_1 \ldots v_n$:

Ordinal case: According to the industrial utility functions, the social choice mechanism ranks z over y and y over x.
Cardinal case: According to the industrial utility functions, there is a larger sum of utility in z than in y, and a larger sum in y than in x.

And finally I add for the

Cardinal case: The sum total of utility in x under the pre-industrial utility functions is greater than the sum total of utility in y under the industrial utility functions.

This means that before industrialisation, in both the ordinal and the cardinal case, the individuals live in the best of all possible worlds. After industrialisation, this is no longer true, as the social choice would now be an even more industrialised world. Nevertheless the industrialised state is socially preferred over the pre-industrial one, even though (assuming the cardinal case) people are in fact worse off than they used to be. The intuitive meaning is that for everybody z is better than y on some objective dimension (actual or expected income) and y better than x; indeed y is sufficiently much better than x to create a new level of desires, and z sufficiently much better than y to engender a level of frustration that actually makes people (cardinally) worse off in y than they were in x,

although, to repeat, the social choice in y is y rather than x. 'We were happier before we got these fancy new things, although now we would be miserable without them.' Clearly the story is not an implausible one.

What in this case should the utilitarian recommend? The ordinal utilitarian has, I believe, no grounds for any recommendation at all. State x is socially better than y according to the x-preferences, and y better than x according to the y-preferences, and no more can be said. The cardinal utilitarian, however, would unambiguously have to recommend x over y on the stated assumptions. But this, I submit, is unacceptable. It cannot be true that the smallest loss in welfare always counts for more than the largest increase in autonomy. There must be cases in which the autonomy of wants overrides the satisfaction of wants. And the release from adaptive preferences has exactly these consequences in the case that we have described; inducement of frustration and creation of autonomous persons. We do not want to solve social problems by issuing vast doses of tranquillisers, nor do we want people to tranquillise themselves through adaptive preference change. Engels may have overestimated the mindless bliss of pre-industrial society and underrated the mindless misery, but this does not detract from his observation that 'this existence, cosily romantic as it was, was nevertheless not worthy of human beings'.[27]

I am not basing my argument on the idea that frustration in itself may be a good thing. I believe this to be true, in that happiness requires an element of consummation and an element of expectation that reinforce each other in some complicated way. 'To be without some of the things you want is an indispensable part of happiness.'[28] But a utilitarian would then be happy to plan for optimal frustration. I am saying that even more-than-optimal frustration may be a good thing if it is an indispensable part of autonomy. Nor am I arguing that the search for ever larger amounts of material goods is the best life for man. There certainly comes a point beyond which the frustrating search for material welfare no longer represents a liberation from adaptive preferences, but rather an enslavement to addictive preferences. But I do argue that this point is not reached in the early stages of industrialisation. Only the falsely sophisticated would argue that to strive for increased welfare was non-autonomous from its very inception.

I should now explain exactly how this example provides an objection to utilitarian theory. Generally speaking, a theory of justice or of social choice should satisfy two criteria (among others). Firstly, it should be a guide to action, in the sense that it should enable us to make effective choices in most important situations. If in a given case the theory tells us that two or more alternatives are equally and maximally good, then this

[27] Engels 1975, p. 309.
[28] Bertrand Russell, quoted after Kenny 1965–6.

should have a substantive meaning and not simply be an artifact of the theory. The latter is true, for example, of the Pareto principle that x is socially better than y if and only if one person strictly prefers x and y and no one strictly prefers y over x, whereas society is 'indifferent' between x and y if some person strictly prefers x over y and some other person strictly prefers y over x. Even though this principle formally establishes a ranking, it is hopelessly inadequate as a guide to action. A theory should not tell us that some alternatives are non-comparable, nor try to overcome this problem by stipulating that society is indifferent between all non-comparable alternatives.

Secondly, we must require of a theory of justice that it does not strongly violate our ethical intuitions in particular cases. If a theory suggests that people should take tranquillisers when the Coase theorem requires them to,[29] then we *know* that it is a bad theory. True, the proper role of such intuitions is not well understood. If they are culturally relative, one hardly sees why they should be relevant for a non-relative theory of justice. And if they are culturally invariant, one suspects that they might have a biological foundation,[30] which would if anything make them even less relevant for ethics. Perhaps one could hope that persons starting from different intuitions might converge towards a unique reflective equilibrium,[31] which would then represent man as a rational rather than a culturally or biologically determined being. Such problems notwithstanding, I do not see how a theory of justice can dispense with intuitions altogether.

My argument against utilitarianism then is that it fails on both counts. Ordinal utilitarianism in some cases fails to produce a decision, and cardinal utilitarianism sometimes generates bad decisions. The indecisiveness or ordinal utilitarianism is due, as in other cases, to the paucity of information about the preferences. Cardinal utilitarianism allows for more information, and therefore ensures solutions to the decision problem. But even cardinalism allows too little information. Satisfaction induced by resignation may be indistinguishable on the hedonometer from satisfaction of autonomous wants, but I have argued that we should distinguish between them on other grounds.

The distinctions elaborated in section I may now be brought to bear on these issues. The reason why counteradaptive preferences are less problematic for ethics than adaptive ones is that release from counteradaptive preferences simultaneously improves autonomy and welfare. When I no longer possess (or no longer am possessed by) the perverse drive for novelty and change, the non-satisfaction of non-autonomous wants may

[29] As suggested by Nozick 1974, p. 76 n.
[30] As suggested by Rawls 1971, p. 503.
[31] Rawls 1971 is at the origin of this notion.

turn into the satisfaction of autonomous ones. The destructive character of counteradaptive preferences is well illustrated in an example due to von Weiszäcker.[32] Here a person obsessed by the quest for novelty is bled to death by a series of stepwise changes, each of which is perceived as an improvement in terms of the preferences induced by the preceding step. Clearly, to be released from this obsession is both a good thing in itself and has good consequences for welfare. Release from adaptive preferences, however, may be good on the autonomy dimension while bad on the welfare dimension.

Similar remarks apply to character planning, which may improve welfare without loss of autonomy. I am not arguing that character planning is *ipso facto* autonomous, for surely there are non-autonomous second-order wants, e.g. being addicted to will-power.[33] But I do not believe these cases to be centrally important, and in any case I am here talking about *changes* in the degree of autonomy. Character planning may improve welfare compared both to the initial problematic situation and to the alternative solution, which is adaptive preference change. First, recall that character planning tends to upgrade the possible, which cardinally speaking is better than a downgrading of the impossible. Both solutions reduce frustration, but character planning leaves one cardinally better off. Secondly, observe that the strategy of character planning is fully compatible with the idea that for happiness we need to have wants somewhat (but not too much) beyond our means. True, this notion is incompatible with the Buddhist version of character planning that sees in frustration *only* a source of misery.[34] But I believe that this is bad psychology, and that Leibniz was right in that 'l'inquiétude est essentielle à la félicité des créatures'.[35] And this means that character planning should go for optimal frustration, which makes you better off than in the initial state (with more-than-optimal frustration) and also better off than with adaptive preferences, which tend to limit aspirations to, or even below, the level of possibilities, resulting in a less-than-optimal level of frustration.

Endogenous preference change by learning not only creates no problems for ethics, but is positively required by it. If trying out something you believed you would not like makes you decide that you like it after all, then

[32] von Weiszäcker 1971; also Elster 1978a, p. 78, who gives as an illustration the sequence (1/2,3/2), (3/4,1/2), (1/4,3/4), (3/8,1/4) . . . in which each bundle is seen as an improvement over the preceding one because it implies an increase in the smallest component. A very conservative person, conversely, might reject each change in the opposite direction because it implies a reduction in the largest component. Such conservatism is akin to adaptive preference change, since it implies that you systematically upgrade what is most abundantly available (or downgrade the relatively unavailable).
[33] Elster 1979, p. 40.
[34] See Kolm 1979.
[35] Leibniz 1875–90, vol. V, p. 175.

the latter preferences should be made into the basis for social choice, and social choice would not be adequate without such a basis. This is, of course, subject to the qualifications mentioned above: the new preferences should not be reversible simply by making the preferred object inaccessible, and the need for knowledge may be overridden by the need for substance of character. Nor does precommitment create any difficulties. If the wants are prior to and actually shape the feasible set, then the coincidence of aspirations and possibilities is in no way disturbing. As to the deliberate (exogenous) manipulation of wants, it can be condemned out of hand on grounds of autonomy, and possibly on grounds of welfare as well.

Hard problems remain, however, concerning the relation between misperception of the situation and misformation of the preferences. Consider again the alternative interpretation of the Industrial Revolution, in terms of excessive anticipations rather than of rising aspirations. From the work of Tocqueville, Merton and Veyne, it would appear that below a certain threshold of actual mobility, expected mobility is irrationally low, in fact zero. Above this threshold, expected mobility becomes irrationally high, close to unity. And so, in society with little actual mobility, preferences may adapt to the perceived rather than to the actual situation, a contributing factor to what I have called overshooting or over-adaptation. Similarly, once a society has passed the mobility threshold, irrational expectations are generated, with a corresponding high level of wants. The intensity of the desire for improvement grows with the belief in its probability, and the belief in turn through wishful thinking feeds on the desire.

This view, if correct, implies that one cannot sort out in any simple way the frustration due to irrational expectations from the one due to a new level of aspirations. Let us imagine, however, that there was no tendency to wishful thinking. Then the actual and the expected rates of mobility would coincide (or at least not differ systematically), and the rational expectation would then generate a specific intensity of desire or aspiration level, with a corresponding level of frustration. The utilitarian might then want to argue that in this counterfactual state with rational expectations there would not be generated so much frustration as to make people actually worse off after the improvement in their objective situation. I am not certain that this is a relevant counterargument, for should one's acceptance of utilitarian theory turn upon empirical issues of this kind? And in any case I am not sure that the counterfactual statement is in fact true. Even when one knows that there is only a modest probability that one will get ahead, it may be sufficient to induce a state of acute dissatisfaction. But I have less than perfect confidence in both of these replies to the utilitarian counterargument, and so there is a gap in my argument. I leave it to the reader to assess for himself the importance of the difficulty.

The criticism I have directed against utilitarian theory is, essentially, that it takes account of wants only as they are *given*, subject at most to a clause about the need for learning about the alternatives. My objection has been what one might call 'backward-looking', arguing the need for an analysis of the *genesis* of wants. Before I spell out some methodological implications of this objection, I would like to point out that the assumption of given wants may also be questioned from two other directions, which for mnemonic purposes I shall call 'upward-looking' and 'forward-looking' respectively.

The language of directions suggests that preferences may be viewed along two dimensions. One is the temporal dimension: the formation and change of preferences. The other is a hierarchical dimension: the ranking of preferences according to higher-order preferences. If, in addition to information about the first-order preferences of individuals, we have information about their higher-order preferences, we may be able to get out of some of the paradoxes of social choice theory. This approach has been pioneered by Amartya Sen.[36] For some purposes this 'upward-looking' correction of preferences may be useful, but it can hardly serve as a general panacea.

Preferences, however, may also be corrected in a more substantial manner. Instead of looking at politics as the *aggregation* of given preferences, one may argue that the essence of politics is the *transformation* of preferences through public and rational discussion. This 'forward-looking' approach has been pioneered by Jürgen Habermas in numerous recent works. On his view, the multifarious individual preferences are not a final authority, but only idiosyncratic wants that must be shaped and purged in public discussion about the public good. In principle this debate is to go on until unanimity has been achieved, which implies that in a rationally organised society there will be no problem of social choice as currently conceived. Not optimal compromise, but unanimous agreement is the goal of politics. The obvious objection is that unanimity may take a long time to achieve, and in the meantime decisions must be made – and how can we then avoid some kind of aggregation procedure? In addition the unanimity, even if sincere, could easily be spurious in the sense of deriving from conformity rather than from rational conviction. There is no need to assume force or manipulation as the source of conformity, for there is good psychological evidence that a discordant minority will fall into line simply to reduce dissonance.[37] Habermas assumes crucially that in the absence of force rationality will prevail, but this is hardly borne out by the facts. I have argued that the containment of wants within the limits

[36] Sen 1974; 1977b.
[37] Asch 1956.

of the possible should make us suspicious about their autonomy, and similarly I believe that unanimity of preferences warrants some doubts about their authenticity. This implies, at the very least, that the forward-looking approach must be supplemented by the backward-looking scrutiny. The end result of unanimity does not in itself ensure rationality, for we must also ascertain that agreement is reached in an acceptable way.

The backward-looking approach in all cases involves an inquiry into the history of the actual preferences. One should note, however, that there are other ways of taking historical information into account. Thus we may make present decisions a function of present and past preferences, rather than of present preferences together with their past history. The rationale for using sequences of preferences as input to the social choice process could only be that they would somehow capture the relevant historical aspects of present preferences, and this they might well do. Persons tending to have adaptive preferences might be detected if they exhibit systematic variation of preferences with changing feasible sets. But the correlation would at best be a crude one, since the tendency towards adaptive preferences need not be a constant feature of a person's character.

The backward-looking principle is one of *moral hysteresis*.[38] Since information about the present may be insufficient to guide moral and political choice in the present, we may have to acquire information about the past as well. In Robert Nozick's terminology, I have been engaged in a polemic against end-state principles in ethical theory.[39] In Nozick's own substantive theory of justice, we need information about the historical sequence of transfers in order to determine what is a just distribution in the present. In Marxist theories of justice we also need to go beyond present ownership of capital goods, in order to determine whether it is justified by past labour.[40] And Aristotle argued that in order to blame or condone actions in the present, it is not enough to know whether the person was free to do otherwise in the present: we also need to know whether there was freedom of choice at some earlier stage.[41] In the present article, I have raised a more elusive problem, the historical dimension of wants and preferences. Adaptive preference formation is relevant for ethics, and it is not always reflected in the preferences themselves, and so it follows that ethics needs history.[42]

[38] Elster 1976 has a discussion of the more well-known notion of causal hysteresis.
[39] Nozick 1974, pp. 153 ff.
[40] Elster 1978b,c.
[41] *Nicomachean Ethics*, 1114a.
[42] This conclusion parallels the conclusion of my forthcoming essay on 'Belief, bias and ideology': 'Since epistemology deals with the rationality of beliefs, and since the rationality of a belief can neither be read off it straight away nor be assessed by comparing the belief with the evidence, we must conclude that epistemology needs history.'

12 Liberty and welfare

ISAAC LEVI

According to A. K. Sen, liberalism (or 'libertarianism' as he now prefers to call it[1]) permits each individual in society 'the freedom to determine at least one social choice, for example having his own walls pink rather than white, other things remaining the same for him and the rest of society'.[2]

Sen contends that the value involving individual liberty illustrated by this example imposes a constraint on social welfare functions – i.e. rules which specify a ranking of social states with respect to whether they serve the general welfare better or worse given information about the preferences of individual members of society for these social states or their welfare levels in these social states. This constraint 'represents a value involving individual liberty that many people would subscribe to' regardless of whether it captures all aspects of the presystematic usage of the terms 'liberalism' or 'libertarianism'.[3]

Sen's condition *L* asserts that each citizen ought to have his preference ranking of at least one pair of social states determine the social ranking of the same pair of states with respect to welfare.[4]

P. Bernholz pointed out that libertarians do not concede individuals rights to determine the social ranking of social states but to determine aspects of social states.[5] P. Gärdenfors has recently combined Bernholz's observation with R. Nozick's suggestion that granting rights to individuals cedes to them the ability to constrain the domain of social choice to a given class of social states.[6]

In his interesting discussion of Nozick's idea, Sen points to an ambiguity in the interpretation of a social ordering. He suggests that a social ordering can be construed 'to be purely a mechanism for choice' or as 'reflecting a view of social welfare'.[7]

[1] Sen 1976, p. 218.
[2] Sen 1970b, p. 153.
[3] *loc.cit.*, n1.
[4] *loc. cit.*
[5] Bernholz 1974, pp. 100–1.
[6] Gärdenfors 1978; and Nozick 1974, pp. 165–6.
[7] Sen 1976, pp. 229–31.