Introduction

JON ELSTER AND AANUND HYLLAND

Any body of scientific knowledge may grow in two directions. On the one hand it may expand upwards and outwards, adding ever more branches to a common trunk of assumptions and axioms. On the other hand it may endeavour to strike its roots ever more deeply, or, to change the metaphor, to secure more solid foundations for the superstructure. In every discipline, these seem to be alternate stages of development. After a period of expansion, there follows a period of consolidation and rethinking that in turn enables expansion along new lines. The inward-looking turn can be provoked in several ways. The theory may encounter anomalies, counter-intuitive or counter-empirical results. It may suffer decreasing marginal productivity. Or it may increasingly lose touch with the problems that originally constituted its *raison d'être*. This last danger is especially acute in the formal sciences, like mathematics. As John von Neumann once remarked, mathematics out of touch with physical sciences tends to become *baroque*, a term used in contrast to the *classical* style of thinking that is constantly revitalized by contact with the empirical sciences.

Formal theorizing in the social sciences is today in some danger of becoming baroque. A frequent scenario seems to be the following. In a first stage, there exists a theoretical problem with immediate economic, social or political significance. It is, however, ill-understood, perhaps even ill-defined. In the second stage, a proposal is put forward to conceptualize the problem in a way that dispels confusion and permits substantive conclusions to be drawn. In a third stage the conceptual apparatus ceases to have these liberating effects, and becomes a new, independent source of problems. An illustration could be the notion of supply and demand schedules. Before the invention of this simple diagrammatic device, the notion of a *change* in supply or

demand was hopelessly confused. With this tool, on the other hand, the distinction between price-induced changes and changes induced by a shift in the schedules followed immediately. In a further stage, the refinement of the concepts led to the propositions of general equilibrium theory, which has at best a very tenuous connection with actual market operations.

Social choice theory sprang from two distinct problems: that of finding an adequate, robust voting system, and that of finding a measure for aggregate social welfare. The very idea that there was a close relation between these problems was not clearly perceived. As usual, a notational breakthrough was decisive. By stating the problem as one of finding a *function from a set of individual preferences to a social preference order*, connections could be seen and precise questions could be asked that earlier were only an inchoate possibility. Following Kenneth Arrow's pioneering work, major results were proved at a gratifying rate. Today, social choice theory may be approaching the baroque stage. Breakthroughs are dwindling, while minor embellishments are accelerating. Formalism is gaining the upper hand, as in what Ragnar Frisch used to refer to as 'playometrics'. From a means, formal modelling is becoming an end in itself. These comments are not intended as dismissive of the extensions and refinements that are currently being produced. Indeed, in many cases it is only by stretching the concepts to the limits that their weak points appear. The interplay between intensive and extensive development is part and parcel of scientific development, and it would be absurd to assign one to a higher position than the other. The time may nevertheless be appropriate for a new look at the foundations of social choice theory. The present volume, to be sure, can only survey a few of the issues, and, moreover, very inadequately. It is hoped that it will contribute as much to shifting the emphasis of discussion towards foundational issues as to illuminating the specific problems raised in the various contributions.

When discussing foundations, one may proceed in several ways. First, one may examine the relation between a given discipline and adjoining fields, in the hope that generalizations and simplifications will emerge. Secondly, one may concentrate on the substantive interpretation of the formal results, with a view to finding inspiration for new theoretical developments. Thirdly, one may reconsider the basic

assumptions—implicit and explicit—of the theory, in order to find and strengthen the weak links in the chain. The contributions to the present volume illustrate the second and third procedures, while the first is largely neglected. In what follows, the first and second are covered rather briefly, while the third is the subject of somewhat more extensive comment. In particular, we discuss some possible *self-supporting features* of aggregation mechanisms.

Lateral connections

Social choice theory has obvious links to cooperative and non-cooperative game theory. Also, the connection with theories of distributive justice is increasingly intimate. The structure of the theory has been considerably clarified by the explicit consideration of these connections.

Social choice theory and cooperative game theory have many common features. In particular, n-person bargaining theory and social choice theory are both concerned with deriving a social outcome from individual preferences, with Pareto optimality as one main constraint. The arbitrator (in bargaining theory) or the constitutional designer (in social choice theory) ask themselves what Pareto-optimal outcome ought to be preferred, given what they know about individual preferences and their notions of how the outcome ought to reflect or respect them. These notions have, in both theories, an essential counterfactual element. They do not simply postulate relations between the actual preferences of the actual individuals over the actual alternatives, on the one hand, and the socially preferred outcome on the other. They also impose consistency conditions on how the preferred outcome ought to change when we vary the preferences of the individuals, the number of individuals, or the set of feasible alternatives. Such consistency is a necessary if not sufficient condition of both rationality and morality.

These broadly similar features are overshadowed, however, by the many dissimilarities in the framework of the theories. Bargaining theories impose more structure on the problem than the standard Arrowian social choice theory, with respect both to the feasible set and to the individual preferences. Also, the outcome to a bargaining problem is supposed to reflect the bargaining strength of the partici-

pants, i.e. what would happen to them if they failed to reach agreement. There is no analogous feature in the formulation of the social choice problem.[1]

Non-cooperative game theory seemed rather foreign to Arrow's original statement, but has recently become increasingly central. This shift in emphasis went together with increasing attention to the fact that expressing inputs to the social decision mechanism is an *action*, one which is guided by the preferences of the individual. He may or may not find it in his interests, as defined by his 'real preferences', to use the latter as his input to the decision process. Clearly, the question of what preferences to express is a problem for non-cooperative game theory, since the answer may depend on what other people can be expected to do. Specifically, one may ask under what circumstances, if any, honest voting can be a dominant strategy; more weakly, whether the voting game has equilibria in which all players express their real preferences; moreover, whether the solution to the game is an equilibrium point of this type.[2] Similarly, whether the agents should use or waive their *rights* turns out to be an important question in some formulations of the Liberal Dilemma. This is also a game-theoretic issue, since the decision may depend on whether others can be expected to waive *their* rights.

Arrow's theory has been characterized by himself as 'ordinal utilitarianism', i.e. a conception of distributive justice. It may then be usefully contrasted with other conceptions of this kind, e.g. classical utilitarianism and Rawls's difference principle. One of the more important developments in social choice theory has been the proof that by an extension of its framework both the latter theories can be subsumed as special cases. Specifically, by allowing more utility information than is possible within the ordinal framework, one can give axiom sets that uniquely characterize the average utility or the minimum utility as the proper maximand for the social welfare function.[3] This broad group of conceptions may be contrasted with those which use bargaining theory as the proper framework for settling issues of distributive justice.[4]

This cluster of theoretical clusters—games, bargaining, justice, and

[1] For discussions of the relation between bargaining theory and social choice theory, see Luce and Raiffa (1957, pp. 349 ff.) and Sen (1970, Ch. 8).
[2] See Dasgupta, Hammond, and Maskin (1979) for an overview.
[3] See d'Aspremont and Gevers (1977).
[4] See, for instance, Gauthier (forthcoming).

social choice—exhibits all sorts of intricate internal connections. Problems from one cluster can often be reformulated in the language of another. Yet, we are a far cry from the transparency and simplicity that would characterize a truly unified theory. Indeed, unification may be a will-o'-the-wisp, although one may at least expect partial simplifications and generalizations to occur as the result of further cross-disciplinary work.

Interpretations

Any formal structure can be interpreted in several distinct ways. One specific interpretation is often that intended by the creators of the structure. Later, it is seen that other, substantively different problems can be subsumed under the same set of axioms. These problems may suggest additions to or modifications of the formal structure, and thus offer new insight into the original set of problems. This interplay between substantive issues and formal modelling is important in securing better foundations for the theory.

In the social sciences and in philosophy there is one particular instance of this interplay that has been especially important. When studying some kind of interaction *between* persons, from a normative or a behavioural point of view, it is often instructive to look into whether a similar interaction can be observed *within* a single person. It will often be found that the concepts developed for the interpersonal case also apply to the intrapersonal one, but not in a wholly unstrained way. To improve the fit, changes are needed that, on further reflection, may also throw light on the original interpersonal problem.

In social choice theory this extension to intrapersonal cases can occur in several ways. One is instantiated in the contribution of Ian Steedman and Ulrich Krause to *The Multiple Self*, a companion volume to the present work. They observe that an individual often evaluates a choice situation from many different perspectives, each of which has associated with it a preference ordering of the options. My egoistic, altruistic, moral and social selves may rank the alternatives differently, so that the need arises for some aggregation mechanism. Some features of standard social choice theory appear somewhat differently in this setting. For instance, there might not be anything objectionable in having a dictator. We might want to say that the moral

preference structure ought to win out in all cases. Yet, since we know that this outcome is far from always observed in reality, we may reformulate the question as a second-best problem: given that some intrapersonal democracy is unavoidable, how ought it to be set up? Another intrapersonal extension is to intertemporal decisions. We may conceive of each time slice of the individual as a separate 'self' with interests that range forward into the future and backward into the past. This procedure would enable us to ask questions such as the following: Ought there to be a condition of 'liberalism' that would block preferences among past options to influence the choice between them? That is, given two consumption streams that are identical from time t onwards, should we allow post-t selves any say in defining the preference between them?

In addition to the distinction between interpersonal and intrapersonal applications, there is the better-known distinction between interprofile and intraprofile applications. The former involve consistency conditions of a counterfactual sort, whereas the latter consider just one actual preference profile. Impossibility theorems may be proved for both cases,[5] although of course one would have to be careful when interpreting the terms. For instance, the existence of a 'dictator' has a very special meaning in the intraprofile case, since it cannot have any implication about *power*, which is an essentially counterfactual notion. By contrast, in the interprofile case a dictator is defined as someone who gets his way whatever his preferences and those of all others might happen to be. By exploring partial analogies of these kinds, we acquire a firmer grasp of what was involved in the original theory, and may even be motivated to modify it to incorporate some of the features from the analogous cases.

Assumptions

In social choice theory under the standard (interpersonal and interprofile) interpretation, various assumptions are made explicitly or implicitly. Here we focus on some of the unstated assumptions, to see whether an explicit consideration can suggest some further lines of research. Most importantly, the framework takes a number of features as exogenously given that are in reality the outcome of complex

[5] Roberts (1980).

economic, social and political processes. Also, whether the social decision mechanism ought not itself to be socially chosen is a question frequently asked, but rarely answered.

Among the features usually taken as given are (i) the set of individuals, (ii) their preferences, (iii) the options confronting them, and (iv) the rights assigned to them. In the present volume, the contributions by Elster and Goodin consider in some detail the question of preference formation and preference filtering. The question of endogenous agenda formation is receiving a good deal of attention in the literature on voting systems. This is closely related to the problems of strategic expression of preferences. (Given the Borda rule, for example, one may distort the decision process either by misrepresenting one's preferences or by adding items to the agenda.) In what follows we consider the endogenous formation of the political unit, the endogenous formation of rights, and the endogenous emergence of the social choice mechanism itself. The discussion of these various questions is not, of course, presuppositionless. It rests on further assumptions that may well be questioned. Also, it is formulated within a less general framework than the standard Arrowian theory. Since the intention is mainly to illustrate a line of argument, this does not matter too much.

Much of politics concerns the distribution of burdens and benefits within a given political unit. Throughout history, however, the question of where to draw the boundaries between political units has also been of overwhelming importance. For social choice theory the problem is whether it is possible to suggest a normative justification for drawing these boundaries in one way rather than in another. We approach this question somewhat indirectly. We first mention a couple of respects in which a political system might be said to be self-supporting or self-justifying, and then go on to ask whether a similar argument could hold for the drawing of boundaries.

Consider first the question of suffrage. In any polity, some individuals lack the right to vote: children, criminals, foreign citizens, etc. The question then arises as to how the right to vote is to be assigned. There may not exist a general answer, but in one particular case it seems possible to use consistency arguments to reach a solution. This is the case of voting age. Let us assume that we ask all individuals more than x years old what, in their opinion, ought to be the voting age.

Specifically, for each n we ask them whether persons at least n years old ought to have the right to vote. The smallest n such that there is a simple majority to let people at least n years old have the right to vote, we call $f(x)$. It is highly likely that, say, $f(6)>6$ and $f(50)<50$. Also, it is probable that for all x, $f(x+1) \geqslant f(x)$. This ensures that there is some x such that $f(x)$ approximately equals x. We submit that the voting age ought to be the smallest x with this property. The proposal assumes that democracy ought to be extended as far as can consistently be defended. Thus, if in the group above 16 years there is a majority against letting 16-year-olds have the right to vote, this ought not to be chosen as the voting age. Conversely, one ought not to set the voting age at 21 if there is a majority among persons above 18 years of age for the view that people above 18 ought to have the right to vote.

Consider next the question of majority voting. For various reasons one may sometimes desire to use qualified majority voting on fundamental constitutional issues. The obvious question is: how does one decide which majority to require? In the light of the preceding paragraph, the answer should be clear. At any given time there is a well-defined percentage $g(x)$ of the voters that want the majority required for a certain constitutional change to be at least x per cent. $g(x)$ must be a decreasing function of x; also we may assume $g(50)>50$ and $g(100)<100$. Then, again, we ought to choose a self-supporting majority: the unique $\bar{x}$ such that $g(\bar{x}) = \bar{x}$.

Against the background of these two examples, we may consider the question of boundary-drawing. The obvious suggestion now is the following. One ought to define as one political unit the largest territorial unit such that within the unit there is a simple majority for considering it as one voting community rather than several. It may also be obvious why this proposal doesn't work. Unlike the other examples, one cannot make the crucial monotonicity assumption that allows the argument to go through. One cannot assume, that is, that the feeling of belonging to one community becomes progressively weaker as one moves outward from the centre. The following example suffices to prove the point:

> Consider, for example, the Irish question as it stood between 1918 and 1922. Simplifying somewhat, there was (probably) a majority in the U.K. as a whole (i.e. the British Isles) for the maintenance of the

union; within Ireland (i.e. the whole island) there was a majority in favour of independence for the whole of Ireland; within the six provinces that became Northern Ireland there was a majority for partition as a second best to union; but within two of those six counties there was a majority for unification with the south as a second best to independence for the whole of Ireland.[6]

In cases such as these, the only solution may be to assign rights to the geographically dispersed minority group. The protection of these rights, moreover, could be ensured by constitutional guarantees that make them very hard to abolish. Yet once again we encounter a difficulty. Rights must be *assigned* by someone; they do not fall from heaven. By whom, and by what procedure, ought they to be assigned? One answer has already been suggested. For a right to receive strong constitutional guarantees, it must be strongly supported in the population. If people feel strongly about rights that a majority is not prepared to uphold and if, moreover, the minority is geographically dispersed, emigration and civil war are the only options. And of these, the former may not be available.

The preceding analysis was deliberately brief and incomplete. The intention was neither to present a theory of self-supporting social institutions, nor to explain the endogenous emergence of the variables that the standard theory takes as exogenously given. The suggestions developed above are offered mainly with a view to illustrating one line of foundational research, which proceeds by moving one step further back in the analysis.

[6] Barry (1979, p. 169).

REFERENCES

d'Aspremont, C. and Gevers, L. (1977) 'Equity and the informational basis of collective choice', *Review of Economic Studies* 44, 199–209.

Barry, B. (1979) 'Is democracy special?', in P. Laslett and J. Fishkin (eds.), *Philosophy, Politics and Society*, 5th series, Oxford: Blackwell.

Dasgupta, P., Hammond, P., and Maskin, E. (1979) 'The implementation of social choice rules: some general results on incentive compatibility', *Review of Economic Studies* 46, 185–216.

Gauthier, D. (forthcoming) 'Bargaining and justice'. To appear in *Social Philosophy and Policy*.

Luce, R. D. and Raiffa, H. (1957) *Games and Decisions*, New York: Wiley.

Roberts, K. (1980) 'Social choice theory: the single-profile and multi-profile approaches', *Review of Economic Studies* 47, 441–50.

Sen, A. K. (1970) *Collective Choice and Social Welfare*, San Francisco: Holden Day.

Steedman, I. and Krause, U. (1985) 'Goethe's Faust, Arrow's possibility theorem and the individual decision taker', in J. Elster (ed.), *The Multiple Self*, Cambridge: Cambridge University Press.