CHAPTER 1

Prologue

> [T]he social order is a sacred right which serves as a foundation for all other rights. This right, however, since it comes not by nature, must have been built upon convention. To discover what these conventions are is the matter of our inquiry.
>
> Jean-Jacques Rousseau*

1 The problem of rule selection

In trying to lay the foundations of the new welfare economics, J. R. Hicks brought forward the following problem:

> Although the economic system can be regarded as a mechanism for adjusting means to ends, the ends in question are ordinarily not a single system of ends, but as many independent systems as there are "individuals" in the community. This appears to introduce a hopeless arbitrariness into the testing of efficiency [of any given economic organization]. You cannot take a temperature when you have to use, not one thermometer, but an immense number of different thermometers, working on different principles, and with no necessary correlation between their registrations. How is this difficulty to be overcome? [Hicks, 1939, p. 699]

According to Hicks, there are three possible ways of dealing with this problem:

1. Instead of using the preference scales of the individuals in the community, the investigator may use his own "thermometer," that is, decide for himself what is good or bad for society in judging the relative performances of the alternative economic organizations.†
2. The investigator may seek a method of aggregating the (possibly) conflicting reports of the various thermometers so as to construct an "average" or "social" registration.

* J.-J. Rousseau, *The Social Contract or Principle of Political Right*. In *Social Contract: Essays by Locke, Hume and Rousseau* (with an introduction by Sir Ernest Barker). London: Oxford University Press, 1947, p. 240.

† To avoid awkward wording, the pronoun "he" will often be used in the generic sense to mean "he or she."

3. The investigator may focus attention on those cases in which the difficulty due to the heterogeneity of individual preferences may be circumvented by allowing *hypothetical* compensation payments between gainers and losers.

The first of these methods was rejected by Hicks because of its paternalistic and "unscientific" flavor, whereas the second method was identified with the traditional method of Marshall, Edgeworth, and Pigou, of "weighting" the component parts under the assumption of interpersonal comparability of welfare units, and was likewise rejected on the Robbinsian ground that we had no "scientific" means for interpersonal welfare comparisons. This left only the third method, which Hicks did adopt, and the foundations of the new welfare economics were based on this method.

Soon after removal of the scaffold for construction, however, demolition activity began. The hypothetical compensation principles proposed by Pareto (1909), Barone (1908), and Kaldor (1939), and used extensively by Hicks (1939, 1940), were found to contain grave logical difficulties.[1] The very foundations of the new welfare economics were thereby revealed to be quite shaky. Because Hicks's third method involving the hypothetical compensation principle turned out to be will-o'-the-wisp, we must once again face up to the Hicksian problem of making social judgments in the face of heterogeneous individual preferences.

As a matter of fact, the second method (which Hicks immediately threw away), which involves constructing a rule that aggregates individual preferences, commands more careful scrutiny, because the weighted-sum-of-utilities method is just a particular instance thereof, and the class of possible rules is very broad indeed. It was Arrow (1950, 1963, 1967*a*, 1967*c*) who formalized this preference aggregation problem in a way that was in concordance with the Robbins–Hicks–Allen *ordinalist* approach to individual utility and welfare. His studies had an even more devastating effect on the prior effort to lay the foundations of welfare economics, because Arrow's central conclusion was that *there cannot possibly exist a "rational," "efficient," and "democratic" rule of preference aggregation.* How is this difficulty to be overcome? The following chapters will examine this and related problems.

The first order of business is to further clarify the nature of our problem. We have said that we want to construct a rule that aggregates individual preferences with a view toward forming social judgments regarding the comparative virtues of alternative economic organizations. A point to keep in mind is that this rule selection or system design should in itself be a matter of concern for the individuals in the society. The reason is that an acknowledged rule associates to each possible set of individual preference orderings a social choice procedure for selecting a socially preferred decision out of any set of feasible alternatives. Now, "the principles of practical reasoning for an entity capable of decision are to be decided upon by that entity. Thus the principle of choice for a rational individual is his to make; but, similarly, the principles of social choice must be adopted by the association itself, that is, by the individuals that constitute it" (Rawls, 1968, p. 69). This granted, we seem to be trapped by the prima facie infinite regression: We need a rule for aggregating individual preferences over primitive social states, and this presupposes a higher-order rule for aggregating individual preferences over alternative rules of the first order, and so on ad infinitum. How can we break this impasse?

The clue is found in the fact that it is much easier to propose and then accept or reject an abstract performance criterion of a rule (such as that the outcome generated by a rule should not invidiously favor or disfavor any individual, because otherwise those who feel that their wishes are unfairly treated by the rule will have legitimate resentment against the social decisions thereby arrived at and will fail to be motivated to cooperate with others to attain the socially "best" position identified by that rule), even though preferences among rules (of any order) may be prohibitively difficult to form in the vacuum. Motivated by this simple observation, we proceed as follows.

Let each individual propose the performance criteria that he wishes the rule to satisfy, with the understanding that the rule acknowledged now will be binding on indefinite future contingencies, the peculiarities of which cannot be known in the primordial stage of rule selection. Each individual recognizes that he is required, along with everybody else, to make a firm commitment in advance. Because each individual is deprived of any opportunity to tailor the rule to favor him if it is adopted and to reject the rule if it comes to act to his disadvantage, it is likely that everyone will have every incentive to propose performance criteria of a general nature.[2]

The next step is to single out certain performance criteria from among those that were initially proposed, the selection being made on the basis of unanimous acknowledgment by all individuals. Given this set of unanimously acknowledged criteria, an investigation should then be made to identify those rules for social choice that are thereby qualified.

A point of some importance is that a rule need not necessarily be unanimously acknowledgeable simply because each one of its component axioms elicits agreement among individuals in isolation. As Samuelson (1977, p. 85) observed, the classic phrase "by their fruit must ye know them" applies above all to axiom systems, and the joint effect of a set of individually appealing axioms on a rule might turn out to yield an unappealing rule, one eventually unacceptable to the participating individuals. Even worse, a combination of individually innocent and persuasive axioms might annihilate *all* rules, as Arrow's general impossibility theorem exemplifies.

Let us notice that the message of these paradoxes and logical contradictions need not necessarily be negative. They may serve as signals that force us to reexamine the individual axioms (performance criteria) and, going back one step further, the conceptual framework in which these axioms are phrased. It is to be hoped that we can thereby penetrate to a more appropriate understanding of the factors that are responsible for the institutional stability of voluntary associations of free individuals. This is the central frame of thought that guides us throughout this work.[3]

2 Plan of the book

Following this introductory chapter and its Appendix, which expounds an elementary theory of binary relations, there are seven chapters that constitute the main body of this book. In Chapter 2 we introduce the formal concept of a choice function and the rationality thereof, with a view toward characterizing the concept of a rational choice function in terms of the revealed preference and related axioms. In the main body of the chapter we consider this rationalization problem, which is the choice-functional counterpart of the integrability problem in demand theory, when the domain of a choice function contains all finite subsets taken from the universal set. This restriction is imposed because of the intended application of this theory in the rest of the book. In the Appendix to Chapter 2 we develop a more general theory of a rational choice function.

In Chapter 3 we set about analyzing the problem of the primordial rule selection, and we examine Arrow's general impossibility theorem on the "rational," "efficient," and "democratic" collective choice rules. Special care will be taken here with the robustness of the Arrovian impossibility theorems with respect to the successive weakening of the collective rationality requirement.

In Chapter 4 we focus our attention on a particular class of rules: The simple majority decision rule and its extensions.

Up to this point, our investigation into the structure of a social choice rule makes use only of an informational basis that excludes any form of interpersonal welfare comparisons. In the rest of this book, we examine the effect of widening the informational basis for collective choice. Instead of confining our attention to the *intrapersonally ordinal* and *interpersonally noncomparable* preference orderings by each and every individual, we now ask each individual to put himself in the position of each other individual through the imaginary exchange of circumstances to determine if that is a better or worse position than his actual position. It can hardly be denied that individuals may make, and do make, interpositional comparisons of this type, and this provides us with one operational interpretation of what one may mean by interpersonal comparisons of welfare.

In Chapter 5 we formulate a social choice version of the "equity-as-no-envy" approach of Foley (1967), Varian (1974, 1975, 1976), and others, and in Chapter 6 we examine two versions of the constrained majoritarian social choice rules. Both chapters make extensive use of the interpersonal comparisons of welfare mentioned earlier and try to gauge their effects on the Arrovian impossibility theorems.

In Chapter 7 we examine another class of impossibility theorems and the resolutions thereof. The impossibility theorems in question maintain the incompatibility of democratic values and libertarian claims and assert that unadulterated exercise of libertarian rights, coupled with mechanical use of the Pareto unanimity principle, may disqualify all collective choice rules with unrestricted domain. Two resolution schemes will be presented, the first of which restricts mechanical use of the Pareto unanimity principle, whereas the second restricts the exercise of rights by the impartial principles of justice. It is clear that they differ in their informational requirements as well as in their basic attitudes toward libertarian rights and democratic values. Nevertheless, they seem to share a common general moral: The ultimate guarantee of a minimal level of libertarian rights in a democratic society lies in an attitude whereby individuals respect and care for equal basic liberty and justice for one another in the conflict situation.

Finally, Chapter 8 summarizes the broad implications of our analysis and puts forward several qualifications.

3 Concluding remarks

Before concluding this introductory chapter, two final remarks are necessary regarding terminology. First, in the theory of collective choice and social welfare, the term "individual" may be interpreted flexibly, depending on the context. In some contexts it may mean a human individual, whereas in others it may refer to a nation, a social group within a large committee, a team, and so on. In the main, we shall refer to "individual" as if we mean human individual, but the theory makes good sense under the different interpretations of the term.

Second, a "social state" is a primitive concept of our analysis; it can stand for almost anything that can be construed as the outcome of a social decision process. Depending on the context, it may denote "a set of bills passed and bills failed" (Plott, 1972, p. 84) or "a vector whose components are values of the various particular decisions actually made by the government, such as tax rates, expenditures, antimonopoly policy, and price policies of socialized enterprise" (Arrow, 1963, p. 87) or "a complete description of the amount of each type of commodity in the hands of each individual, the amount of labor to be supplied by each individual, the amount of each productive resource invested in each type of productive activity, and the amounts of various types of collective activity" (Arrow, 1963, p. 17), and so on. It is hoped that the reader will keep in mind this flexibility of interpreting the concept of a social state.

Appendix: Elementary properties of binary relations

1. In the whole body of this book, we shall make extensive use of the theory of binary relations in general, as well as the extension theorems for binary relations in particular. Let us therefore gather some elementary properties of binary relations in a form that is convenient for later reference. The following notation will be used throughout.

$\exists$	existential quantifier ("there exists")
$\forall$	universal quantifier ("for all")
$\rightarrow$	implication ("if ..., then ...")
$\leftrightarrow$	equivalence ("if and only if")
$\sim$	negation ("not")
$\&$	conjunction ("and")
$\vee$	alternation (the inclusive "or")
$=$	identity
$x \in A$	x belongs to (is an element of) A
$x \notin A$	x does not belong to (is not an element of) A
$A \subset B$	A is contained in (is a subset of) B
$A \supset B$	A contains (is a superset of) B
$A \subset\subset B$	A is properly contained in B ($A \subset B \,\&\, A \neq B$)
$\{x \in A \mid P(x)\}$	set of all $x \in A$ having a property P
$A \cap B$	intersection of A and B (elements belonging to both A and B)
$A \cup B$	union of A and B (elements belonging to either A or B)
$A \times B$	Cartesian product of A and B, namely the set of ordered pairs (x, y) such that $x \in A$ and $y \in B$
$A \setminus B$	set-theoretic difference, namely the set of points x such that $x \in A$ and $x \notin B$
$\mathscr{P}(A)$	power set of A, namely the set of all subsets of A
Z^+	set of all positive integers
$\varnothing$	empty set

Other symbols will be explained when necessity dictates.

2. Let X be the universal set of our discourse. A *binary relation* on X is a proposition (R) such that, for any ordered pair (x, y), where $x, y \in X$, we can unambiguously say that (R) is true or false. The set of all ordered pairs (x, y) that satisfy (R), to be written as R, is then a well-defined subset of $X \times X$. Conversely, given a set $R \subset X \times X$, we can define a binary relation (R) by

$$(R) \text{ is true for } (x, y) \leftrightarrow (x, y) \in R$$

We can thus identify a binary relation on X with a subset of $X \times X$. It is customary to write, for all $x, y \in X$, xRy if and only if $(x, y) \in R$ holds true. In this Appendix (and, indeed, in most of this book) we shall use the set-theoretic notation for the sake of convenience of mathematical operations on a class of binary relations. We shall have more to say on this matter in the final section.

Given a binary relation R on X, we define the *symmetric part* $I(R)$ and the *asymmetric part* $P(R)$ thereof by

$$I(R) = \{(x, y) \in X \times X \mid (x, y) \in R \,\&\, (y, x) \in R\} \tag{A.1}$$

and

$$P(R) = \{(x, y) \in X \times X \mid (x, y) \in R \,\&\, (y, x) \notin R\} \tag{A.2}$$

respectively. Clearly, $I(R) \cap P(R) = \varnothing$ and $I(R) \cup P(R) = R$ hold true for every R.

A notable example of a binary relation is what we call a *weak preference relation* R of an agent, which is defined on a set X of all conceivable options by

$$(x, y) \in R \leftrightarrow \text{According to the agent's view, } x \text{ is at least as good as } y$$

$I(R)$ and $P(R)$ will then denote the *indifference relation* and the *strict preference relation*, respectively, of this agent. To the extent that it may facilitate understanding, we provide interpretations of a variety of properties of a binary relation to be introduced later in terms of a weak preference relation, although these properties can apply to any binary relation we may specify.

3. Let us now enumerate several properties of a binary relation R that have been found relevant in various contexts of social choice theory.[4]

(a) *Completeness*

$$\forall x, y \in X: x \neq y \rightarrow (x, y) \in R \vee (y, x) \in R$$

(b) *Reflexivity*

$$\forall x \in X: (x, x) \in R$$

(c) *Irreflexivity*

$$\forall x \in X: (x, x) \notin R$$

(d) *Transitivity*

$$\forall x, y, z \in X: [(x, y) \in R \,\&\, (y, z) \in R] \rightarrow (x, z) \in R$$

(e) *Quasi transitivity*

$$\forall x, y, z \in X: [(x, y) \in P(R) \,\&\, (y, z) \in P(R)] \rightarrow (x, z) \in P(R)$$

(f) *Triple acyclicity*

$$\forall x, y, z \in X: [(x, y) \in P(R) \,\&\, (y, z) \in P(R)] \rightarrow (z, x) \notin P(R)$$

(g) *Acyclicity*

$$\forall t \in Z^+, \forall x^1, x^2, \ldots, x^t \in X: [\forall \tau \in \{1, 2, \ldots, t-1\}: (x^\tau, x^{\tau+1}) \in P(R)] \rightarrow (x^t, x^1) \notin P(R)$$

(h) *Consistency*

$$\forall t \in Z^+, \forall x^1, x^2, \ldots, x^t \in X: [(x^1, x^2) \in P(R) \,\&\, \forall \tau \in \{2, 3, \ldots, t-1\}: (x^\tau, x^{\tau+1}) \in R] \rightarrow (x^t, x^1) \notin R$$

(i) *Symmetry*

$$\forall x, y \in X: (x, y) \in R \rightarrow (y, x) \in R$$

(j) *Asymmetry*

$$\forall x, y \in X: (x, y) \in R \rightarrow (y, x) \notin R$$

(k) *Equivalence:* R satisfies (b), (d), and (i)
(l) *Ordering:* R satisfies (a), (b), and (d)
(m) *Quasi ordering:* R satisfies (b) and (d)

Consider an agent with a weak preference relation R on the set X of all conceivable options. If he is not like Buridan's ass, whose inability to compare two haystacks resulted in starvation, then R is a complete binary relation. It is also the case that R is reflexive, because every option is at least as good as itself, and the strict preference relation $P(R)$ is irreflexive.

The properties (d) to (h), when they are applied to a weak preference relation R, represent various degrees of the internal coherence of the preference structure. Let us explain their respective meanings as well as their mutual relations.

There are three distinct possibilities of the three-term cycles generated by R. First, we have a situation where we start from one option and climb up the preference ladder three times only to find ourselves in the initial position. If we denote $(u, v) \in P(R)$ by an arrow with two heads from v to u, this three-term cycle can be depicted as in Figure 1.1a. The second

Cambridge University Press
978-0-521-12255-9 - Rational Choice, Collective Decisions, and Social Welfare
Kotaro Suzumura
Excerpt
More information

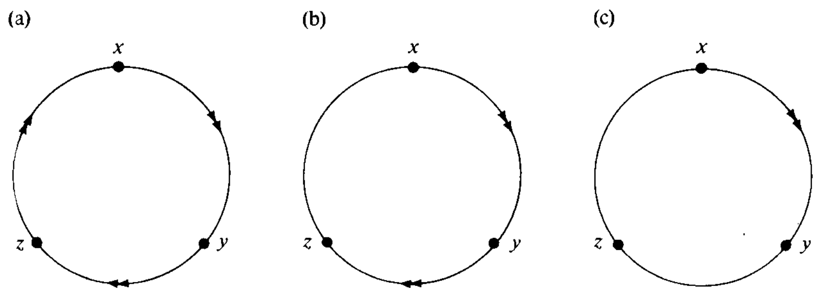

Figure 1.1. Three-term cycles.

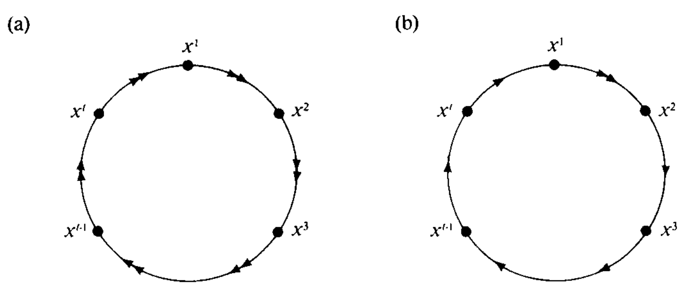

Figure 1.2. t-Term cycles.

possibility is that, starting from one option, we climb up the preference ladder twice and then move to another option on the same preference plateau only to realize that we are back to the initial position. Denoting $(u, v) \in I(R)$ by an arc without arrow connecting u and v, this three-term cycle can be shown as in Figure 1.1b. The third and last possibility is described in Figure 1.1c, where we start from an option and come back to that option again by climbing up the preference ladder once and then moving to another option on the same preference plateau twice. It is easy to verify that these three possibilities exhaust the three-term cycles that R can generate.

Referring back to the definitions (d), (e), and (f), we can easily verify that a transitive R excludes all of these three-term cycles. It can also be seen that a quasi-transitive R excludes the three-term cycles of the types in Figures 1.1a and 1.1b, whereas an acyclic R excludes the three-term cycles of the Figure 1.1a type only.

Going one step further, let us consider two possibilities of the t-term cycles generated by R, where $t \in Z^+$ satisfies $3 \leq t < +\infty$, that are described by Figures 1.2a and 1.2b. The first possibility described in

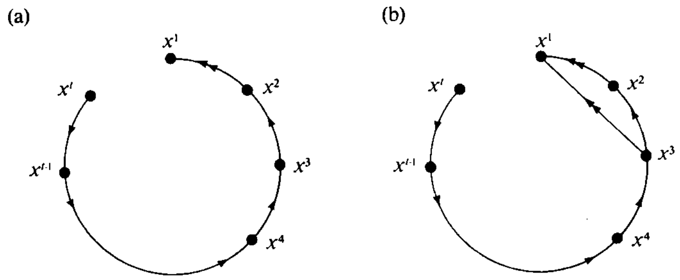

Figure 1.3. Transitivity and t-term coherence.

Figure 1.2a obtains when we come back to where we started after climbing up the preference ladder exactly t times. The second possibility is a bit more subtle. It obtains when we come back to where we started after climbing up the preference ladder just once and then visit $(t-1)$ other options subject to the condition that we should never follow the downhill road. This situation is described in Figure 1.2b, where an arc with a single-headed arrow from v to u denotes $(u, v) \in R$.

Referring back to the formal definitions (g) and (h), we can assert that an acyclic [resp. a consistent] R excludes the occurrence of the t-term cycles of the Figure 1.2a [resp. Figure 1.2b] type for every $t \in Z^+$. Because the t-term cycles of the Figure 1.2a type constitute a special case of the t-term cycles of the Figure 1.2b type, it follows that *consistency implies acyclicity*. It is also clear that *acyclicity implies triple acyclicity*.

How is transitivity, which is the strongest three-term coherence condition, related to the t-term coherence conditions? To answer this question, suppose that R is transitive and assume that we have a preference chain described by Figure 1.3a, where $3 \leq t < +\infty$, so that we have

$$(x^1, x^2) \in P(R) \,\&\, \forall \tau \in \{1, 2, \ldots, t-1\}: (x^\tau, x^{\tau+1}) \in R \tag{A.3}$$

R being transitive, it follows from $(x^1, x^2) \in P(R)$ and $(x^2, x^3) \in R$ that $(x^1, x^3) \in R$. We show that a contradiction ensues if we have $(x^1, x^3) \in I(R)$. Suppose that $(x^2, x^3) \in R$ is in fact $(x^2, x^3) \in P(R)$ [resp. $(x^2, x^3) \in I(R)$]. We then have a three-term cycle of the Figure 1.1b type [resp. the Figure 1.1c type], in contradiction with the assumed transitivity of R. Now that we have $(x^1, x^3) \in P(R)$, we can connect x^1 and x^3 with a double-headed arrow from the latter to the former to obtain Figure 1.3b. Using x^3 in place of x^2 and x^4 in place of x^3, we can repeat the foregoing argument to obtain $(x^1, x^4) \in P(R)$. Proceeding in this way, we can eventually assert that (A.3) entails $(x^1, x^t) \in P(R)$, which in effect excludes the possibility of a t-term