To verify transitivity, suppose that $x\ F(\succsim_1, \ldots, \succsim_I)\ y$ and $y\ F(\succsim_1, \ldots, \succsim_I)\ z$. We can assume that the three alternatives $\{x, y, z\}$ are distinct. Let $(\succsim_1'', \ldots, \succsim_I'') \in \mathscr{A}$ be a profile that takes $\{x, y, z\}$ to the top from $(\succsim_1, \ldots, \succsim_I)$. Because $f(\cdot)$ is weakly Paretian, we have $f(\succsim_1'', \ldots, \succsim_I'') \in \{x, y, z\}$.

Suppose that we had $y = f(\succsim_1'', \ldots, \succsim_I'')$. Consider a profile $(\succsim_1', \ldots, \succsim_I') \in \mathscr{A}$ that takes $\{x, y\}$ to the top from $(\succsim_1'', \ldots, \succsim_I'')$. Since y maintains its position from $(\succsim_1'', \ldots, \succsim_I'')$ to $(\succsim_1', \ldots, \succsim_I')$, it follows from monotonicity that $f(\succsim_1', \ldots, \succsim_I') = y$. But $(\succsim_1', \ldots, \succsim_I')$ also takes $\{x, y\}$ to the top from $(\succsim_1, \ldots, \succsim_I)$: the relative ordering of x and y, the two alternatives at the top, has not been altered in any individual preference in going from $(\succsim_1, \ldots, \succsim_I)$ to $(\succsim_1', \ldots, \succsim_I')$. Therefore we conclude that $y\ F(\succsim_1, \ldots, \succsim_I)\ x$, which contradicts the assumption that $x\ F(\succsim_1, \ldots, \succsim_I)\ y$, $x \neq y$. Hence, $y \neq f(\succsim_1'', \ldots, \succsim_I'')$.

Similarly, we obtain $z \neq f(\succsim_1'', \ldots, \succsim_I'')$. We only need to repeat the same argument using the pair $\{y, z\}$ (you are asked to do so in Exercise 21.E.3).

The only possibility left is $x = f(\succsim_1'', \ldots, \succsim_I'')$. Thus, let $(\succsim_1', \ldots, \succsim_I') \in \mathscr{A}$ take $\{x, z\}$ to the top from $(\succsim_1'', \ldots, \succsim_I'')$. Since x maintains its position in going from $(\succsim_1'', \ldots, \succsim_I'')$ to $(\succsim_1', \ldots, \succsim_I')$, it follows that $x = f(\succsim_1', \ldots, \succsim_I')$. But $(\succsim_1', \ldots, \succsim_I')$ also takes $\{x, z\}$ to the top from $(\succsim_1, \ldots, \succsim_I)$. Thus, $x\ F(\succsim_1, \ldots, \succsim_I)\ z$, and transitivity is established.

Step 4: The social welfare functional $F: \mathscr{A} \to \mathscr{P}$ rationalizes $f: \mathscr{A} \to X$; that is, for every profile $(\succsim_1, \ldots, \succsim_I) \in \mathscr{A}$, $f(\succsim_1, \ldots, \succsim_I)$ is a most preferred alternative for $F(\succsim_1, \ldots, \succsim_I)$ in X.

This is intuitive enough since $F(\cdot)$ has been constructed from $f(\cdot)$. Denote $x = f(\succsim_1, \ldots, \succsim_I)$ and let $y \neq x$ be any other alternative. Consider a profile $(\succsim_1', \ldots, \succsim_I') \in \mathscr{A}$ that takes $\{x, y\}$ to the top from $(\succsim_1, \ldots, \succsim_I)$. Since x maintains position from $(\succsim_1, \ldots, \succsim_I)$ to $(\succsim_1', \ldots, \succsim_I')$, we have $x = f(\succsim_1', \ldots, \succsim_I')$. Therefore, $x\ F(\succsim_1, \ldots, \succsim_I)\ y$.

Step 5: The social welfare functional $F: \mathscr{A} \to \mathscr{P}$ is Paretian.

Clear if $x \succ_i y$ for every i then, by the Paretian property of $f(\cdot)$, we must have $x = f(\succsim_1', \ldots, \succsim_I')$ whenever $(\succsim_1', \ldots, \succsim_I')$ takes $\{x, y\}$ to the top from $(\succsim_1, \ldots, \succsim_I)$. Hence $x\ F(\succsim_1, \ldots, \succsim_I)\ y$, and by step 3 we conclude that $x\ F_p(\succsim_1, \ldots, \succsim_I)\ y$.

Step 6: The social welfare functional $F: \mathscr{A} \to \mathscr{P}$ satisfies the pairwise independence condition.

This follows from step 1. Suppose that $(\succsim_1, \ldots, \succsim_I) \in \mathscr{A}$ and $(\succsim_1', \ldots, \succsim_I') \in \mathscr{A}$ have the same ordering of $\{x, y\}$ for every i (that is, for every i, $x \succsim_i y$ if and only if $x \succsim_i' y$). Suppose that $(\succsim_1'', \ldots, \succsim_I'') \in \mathscr{A}$ takes $\{x, y\}$ to the top from $(\succsim_1, \ldots, \succsim_I)$ and that, say, $x = f(\succsim_1'', \ldots, \succsim_I'')$. Then $x\ F(\succsim_1, \ldots, \succsim_I)\ y$. But $(\succsim_1'', \ldots, \succsim_I'')$ also takes $\{x, y\}$ to the top from $(\succsim_1', \ldots, \succsim_I')$. Hence, $x\ F(\succsim_1', \ldots, \succsim_I')\ y$, as we wanted to prove.

Step 7: The social choice function $f: \mathscr{A} \to X$ is dictatorial.

By Arrow's theorem (Proposition 21.C.1) there is an agent $h \in I$ such that for every profile $(\succsim_1, \ldots, \succsim_I) \in \mathscr{A}$ we have $x\ F_p(\succsim_1, \ldots, \succsim_I)\ y$ whenever $x \succ_h y$. Therefore, $f(\succsim_1, \ldots, \succsim_I)$ [which by step 4 is a most preferred alternative for

$F(\succsim_1, \ldots, \succsim_I)$ in X] must also be a most preferred alternative for h; that is, $f(\succsim_1, \ldots, \succsim_I) \succsim_h x$ for every $x \in X$. Hence agent h is a dictator. ■

Finally, we mention the following corollary (Proposition 21.E.2) to hint at the connection between Proposition 21.E.1 and the issue of incentives to truthful preference revelation, a topic that is studied extensively in Chapter 23.

Proposition 21.E.2: Suppose that the number of alternatives is at least three and that $f: \mathscr{P}^I \to X$ is a social choice function that is weakly Paretian and satisfies the following *no-incentive-to-misrepresent* condition:

$$f(\succsim_1, \ldots, \succsim_{h-1}, \succsim_h, \succsim_{h+1}, \ldots, \succsim_I) \succsim_h f(\succsim_1, \ldots, \succsim_{h-1}, \succsim'_h, \succsim_{h+1}, \ldots, \succsim_I)$$

for every agent h, every $\succsim'_h \in \mathscr{P}$, and every profile $(\succsim_1, \ldots, \succsim_I) \in \mathscr{P}^I$. Then $f(\cdot)$ is dictatorial.

Proof: In view of Proposition 21.E.1 it suffices to show that $f: \mathscr{P}^I \to X$ must be monotonic.

Suppose that it is not. Then without loss of generality we can assume that, for some agent h, there are preferences $\succsim_i \in \mathscr{P}$ for agents $i \neq h$, and preferences $\succsim''_h$, $\succsim'''_h \in \mathscr{P}$ for agent h, such that, denoting

$$x = f(\succsim_1, \ldots, \succsim_{h-1}, \succsim''_h, \succsim_{h+1}, \ldots, \succsim_I)$$

and

$$y = f(\succsim_1, \ldots, \succsim_{h-1}, \succsim'''_h, \succsim_{h+1}, \ldots, \succsim_I),$$

we have that $x \succsim''_h z$ implies $x \succsim'''_h z$ for every $z \in X$, and yet $y \neq x$.

There are two possibilities: Either $y \succ''_h x$ or $x \succsim''_h y$.

If $y \succ''_h x$ then the no-incentive-to-misrepresent condition is violated for the "true" preference relation $\succsim_h = \succsim''_h$ and the misrepresentation $\succsim'_h = \succsim'''_h$.

If $x \succsim''_h y$ then $x \succsim'''_h y$. Therefore, since no two distinct alternatives can be indifferent, $x \succ'''_h y$. But if $x \succ'''_h y$ then the no-incentive-to-misrepresent condition is violated for the "true" preference relation $\succsim_h = \succsim'''_h$ and the misrepresentation $\succsim'_h = \succsim''_h$. ■

REFERENCES

Arrow, K. J. (1963). *Social Choice and Individual Values*, 2d ed. New York: Wiley.

Austen-Smith, D., and J. S. Banks. (1996). *Positive Political Theory*. Ann Arbor: University of Michigan Press.

Caplin, A., and B. Nalebuff. (1988). On 64%-majority voting. *Econometrica* **56**: 787–814.

Grandmont, J-M. (1978). Intermediate preferences and majority rule. *Econometrica* **46**: 317–30.

May, K. (1952). A set of independent, necessary and sufficient conditions for simple majority decision. *Econometrica* **20**: 680–84.

Moulin, H. (1988) *Axioms of Cooperative Decision Making*. Cambridge, U.K.: Cambridge University Press.

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

Sen, A. (1986). Social choice theory. Chap. 22 in *Handbook of Mathematical Economics*, edited by K. Arrow, and M. Intriligator. Amsterdam: North-Holland.

Shepsle, K. A., and M. Boncheck. (1995), *Analyzing Politics*. New York: W. W. Norton.

Tullock, G. (1967). The general irrelevancy of the general possibility theorem. *Quarterly Journal of Economics* **81**: 256–70.

EXERCISES

21.B.1[A] Verify that majority voting between two alternatives satisfies the properties of symmetry among agents, neutrality between alternatives, and positive responsiveness.

21.B.2[A] For each of the three properties characterizing majority voting between two alternatives according to Proposition 21.B.1 (symmetry among agents, neutrality between alternatives, and positive responsiveness) exhibit an example of a social welfare functional $F(\alpha_1, \ldots, \alpha_I)$ distinct from majority voting and satisfying the other two properties. This shows that none of the three properties is redundant for the characterization result.

21.B.3[A] Suppose there is a public good project that can take two levels $k \in \{0, 1\}$, where $k = 0$ can be interpreted as the status quo. The cost, in dollars, of any level of the public good is zero. There is a population I of agents having quasilinear preferences (with dollars as numeraire) over levels of the public good and money holdings. Thus, the preferences of agent i are completely described by the willingness to pay $v_i \in \mathbb{R}$ for the level $k = 1$ over the level $k = 0$. The number v_i may be negative (in this case it amounts to the minimum compensation required).

Show that a majority rule decision over the two levels of the public project guarantees a Pareto optimal decision over the set of policies constituted by the two levels of the public project (with no money transfers taking place across agents) but not over the larger set of policies in which transfers across agents are also possible.

Compare and contrast the majority decision rule (a "median") with the Pareto optimum decision rule for the case in which transfers across agents are possible (a "mean").

21.C.1[A] Provide the requested completion of step 1 of the proof of Proposition 21.C.1.

21.C.2[B] We can list the implicit and explicit assumptions of the Arrow impossibility theorem (Proposition 21.C.1) to be the following:

(a) The number of alternatives is at least 3.
(b) Universal domain: To be specific, the domain of the social welfare functional $F(\cdot)$ is $\mathscr{R}^I$.
(c) Social rationality: That is, $F(\succsim_1, \ldots, \succsim_I)$ is a rational preference relation (i.e. complete and transitive) for every possible profile of individual preferences.
(d) Pairwise independence (Definition 21.C.3).
(e) Paretian condition (Definition 21.C.2).
(f) No dictatorship: That is, there is no agent h that at any profile of individual preferences imposes his strict preference over any possible pair of alternatives (see Proposition 21.C.1 for a precise definition).

For each of these six assumptions exhibit a social welfare functional $F(\cdot)$ satisfying the other five. This shows that none of the conditions is redundant for the impossibility result.

21.C.3[A] Show that there are social welfare functionals $F: \mathscr{R}^I \to \mathscr{R}$ defined on $\mathscr{R}^I$ (i.e., individual indifference is possible) satisfying all the conditions of Arrow's impossibility theorem (Proposition 21.C.1) and for which, however, the social preferences are not *identical* to the preferences of any individual. [*Hint*: Try a *lexical dictatorship* in which the nth-ranked dictator imposes his preference if and only if every higher ranked dictator is indifferent.]

21.D.1[B] Suppose that X is a finite set of alternatives. Construct a reflexive and complete preference relation $\succsim$ on X with the property that $\succsim$ has a maximal element on every strict subset $X' \subset X$, and yet $\succsim$ is not acyclic.

21.D.2[A] Verify that the social preferences generated by the oligarchy example (Example 21.D.1) are quasitransitive but that social indifference may not be transitive. Interpret.

21.D.3[A] Show that the social preferences generated by the vetoers example (Example 21.D.2) are acyclic but not necessarily quasitransitive. Show also that in spite of the veto power of agent 2 it may happen that alternative x is the only maximal alternative for the social preferences.

21.D.4[A] With reference to Example 21.D.4, show that a continuous preference relation $\succsim$ on $X = [0, 1]$ is single peaked only if it is *strictly convex*.

21.D.5[A] Give a direct proof that none of the six linear orders possible among three alternatives can make the three preferences involved in the Condorcet paradox (Example 21.C.2) into a single-peaked family.

21.D.6[B] Give an example with an even number of agents and single-peaked preferences in which pairwise majority voting fails to generate a fully transitive social preference relation.

21.D.7[C] Suppose that X is a convex subset of $\mathbb{R}^2$ with the origin in its interior. There are three agents $i = 1, 2, 3$. Every i has a continuously differentiable utility function $u_i: X \to \mathbb{R}$. Assume that the cone in $\mathbb{R}^2$ spanned by the set of gradients at the origin $\{\nabla u_1(0), \nabla u_2(0), \nabla u_3(0)\}$ is the entire $\mathbb{R}^2$. Show the following:

(a) There are three alternatives $x, y, z \in X$ that constitute a Condorcet cycle (i.e., there is a strict majority for x over y, y over z, and z over x).

(b) (Harder) Given any $x \in \mathbb{R}^2$, there is a $y \in \mathbb{R}^2$ such that $\|x - y\| < \|x\|$ and y is preferred by two agents to the origin $0 \in \mathbb{R}^2$. That is, if you think of the origin as the status-quo then for any x we can find a strict majority that prefers, over the status-quo, an alternative that moves us closer to x. [*Hint:* You can safely assume that the utility functions are linear.]

21.D.8[C] The situation is as in Exercise 21.D.7 except that now, at the origin, the gradients of the utility functions constitute a pointed cone (i.e. the cone does not contain any half-space). Assume also that utility functions are quasiconcave.

(a) Argue that at the origin there is an agent who is a directional median in the sense that any alternative having a strict majority against the origin must make this agent strictly better off.

(b) Suppose now that at every $x \in X$ the cone spanned by $\{\nabla u_1(x), \nabla u_2(x), \nabla u_3(x)\}$ is pointed. Then according to (a) there is a directional median agent at every $x \in X$. Show that this directional median agent can change with x and that Condorcet cycles are possible.

(c) The situation is as in (b). Show that, if the directional median agent is the same at every $x \in X$, then there can be no Condorcet cycle.

21.D.9[C] (Grandmont) Consider a set of alternatives X. Given three rational preference relations $\succsim$, $\succsim'$, $\succsim''$ on X, one says that $\succsim''$ is *intermediate* between $\succsim$ and $\succsim'$ if $x \succsim y$ and $x \succsim' y$ implies $x \succsim'' y$. That is, for every alternative y the intersection of the upper contour sets for $\succsim$ and $\succsim'$ is contained in the upper contour set for $\succsim''$.

(a) Show that if $u(x)$ and $u'(x)$ are utility functions for preferences on X then, for any positive numbers γ and ψ, the preference relation represented by $\psi u(x) + \gamma u'(x)$ is intermediate between the preference relations represented by $u(x)$ and $u'(x)$.

(b) Suppose we are given N functions $h_1(x), \ldots, h_N(x)$ defined on X. The preferences of agents are represented by utility functions of the form $u_\beta(x) = \beta_1 h_1(x) + \cdots + \beta_N h_N(x)$, where $\beta = (\beta_1, \ldots, \beta_N) \in \mathbb{R}^N_{++}$. Show that for any two alternatives $x, y \in X$, the set $B(x, y) = \{\beta \in \mathbb{R}^N_{++} : u_\beta(x) > u_\beta(y)\}$ is the intersection of $\mathbb{R}^N_{++}$ with a translated half-space.

(c) Argue that the conclusion from (**b**) is still correct if a parametrization of utility functions $u_\beta(x)$ by a $\beta \in \mathbb{R}^N$ is such that whenever β'' is a convex combination of β and β' then the preferences represented by $u_{\beta''}(x)$ are intermediate between the preferences represented by $u_\beta(x)$ and $u_{\beta'}(x)$.

(d) Continuing with the parametrization of (**b**), suppose that we take the limit situation where the population of agents is represented by a density $g(\beta)$ over $\mathbb{R}^N_{++}$. We say that β^* is a *median agent* for $g(\cdot)$ if *every* hyperplane in $\mathbb{R}^N$ passing through β^* divides $\mathbb{R}^N$ into two regions having equal mass according to the density $g(\cdot)$. Show that a median agent for an arbitrary $g(\cdot)$ may or may not exist.

(e) In the framework of (**d**), suppose there is a median agent β^*, that $g(\beta^*) > 0$, and that x^* is the single most preferred alternative of the median agent. Show then that x^* defeats any other alternative in pairwise majority voting.

(f) Show that the Euclidean preferences in Example 21.D.6 can be put into the framework of this exercise by keeping the sets of alternatives and of agents conceptually separated.

21.D.10[B] The purpose of this exercise is to illustrate the use of single-peakedness in a policy problem. Specifically, we consider the problem of determining by majority voting a tax level for wealth redistribution.

Suppose that there is an odd number I of agents. Each agent has a level of wealth $w_i > 0$ and an increasing utility function over wealth levels. The mean wealth is $\bar{w}$, and the median wealth is w^*.

(a) Interpret the distributional significance of a difference between $\bar{w}$ and w^*.

(b) Consider a proportional tax rate $t \in [0, 1]$ identical across agents. The set of alternatives is $X = [0, 1]$, the set of possible levels of the tax rate. Tax receipts are redistributed uniformly. Thus, for a tax rate t, the after-tax wealth of agent i is $(1 - t)w_i + t\bar{w}$. Show that the preferences over X of all agents are single peaked and that the Condorcet winner t_c is $t_c = 0$ or $t_c = 1$ according to whether $w^* > \bar{w}$ or $w^* < \bar{w}$, respectively. Interpret.

(c) Now suppose that taxation gives rise to a deadweight loss. Being very crude about it, suppose that a tax rate of $t \in [0, 1]$ decreases the pretax level of agent i's wealth to $w_i(t) = (1 - t)w_i$ [thus, the average tax receipts are $t(1 - t)\bar{w}$ and the ex post wealth level of agent i is $(1 - t)^2 w_i + t(1 - t)\bar{w}$]. Show that preferences on wealth levels are again single peaked (but notice that the after-tax wealth level may not be a concave function of the tax rate). Show then that $t_c \leq \frac{1}{2}$. Also, $t_c = 0$ or $t_c > 0$ according to whether $w^* > \frac{1}{2}\bar{w}$ or $w^* < \frac{1}{2}\bar{w}$, respectively. Compare with (b) and interpret.

(d) Let us modify (**c**) by assuming that the deadweight loss affects individual wealth differently: A tax rate of $t \in [0, 1]$ decreases pretax wealth of agent i to $(1 - t^2)w_i$ [this is theoretically more satisfactory than the situation in (**c**) since we know from first principles that at $t = 0$ a small increase in t should have a second-order effect on total welfare]. Show then that individual preferences on tax rates need no longer be single peaked.

21.D.11[B] Consider a finite set of alternatives X and a set of preferences $\mathscr{P}_\geq$, single-peaked with respect to some linear order $\geq$ on X (note that we rule out the possibility of individual indifference). The number of agents is odd. As we have seen in Proposition 21.D.2, a possible class of social welfare functionals $F: \mathscr{P}_\geq^I \to \mathscr{P}$ that satisfy the Paretian and pairwise independence conditions are those where we fix a subset $S \subset I$ composed of an odd number of agents (a kind of oligarchy) and let the members of this subset determine social preferences by pairwise majority voting. Show by example that this is *not* the only possible class of social welfare functionals $F: \mathscr{P}_\geq^I \to \mathscr{P}$ that satisfy the Paretian and pairwise independence conditions.

21.D.12[A] Suppose that the total cost $c > 0$ of a project has to be financed by levying taxes from three agents. Therefore, the set of alternatives is $X = \{(t_1, t_2, t_3) \geq 0: t_1 + t_2 + t_3 = c\}$. The financing scheme is to be decided by majority voting.

(a) Show that no strictly positive alternative $(t_1, t_2, t_3) \gg 0$ can be a Condorcet winner.
(b) Discuss what happens with alternatives (t_1, t_2, t_3) where $t_i = 0$ for some i.

21.D.13[B] We have a population of agents (to be simple, a continuum) with Euclidean preferences in $\mathbb{R}^2$. The preferences of the agents fall into a finite number J of types. Each type is indexed by the most preferred point x_j. We assume that the x_j's are in "general position," in the sense that no three of the x_j's line up into a straight line. We denote by $\alpha_j \in [0, 1]$ the fraction of the total mass of agents that are of type j.

(a) Suppose that J is odd and $\alpha_1 = \cdots = \alpha_J$. Prove that if $y \in \mathbb{R}^2$ is a Condorcet winning alternative, then $y \in \{x_1, \ldots, x_J\}$. That is, the Condorcet winning alternative must coincide with the top alternative of some type. Does this remain true if J is even?
(b) (De Marzo) Suppose now that there is a dominant type, that is, a type h such that $\alpha_h > \alpha_j$ for every $j \neq h$. Prove that if there is a Condorcet winning alternative $y \in \mathbb{R}^2$, then $y = x_h$. That is, only the top alternative of the dominant type can be a Condorcet winning alternative.

21.D.14[B] In this exercise we verify that we cannot weaken the definition of single-peakedness to require only that preference be weakly increasing as the peak is approached.

Suppose we have five agents and five alternatives $\{x, y, z, v, w\}$. The individual preferences are

$$x \succ_1 y \sim_1 z \sim_1 v \sim_1 w,$$

$$y \succ_2 x \succ_2 z \succ_2 v \succ_2 w,$$

$$z \succ_3 y \sim_3 v \sim_3 w \succ_3 x,$$

$$v \succ_4 w \succ_4 x \sim_4 y \sim_4 z,$$

$$w \succ_5 x \sim_5 y \sim_5 z \sim_5 v,$$

(a) Show that there is no Condorcet winner among these alternatives; that is, every alternative is defeated by majority voting by some other alternative.
(b) Show that there is a linear order $\geq$ on the alternatives such that the preference relation of the five agents satisfies the following property: "Preferences are weakly increasing as we approach, in the linear order $\geq$, the most preferred alternative of the agent."
(c) Verify that the alternatives could be viewed as points in $[0, 1]$ and that the preferences of each agent could be induced by the restriction to the set of alternatives of a quasiconcave utility function on $[0, 1]$. [Note: $u_i(t)$ is quasiconcave if $\{t \in [0, 1]: u_i(t) \geq \gamma\}$ is convex for every γ.]
(d) (Harder) Extend the previous arguments and constructions into an example with the following characteristics: (i) There are five agents; (ii) the space of alternatives equals the interval $[0, 1]$; (iii) every agent has a quasiconcave utility function on $[0, 1]$ with a single maximal alternative; and (iv) there is no Condorcet winner in $[0, 1]$.

21.E.1[A] Consider a finite set of alternatives X and suppose that there is an odd number of agents. The domain of preferences is $\mathscr{A} = \mathscr{P}^I_{\geq}$, where $\geq$ is a linear order on X (i.e., preferences are single peaked and individual indifferences do not arise). Show that the social choice function that assigns the Condorcet winner to every profile satisfies the weak Pareto and the monotonicity conditions.

21.E.2^{A} Suppose that the set of alternatives X has $N < \infty$ elements and that the alternatives are given to us with labels that go from 1 to N, that is, $X = \{x_1, \ldots, x_N\}$. Consider the social choice function defined on $\mathscr{R}^I$ (i.e., we allow for individual indifference) by letting $f(\succsim_1, \ldots, \succsim_I)$ be the alternative that has the smallest label among all the alternatives that are most preferred by the first agent. Show that this social choice function is dictatorial, weakly Paretian and monotonic. For the sake of completeness, carry out the same verification if the domain of $f(\cdot)$ is $\mathscr{P}^I$.

21.E.3^{A} As requested, complete the proof of step 3 of Proposition 21.E.1.

21.E.4^{A} Suppose that the number of alternatives is finite and that $F: \mathscr{A} \to \mathscr{P}$ is a social welfare functional satisfying the weakly Paretian and the pairwise independence condition on some domain $\mathscr{A} \subset \mathscr{P}^I$. The *induced social choice function* assigns to every profile the socially most preferred alternative. Give two examples where the induced social choice function is not monotonic. One example should be for the two-alternative case and $\mathscr{A} = \mathscr{P}^I$, and the other should be for a case with more than two alternatives. [*Hint:* Choose $\mathscr{A}$ to be very small.]

CHAPTER

22

Elements of Welfare Economics and Axiomatic Bargaining

22.A Introduction

In this chapter, we continue our study of welfare economics. The main difference from Chapter 21 is that here the cardinal aspects of individual utility functions will be at center stage. Moreover, we will not eschew exploring the implications of assuming that utilities are interpersonally comparable.

In Section 22.B, we present the concept of the *utility possibility set*. We also emphasize the distinction between first-best and second-best welfare problems.

In Section 22.C, we first posit the existence of a policy maker, or social planner, endowed with coherent objectives in the form of a *social welfare function*. The role of policy consists, precisely, in the maximization of the social welfare function subject to the constraint represented by the utility possibility set. We then analyze a variety of practically useful examples. The section concludes with a brief discussion of the *compensation criterion*.

In Section 22.D, we probe the extent to which interpersonal comparisons of utility underlie the use of social welfare functions. We do this by analyzing the implications of axioms that postulate the invariance of social preferences to changes in the origins and units of individual utility functions. This section links naturally with Chapter 21 as, again, it relies on the concept of a social welfare functional and, through a different road, it takes us back to Arrow's impossibility theorem.

Sections 22.E and 22.F deal with a somewhat different topic: axiomatic bargaining theory. The aim is now to formulate and analyze reasonable criteria for dividing among several agents the *gains* (or *losses*) from a cooperative endeavor.

In Section 22.E, we study the simplest case: that in which either there is complete cooperation (with the possible outcomes of cooperation described by a utility possibility set) or the outcome is a given *threat point*. We present several solutions for this case, among them the classical *Nash bargaining solution*.

In Section 22.F, we restrict ourselves to the situation in which the utility is transferable among agents. We allow, however, for the possibility of cooperation among subgroups of agents. A classical solution is then the *Shapley value*, of which we give a brief account. We also provide an illustration of an interesting application of the Shapley value to a problem of allocating joint costs to individual projects.

22.B Utility Possibility Sets

As a first step in the study of policy decision problems, this section is concerned with the description of the set of options available to a policy maker. The following section will consider the objectives of the policy maker.[1]

The starting point of the analysis is a nonempty *set of alternatives* X and a collection of I agents. In contrast with Chapter 21, where we used preference relations, we will now assume that agents' tastes are given to us in the form of utility functions $u_i: X \to \mathbb{R}$. One may wonder what is the exact meaning of the utility values $u_i(x)$: Do they have cardinal or ordinal significance? Are they comparable across individuals? These questions will be considered in Section 22.D. For current purposes there is no need to answer them.

It is a traditional, and firm, principle of welfare economics that policy making should not be *paternalistic*. At a minimum, this means that alternatives that cannot be distinguished from the standpoint of agents' tastes should not be distinguished by the policy maker either. We are therefore led to the idea that only the agents' utility values for the different alternative should matter and therefore that the relevant constraint set for the policy maker is the *utility possibility set* [introduced by Samuelson (1947)], which we now define.

Definition 22.B.1: The *utility possibility set* (UPS) is the set

$$U = \{(u_1, \ldots, u_I) \in \mathbb{R}^I : u_1 \leq u_1(x), \ldots, u_I \leq u_I(x) \text{ for some } x \in X\} \subset \mathbb{R}^I.$$

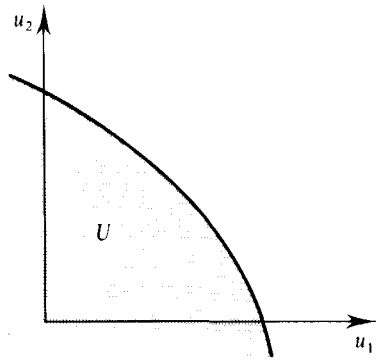
The *Pareto frontier* of U is formed by the utility vectors $u = (u_1, \ldots, u_I) \in U$ for which there is no other $u' = (u'_1, \ldots, u'_I) \in U$ with $u'_i \geq u_i$ for every i and $u'_i > u_i$ for some i.

To gain some insight into the characteristics of UPSs, and in particular, into the important distinction between *first-best* and *second-best* policy problems, we discuss some examples.

Example 22.B.1: *Exchange Economies.* Suppose that we focus on the exchange and production economies with L commodities and I consumers studied in Chapter 10 and in Part IV. The set of alternatives $X \subset \mathbb{R}^{LI}$ then stands for the set of feasible consumption allocations $x = (x_1, \ldots, x_I)$. The utility functions of the different consumers have the form $u_i(x) = u_i(x_i)$; that is, consumer i's utility from an allocation depends only on her own consumption. In Exercise 22.B.1 you are asked to show that under standard conditions (including the concavity of the utility functions), the UPS of this economy is a convex set. In the particular case where the utility functions are quasilinear,[2] we saw in Section 10.D that the boundary of U is a hyperplane. The general case and the quasilinear case are illustrated in Figures 22.B.1 and 22.B.2, respectively. ■

1. For general introductions to public economics, see Atkinson and Stiglitz (1980), Laffont (1988), and Starrett (1988). At a more advanced level, see Guesnerie (1995). Phelps (1973) contains a good compilation of basic articles emphasizing conceptual foundations.

2. As usual, in this case we also neglect the nonnegativity constraints on numeraire.



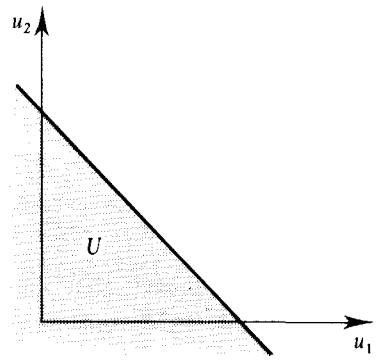


Figure 22.B.1 (left)
A utility possibility set.

Figure 22.B.2 (right)
A utility possibility set: transferable utility.

Example 22.B.1 corresponds to a first-best situation. A *first-best problem* is one in which the constraints defining X are only those imposed by technology and resources. The policy maker cannot produce from a void and, therefore, must respect these constraints, but otherwise she can appeal to any conceivable policy instrument. If, as is often the case, there are other restrictions on the usable instruments, we say that we have a *second-best problem*. The restrictions can be of many sorts: legal, institutional, or, more fundamentally, informational. The last type were amply illustrated in Chapters 13 and 14 (and will be seen again in Chapter 23). We should warn, however, that the conceptual distinction between first-best and second-best problems is not sharp. In a sense, adverse selection or agency restrictions are as primitive as technologies and endowments.

Example 22.B.2: *Ramsey Taxation.* Consider a quasilinear economy with three goods, of which the third is the numeraire. The numeraire good can be freely transferred across consumers (more formally, one of the policy instruments available to the policy maker is the lump-sum redistribution of wealth). The first two goods are produced from the numeraire at a constant marginal cost equal to 1. Consumers face market prices that are equal to marginal cost plus a commodity tax whose level is fixed by the policy maker. Tax proceeds are returned to the economy in lump-sum form. Finally, the amounts consumed are those determined by the demand functions of the different consumers.

We know from the second welfare theorem (Section 16.D) that any utility vector in the first-best UPS can be reached with the above instruments (it suffices to set the tax rates at a zero level and distribute wealth appropriately). But suppose that we now have an unavoidable distortion—the policy maker is constrained to raise a total amount R of tax receipts. This has then become a second-best problem. To determine the corresponding second-best UPS, note first that, since the numeraire is freely transferable across consumers, the boundary of this set is still linear, as in the first-best case (i.e., as in Figure 22.B.2). Hence, to place this boundary it suffices to find the level of prices p_1, p_2 that maximizes $v(p_1, p_2)$, the indirect utility function of a representative consumer (which, up to an increasing transformation, equals the

aggregate consumer surplus; see Section 4.D and Chapter 10 for these concepts).[3]

Denote by $x_1(p_1, p_2)$ and $x_2(p_1, p_2)$ the aggregate demand functions. Then we must solve the problem

$$\begin{aligned} \text{Max} \quad & v(p_1, p_2) \\ \text{s.t.} \quad & (p_1 - 1)x_1(p_1, p_2) + (p_2 - 1)x_2(p_1, p_2) \geq R. \end{aligned}$$

Suppose, to take the simplest case, that the utility functions of the different consumers are additively separable. This means that the two demand functions can be written as $x_1(p_1)$ and $x_2(p_2)$. Then the first-order conditions satisfied by a solution $(\bar{p}_1, \bar{p}_2)$ of the maximization problem are (carry out the calculation in Exercise 22.B.2):

There is $\lambda < 0$, such that

$$\lambda(\bar{p}_1 - 1)\frac{dx_1(\bar{p}_1)}{dp_1} = (1 - \lambda)x_1(\bar{p}_1)$$

and

$$\lambda(\bar{p}_2 - 1)\frac{dx_2(\bar{p}_2)}{dp_2} = (1 - \lambda)x_2(\bar{p}_2).$$

Denoting by $t_\ell = (\bar{p}_\ell - 1)/\bar{p}_\ell$ the tax rate on good ℓ, we can write this condition in elasticity form as

$$t_1 = \frac{\alpha}{\varepsilon_1(\bar{p}_1)} \quad \text{and} \quad t_2 = \frac{\alpha}{\varepsilon_2(\bar{p}_2)} \quad \text{for some } \alpha > 0. \tag{22.B.1}$$

Expression (22.B.1) is known as the *Ramsey taxation formula* [because of Ramsey (1927)]. An implication of it is that if the demand for good 1 is uniformly less elastic than that for good 2, then the optimal tax rate for good 1 is higher. This makes sense: For example, if the demand for good 1 is totally inelastic then there is no deadweight loss from taxation of this good (see Section 10.C) and therefore we could reach the first-best optimum by taxing only this good.[4] ■

Example 22.B.3: *Compensatory Distortion.* The basic economy is as in Example 22.B.2, except that we do not necessarily assume that the utility functions of the consumers are additively separable. The distortion is now of a different type. We

3. Because total surplus equals consumer surplus plus the fixed amount of tax revenues R, by maximizing consumer surplus we maximize total surplus. We note also that the assumption that the amount R must be raised through commodity taxation is somewhat artificial in a context where lump-sum redistribution is possible. We make the assumption, in this and the next example, merely to be pedagogical. Alternatively, we could rule out the possibility of lump-sum transfers. In this case the exercise carried out in this example (and the next) determines the first-order conditions for the problem of maximizing the sum of individual utilities (the "purely utilitarian social welfare function" in the terminology of Section 22.C)

4. We should warn that the formulas in (22.B.1) constitute only first-order conditions. As we shall see in the forthcoming examples, second-best problems are frequently nonconvex and therefore the satisfaction of first-order conditions does not guarantee that we have determined a true maximum.

assume that p_1 is fixed at some level $\hat{p}_1 > 1$.[5] The policy instruments are any transfer of numeraire across agents and the level of a commodity tax on the second good. The net revenue in the two markets is given back to consumers in a lump-sum form. The solution $\bar{p}_2$ of the surplus-maximization problem is then characterized by the first-order conditions (see Exercise 22.B.3)

$$(\hat{p}_1 - 1)\frac{\partial x_1(\hat{p}_1, \bar{p}_2)}{\partial p_2} + (\bar{p}_2 - 1)\frac{\partial x_2(\hat{p}_1, \bar{p}_2)}{\partial p_2} = 0. \tag{22.B.2}$$

Note that except in the separable case, where $\partial x_1(\hat{p}_1, \bar{p}_2)/\partial p_2 = 0$, we have $\bar{p}_2 \neq 1$; that is, even if the initial distortion involves only the first market, second-best efficiency requires creating a compensatory distortion in the second market [this point was emphasized by Lipsey and Lancaster (1956)]. This is an intuitive result: suppose that we were to put $p_2 = 1$; then the last (infinitesimal) unit demanded of the second good makes a contribution $p_2 - 1 = 0$ to the total surplus (recall that p_2 will equal the marginal utility for good 2). Therefore, a small tax on good 2 is desirable because its effect is to divert some demand toward good 1, where the contribution to total surplus of the last unit demanded is $\hat{p}_1 - 1 > 0$. ■

Example 22.B.4: *Few Policy Instruments.* In Examples 22.B.2 and 22.B.3 we have assumed that the unrestricted transfer of numeraire across consumers is one of the instruments available to the policy maker. Because of this, in those two examples the UPS had a "full" frontier, that is, a frontier that is an $(I - 1)$-dimensional surface. In addition, quasilinearity insured that this surface was flat (and therefore that the UPS was convex). We now explore the implications of limiting the extent to which the numeraire is transferable.

We assume that we have two goods and that the utility functions of I consumers are quasilinear with respect to the first good (which is untaxed). Arbitrary transfers of numeraire are not permitted, however. The policy maker now has a single instrument: a commodity tax (or subsidy) on the second good. Again, this good can be produced at unit marginal cost. The policy maker's surplus (or deficit) is given back to the consumers *according to some fixed rule* (hence, no arbitrary transfers of numeraire are permitted). Say, to be specific, that this rule is that the surplus–deficit is absorbed by the first consumer. Then the (second-best) UPS is [denoting by $v_i(p_2)$ the indirect utility function of consumer i]

$$U = \{u \in \mathbb{R}^I : u \leq (v_1(p_2) + (p_2 - 1)\textstyle\sum_i x_i(p_2), v_2(p_2), \ldots, v_I(p_2)) \text{ for some } p_2 > 0\}.$$

Two points are worth observing. The first is that U does not need to be convex (you should show this in Exercise 22.B.4; recall from Proposition 3.D.3 that the indirect utility functions are quasi-convex. An example is represented in Figure 22.B.3. The second is that U is defined by means of a single parameter, p_2, and therefore its Pareto frontier (which, naturally, lies in $\mathbb{R}^I$) is one-dimensional. See Figure 22.B.4 for a case with $I = 3$. This feature is entirely typical. As long as the instruments available to the policymaker are fewer than $I - 1$ in number, the frontier of the UPS cannot be $(I - 1)$-dimensional. Note that when there is free transferability of numeraire across

5. More generally, we could think of the market for good 1 as being beyond the control of the policy maker and giving rise, perhaps because of a monopolistic structure, to a price higher than marginal cost.

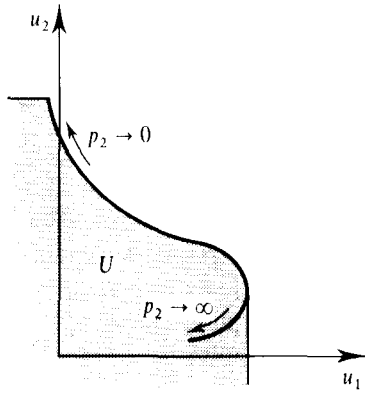


Figure 22.B.3 (left)
A nonconvex second-best utility possibility set (Example 22.B.4).

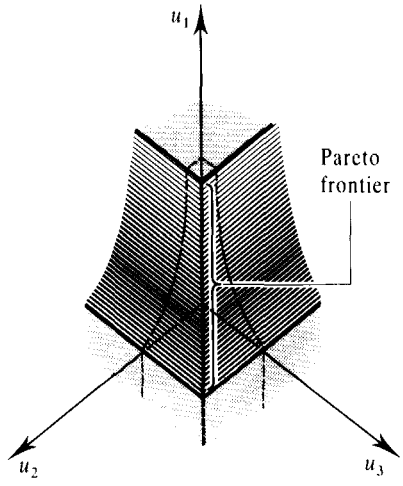


Figure 22.B.4 (right)
A second-best utility possibility set for a case with few instruments: low-dimensional Pareto frontier (Example 22.B.4).

the I consumers, this automatically gives us the necessary minimum of $I - 1$ instruments. ■

Example 22.B.5: *First-best Nonconvexities.* In Example 22.B.4 the possible nonconvexity of the UPS is due to the second-best nature of this set. If lump-sum transfers of numeraire were allowed, then the corresponding first-best UPS would be convex. Yet a first-best UPS may also be nonconvex. Two familiar sources of nonconvexities in first-best problems are indivisibilities and externalities. As for the first, suppose that there are two locations and two agents with identical locational tastes (in particular, they both prefer the same location). There are only two possible assignments of individuals to locations and therefore the UPS will be as in Figure 22.B.5. As for externalities, suppose that there is a single good and that the utility functions of two consumers are $u_1(x_1) = x_1$ and $u_2(x_1, x_2) = x_2/x_1$. Then the UPS is as in Figure 22.B.6 (see Appendix A of Chapter 11 for more on nonconvexities due to externalities). ■

Examples 22.B.4 and 22.B.5 have provided instances where the UPS is nonconvex. There is a procedure that permits one, in principle, to convexify the UPS. It consists of allowing the policy maker to randomize over her set of feasible policies. If random outcomes are evaluated by the different agents according to their expected utility (see

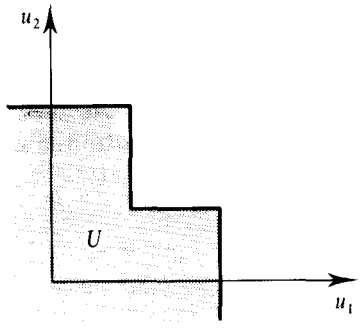


Figure 22.B.5
A nonconvex utility possibility set for a first-best locational problem (Example 22.B.5).

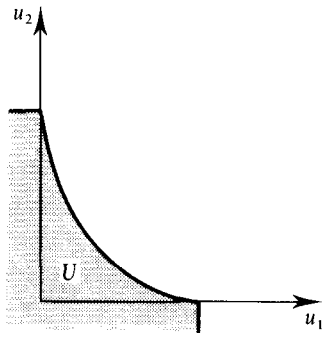


Figure 22.B.6
A nonconvex utility possibility set for a first-best problem with externalities (Example 22.B.5).

Chapter 6), then the (expected, or ex ante) UPS is convex since it is just the set of convex combinations of the utility vectors in the UPS associated with deterministic policies. There is no general theoretical reason to prevent the policy making from randomizing. On the other hand, the *practical* admissibility of stochastic policies cannot be decided on a priori grounds either.

We conclude this section with a final example [borrowed from Atkinson (1973)] that highlights the contrast between first-best and second-best problems.

Example 22.B.6: *Unproductive Taxation.* Suppose that there are two commodities and two consumers. We call the first commodity "labor", or leisure, and the second the "consumption good." There is a total of one unit of labor which is entirely owned by the first consumer. The consumption good can be produced by the first consumer from labor at a constant marginal cost of 1 (there is also free disposal). The first consumer has a utility function $u_1(x_{11}, x_{21})$ and the second has $u_2(x_{22})$. In Figure 22.B.7 we illustrate the construction of the first-best Pareto frontier for this model. Suppose that u_1 is given. Then, subject to attaining the level of utility u_1 for consumer 1, we want to give to consumer 2 as much utility as possible. If consumer 1 gets $(\bar{x}_{11}, \bar{x}_{21})$ then the labor supply is $1 - \bar{x}_{11}$ and the amount of consumption good available for consumer 2 is $1 - \bar{x}_{11} - \bar{x}_{21}$. Thus, we should first determine $(\bar{x}_{11}, \bar{x}_{21})$ by minimizing $x_{11} + x_{21}$ subject to $u_1(x_{11}, x_{21}) \geq u_1$, and then let $u_2 = u_2(1 - \bar{x}_{11} - \bar{x}_{21})$.

We now study the second-best problem where consumer 1 cannot be forced to supply labor. The only available policy instrument for providing consumption good

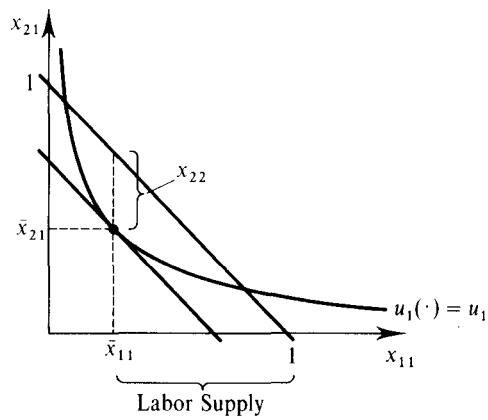
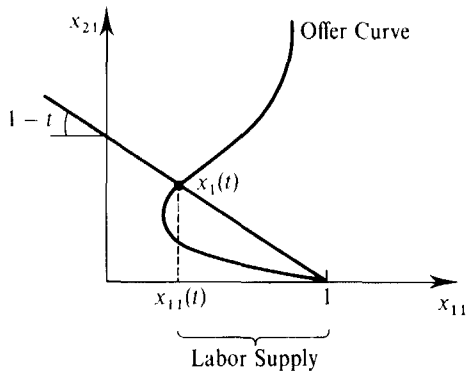


Figure 22.B.7
Construction of the first-best Pareto frontier for Example 22.B.6.



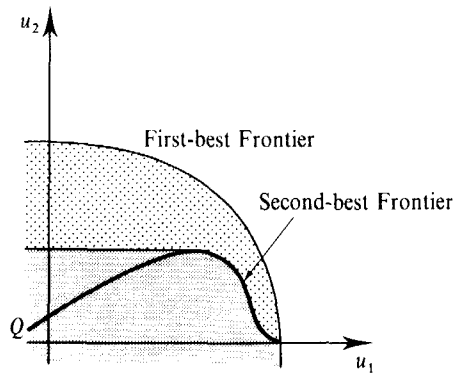


Figure 22.B.8 (left) Construction of the second-best Pareto frontier for Example 22.B.6.

Figure 22.B.9 (right) First-best and second-best utility possibility sets for the unproductive taxation example (Example 22.B.6).

to consumer 2 is a linear tax $t(1 - x_{11})$ on whatever amount of labour the first consumer *decides to supply* given the tax rate. The construction of the second-best frontier is illustrated in Figure 22.B.8. For $t \geq 0$, consumer 1 will choose x_{11} so as to maximize $u_1(x_{11}, (1 - t)(1 - x_{11}))$. Observe that this is as if she had chosen the point in her offer curve corresponding to the price vector $(1, 1/(1 - t))$. Denote this point by $x_1(t) = (x_{11}(t), x_{21}(t))$. The utility of consumer 2 is then $u_2(t(1 - x_{11}(t)))$.

The first-best and second-best UPS are displayed in Figure 22.B.9.[6] In the second-best case the figure also depicts the locus of utility pairs $Q \subset \mathbb{R}^I$ obtained as t ranges from 0 to 1, that is,

$$Q = \{(u_1(x_1(t)), u_2(t(1 - x_{11}(t)))) \in \mathbb{R}^2 : 0 \leq t \leq 1\}.$$

Note that Q does not coincide with the Pareto set of the second-best UPS because it exhibits a characteristic nonmonotonicity. The economic intuition underlying it is clear: if t is low, consumer 2 will get very little of the consumption good; but if t is very high, the situation is not much better. Consumer 2 will now get a large fraction of the labor supplied by consumer 1, but for precisely this reason not much labor will be supplied by consumer 1. ■

We can distill yet another lesson from Example 22.B.6. We see in Figure 22.B.9 that it is quite possible for the first-best and second-best Pareto frontiers to have some points in common; that is, there may well be second-best Pareto optima that are first-best Pareto optima. Yet Figure 22.B.9 tells us that it would be quite silly to select a point in the second-best Pareto frontier merely according to the criterion of proximity to the first-best frontier. The resulting selection may be distributionally very biased.[7] The investigation of more sensible selection criteria will be the purpose of Section 22.C.

6. Again, the second-best frontier may or may not be convex.

7. We may add that it may also be uninteresting from the point of view of policy: in Figure 22.B.9 the only second-best policy that yields a first-best result is $t = 0$, that is, no policy at all!

22.C Social Welfare Functions and Social Optima

In Section 22.B we described the constraint set of the policy maker, or social planner. The next question is which particular policy is to be selected. The application of the Pareto principle eliminates any policy that leads to utility vectors not in the Pareto frontier. Yet this still leaves considerable room for choice,[8] which, by necessity, must now involve trading off the utility of some agent against that of others. In this section we assume that the policy maker has an explicit and consistent criterion to carry off this task. Specifically, we assume that this criterion is given by a *social welfare function* $W(u) = W(u_1, \ldots, u_I)$ that aggregates individuals' utilities into social utilities. We can imagine that $W(u)$ reflects the distributional value judgments underlying the decisions of the policy maker.[9] In Section 22.E (and subsequent ones) we will discuss a somewhat different approach, one that puts more emphasis on the bargaining, or arbitration, aspects of the determination of the final policy selection.

In the current section, we refrain from questioning the assumption of *interpersonal comparability of utility*, which is implicit in our use of levels of individual utility as arguments in the aggregator function $W(u_1, \ldots, u_I)$. Section 22.D, which links with the analysis of Chapter 21, is devoted to investigating this matter.

Thus, for a given social welfare function $W(\cdot)$ and utility possibility set $U \subset \mathbb{R}^I$, the policy maker's problem is

$$\begin{aligned} \text{Max} \quad & W(u_1, \ldots, u_I) \\ \text{s.t.} \quad & (u_1, \ldots, u_I) \in U. \end{aligned} \tag{22.C.1}$$

A vector of utilities, or the underlying policies, solving problem (22.C.1) is called a *social optimum*. If the problem has a second-best nature, and we want to emphasize this fact, then we may refer to a *constrained social optimum*.

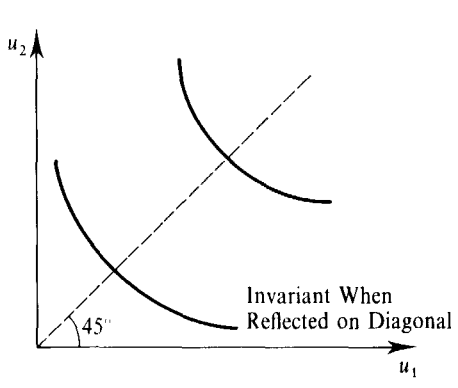
We now present and discuss some of the interesting properties that a social welfare function (SWF) may, or may not, satisfy.

(i) *Nonpaternalism.* This first property is already implicit in the concept itself of a SWF. It prescribes that in the expression of social preferences only the individual utilities matter: Two alternatives that are considered indifferent by every agent should also be socially indifferent. The planner does not have direct preferences on the final alternatives.

(ii) *Paretian property.* Granted the previous property, the *Paretian* property is an uncontroversial complement to it. It simply says that $W(\cdot)$ is increasing; that is, if $u_i' \geq u_i$ for all i, then $W(u') \geq W(u)$, and if $u_i' > u_i$ for all i, then $W(u') > W(u)$. We also say that $W(\cdot)$ is *strictly Paretian* if it is strictly increasing; that is, if $u_i' \geq u_i$ for all i and $u_i' > u_i$ for at least one i, then $W(u') > W(u)$. If $W(\cdot)$ is strictly Paretian then a solution to (22.C.1) is necessarily a Pareto optimum.

8. Only exceptionally will the Pareto frontier consist of a single point. Recall also that, as we saw in Example 22.B.3, in second-best situations with few instruments, the requirement of Pareto optimality may not succeed in ruling out many policies.

9. This approach to welfare economics was first taken by Bergson (1938) and Samuelson (1947).



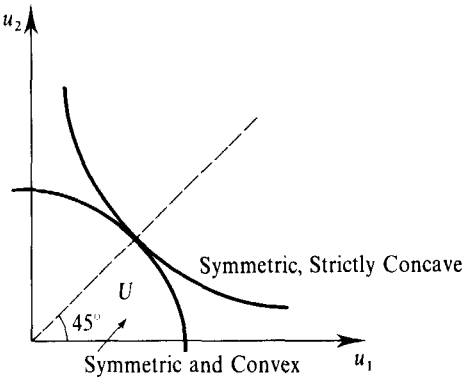


Figure 22.C.1 (left)
A symmetric social welfare function.

Figure 22.C.2 (right)
The optimum of a symmetric, strictly concave social welfare function on a symmetric and convex utility possibility set is egalitarian.

(iii) *Symmetry.* The *symmetry* property asserts that in evaluating social welfare all agents are on the same footing. Formally, $W(\cdot)$ is symmetric if $W(u) = W(u')$ whenever the entries of the vector u [e.g., $u = (2, 4, 5)$] constitute a permutation of the entries of the vector u' [e.g., $u' = (4, 5, 2)$]. In other words, the names of the agents are of no consequence, only the frequencies of the different utility values matter. The indifference curves of a symmetric $W(\cdot)$ are represented in Figure 22.C.1 for a two-agent case. Geometrically, each indifference curve is symmetric with respect to the diagonal. Note also that, because of this, if the indifference surfaces are smooth then the marginal rates of substitution at any $u = (u_1, \ldots, u_I)$ with identical coordinates are all equal to 1.

(iv) *Concavity.* Finally, a most important property is the *concavity* of $W(\cdot)$. We saw in Chapter 6 that, in the context of uncertainty, the (strict) concavity of a utility function implies an aversion to risk. Similarly, in the current welfare-theoretic context it can be interpreted as an *aversion to inequality* condition. A straightforward way to see this is to simply note that if $W(\cdot)$ is concave and $W(u) = W(u')$, then $W(\frac{1}{2}u + \frac{1}{2}u') \geq W(u)$ [with the inequality strict if $u \neq u'$ and $W(\cdot)$ is strictly concave]. Another is to observe that if the UPS is convex and symmetric, then the utility vector that assigns the same utility value to every agent is a social optimum of any symmetric and concave SWF (see Figure 22.C.2 and Exercise 22.C.1).[10] Thus, with convex UPSs and concave, symmetric SWFs some inequality is called for only if, as will typically be the case, the UPS is not symmetric.

It is to be emphasized that in general, and especially for second-best problems, the UPS may not be convex. This means that even if $W(\cdot)$ is concave the identification of social optima is not an easy task. A utility vector that satisfies the first-order conditions of problem (22.C.1) may not satisfy the second-order conditions or, if it does, it still may not constitute a global maximum.

We can gain further insights by discussing some important instances of social welfare functions.

10. The set $U \subset \mathbb{R}^I$ is symmetric if $u \in U$ implies $u' \in U$ for any $u' \in \mathbb{R}^I$ that differs from u only by a permutation of its entries. The interpretation of the symmetry property of a UPS is that there is no bias in the ability to produce utility for different agents. In other words, from the point of view of their possible contributions to social welfare, all agents are identical.

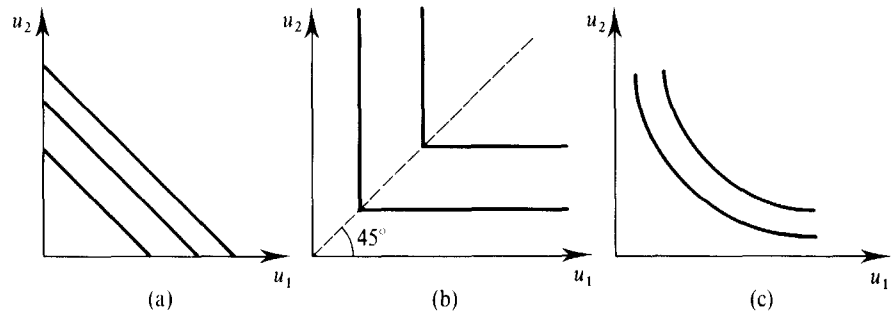


Figure 22.C.3
Social welfare functions.
(a) Purely utilitarian.
(b) Maximin or Rawlsian.
(c) Generalized utilitarian.

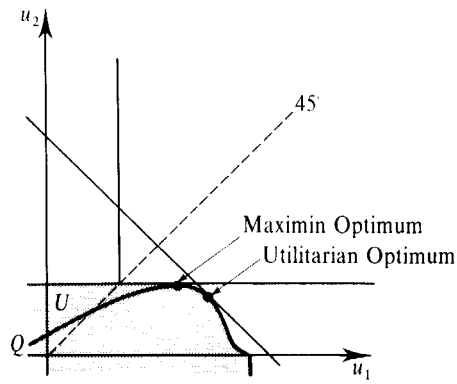
Example 22.C.1: *Utilitarian.* A SWF $W(u)$ is *purely utilitarian* if it has the form $W(u) = \sum_i u_i$ [or, in the nonsymmetric situation, $W(u) = \sum_i \beta_i u_i$]. In this case, the indifference hypersurfaces of $W(\cdot)$ are hyperplanes. They are represented in Figure 22.C.3(a). Note that $W(\cdot)$ is strictly Paretian.

In the purely utilitarian case, increases or decreases in individual utilities translate into identical changes in social utility. The use of the purely utilitarian principle goes back to the very birth of economics as a theoretical discipline. In Exercise 22.C.2 you are asked to develop an interpretation of the purely utilitarian SWF as the expected utility of a single individual "behind the veil of ignorance." Another line of defense, based also on expected utility theory, has been offered by Harsanyi (1955); see Exercise 22.C.3.

Because only the total amount of utility matters, the purely utilitarian SWF is *neutral* towards the inequality in the *distribution of utility*. It is important not to read into this statement more than it says. In particular, it does not say "distribution of wealth." For example, if there is a fixed amount of wealth to be distributed among individuals and these have strictly concave utility functions for wealth, then the purely utilitarian social optimum will be unique and distribute wealth so as to equalize the marginal utility of wealth across consumers. If, say, the utility functions are identical across individuals then this will choose as the unique social optimum the vector in the Pareto frontier that assigns the same utility to every agent (see Exercise 22.C.1 for generalizations). ■

Example 22.C.2: *Maximin.* A SWF is of *maximin* or *Rawlsian type* [because of Rawls (1971)] if it has the form $W(u) = \text{Min}\{u_1, \ldots, u_I\}$ [or, in the nonsymmetric case, $W(u) = \text{Min}\{\beta_1 u_1, \ldots, \beta_I u_I\}$]. In other words, social utility equals the utility value of the worst-off individual. It follows that the social planning problem becomes one of maximizing the utility of the worst-off individual.[11] The (L-shaped) indifference curves of the maximin SWF are represented in Figure 22.C.3(b).

11. One could refine this criterion by adopting a *lexical*, or *serial*, maximin decision rule. First maximize the utility of the worst-off, then choose among the solutions of this first problem by maximizing the utility of the next worst-off, and so on. With this, the objectives of the policy maker can still be expressed by a *leximin* social welfare ordering of utility vectors, but the ordering is not continuous and cannot be represented by a SWF (compare with Example 3.C.1). Even so, the refinement is natural and important. For example, we are then guaranteed that the social optimum is a Pareto optimum. You are asked to show all this in Exercise 22.C.4. Note that the maximin SWF is Paretian but not strictly Paretian. This makes for some difficulties. In Figure 22.C.4 the



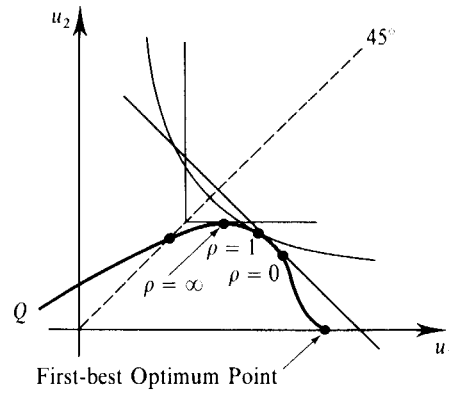


Figure 22.C.4 (left)
A maximin optimum for Example 22.B.6.

Figure 22.C.5 (right)
Range of generalized utilitarian optima for Example 22.B.6 and the constant elasticity SWF of Example 22.C.4 ($\rho \in [0, \infty]$).

It is reasonably intuitive that this concave SWF will have strong egalitarian implications. In fact, the preference for equality is quite extreme. Suppose, in effect, that $U \in \mathbb{R}^I$ is an arbitrary UPS and that $u \in U$ has all its coordinates equal. Then u fails to be the Rawlsian social optimum only if u is not Pareto optimal. Hence, if there is a $u = (u_1, \ldots, u_I)$ in the Pareto frontier of U with all its coordinates equal, then u is a maximin optimum. Note, in contrast, that for a purely utilitarian SWF we reached the social optimum at complete equality only in the case where U is convex and symmetric. In Figure 22.C.4, which continues the analysis of Example 22.B.6, we depict a situation where maximin optimization leads to the selection of a policy (a tax level) that does not yield complete equality. Nonetheless, even in this case, the purely utilitarian social optimum is significantly more unequal than the maximin optimum. ■

Example 22.C.3: *Generalized Utilitarian.* A SWF is *generalized utilitarian* if it has the form $W(u) = \sum_i g(u_i)$ [or, in the nonsymmetric case, $W(u) = \sum_i g_i(u_i)$], where $g(\cdot)$ is an increasing, concave function. The generalized utilitarian SWF is strictly Paretian and could be regarded as an instance of the purely utilitarian case where the individual utility functions $u_i(\cdot)$ have been replaced by $g(u_i(\cdot))$. This is not, however, a conceptually useful point of view. The point is precisely that, given the individual utility functions, there is a deliberate social decision to attach decreasing social weight to successive units of individual utility. The social indifference curves for this case are represented in Figure 22.C.3(c).

We can also verify in Figure 22.C.4 and 22.C.5 that the equality implications of the generalized utilitarian SWF are intermediate between those of the purely utilitarian and of the maximin SWFs. ■

Example 22.C.4: *Constant Elasticity.* An instance of generalized utilitarian functions that is very useful in applications is provided by the family defined by social utility functions $g(\cdot)$ whose marginal utilities have constant elasticity. This is a family in which attitudes towards inequality can be adjusted by means of a single parameter $\rho \geq 0$.

point at the boundary of U with equal coordinates is a maximin optimum but not a Pareto optimum. In the figure we have selected as "maximin optimum" the leximin optimum (which, by definition, is a maximin optimum itself).

For the rest of the example, individual utility values are restricted to be nonnegative. Then, for any $\rho \geq 0$, we let

$$g_\rho(u_i) = (1-\rho)u_i^{1-\rho} \qquad \text{if } \rho \neq 1,$$

and

$$g_\rho(u_i) = \ln u_i \qquad \text{if } \rho = 1.$$

Note that, as claimed, the elasticity of $g'_\rho(u_i)$ is constant because we have $u_i g''_i(u_i)/g'_\rho(u_i) = -\rho$ for all values u_i. Taking into account that, for $\rho \neq 1$, $h(W) = [1/(1-\rho)]W^{1/(1-\rho)}$ is an increasing transformation of W, we can represent the generalized utilitarian social preferences in a particularly convenient manner as

$$W_\rho(u) = (\textstyle\sum_i u_i^{1-\rho})^{1/1-\rho} \qquad \text{for } \rho \neq 1,$$

and

$$W_\rho(u) = \textstyle\sum_i \ln u_i \qquad \text{for } \rho = 1.$$

Thus, we obtain the CES functions that are well known from demand and production theories (see Exercises 3.C.6 and 5.C.10, respectively). Note that for $\rho = 0$ we get $W_0(u) = \sum_i u_i$, the purely utilitarian case, and as $\rho \to \infty$ we get $W_\rho(u) \to \text{Min}\{u_1, \ldots, u_I\}$, the maximin case. (See Exercise 22.C.5.)

In Figure 22.C.5 we depict the range of solutions to Example 22.B.6 as we vary ρ. We see that as the aversion to inequality increases (that is, as $\rho \to \infty$) the optimal tax rate increases. Note, however, that even for very high ρ we do not approach complete equality. On the other hand, none of these second-best solutions corresponds to the point in the Pareto frontier that is also Pareto optimal for the first-best problem. The latter distributes utility so unequally that the equity considerations underlying any symmetric and concave SWF leads us to sacrifice some first-best efficiency for an equity gain. ■

The Compensation Principle

We could ask ourselves to what an extent we can do welfare economics *without* social welfare functions. If the purpose of the SWF is the determination of optimal points in a given Pareto frontier, then resorting to them seems indispensable. This is the usage of social welfare functions that we have emphasized up to now; but in practice this is not the only usage. Often, the policy problem is given to us as one of choosing among several different utility possibility sets; these may correspond, for example, to the UPS associated with different levels of a basic policy variable.[12] If we have a social welfare function $W(\cdot)$, then the choice among two utility possibility sets U and U' should be determined by comparing the social utility of the optimum in U with that of the optimum in U'. However, even if there is no explicit social welfare function one may attempt to say something meaningful about this problem using revealed preference-like ideas. This is the approach underlying the *compensation principle* (already encountered in Sections 4.D and 10.E).

Let us first take the simplest case: that in which we have two utility possibility sets such that $U \subset U'$. Then one is very tempted to conclude that U' should be preferred to U. This would certainly be the case if the points that would be chosen

12. Formally, we can reduce this problem to the previous one by considering the overall UPS formed by the union of the UPSs over which we have to choose. But this may not be the most convenient thing to do because it loses the sequential presentation of the problem (first choose among UPS, then choose the utility vector).

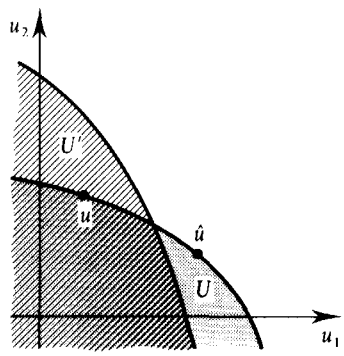


Figure 22.C.6
U' passes the weak compensation test over (U, u).

within each of U and U' were the optima of a social welfare function. But even if no social welfare function is available the set U' might still be considered superior to U according to the following *strong compensation test*: For any possible $u \in U$ there is a $u' \in U'$ such that $u'_i \geq u_i$ for every i. That is, wherever we are in U it is possible to move to U' and compensate agents in a manner that insures that every agent is made (weakly) better off by the change to U'. If the compensation is actually made, so that every agent will indeed be made better off by a switch from U to U', there is no doubt that the switch should be recommended. But if compensation will not occur, matters are not so clear: By choosing U' over U based only on a *potential* compensation we are neglecting quite drastically any distributional implication of the policy change. In fact, it is even possible that the change leads to a purely egalitarian worsening (see Exercise 22.C.6).

Recall from Section 10.D that in the quasilinear case we always have $U \subset U'$ or $U' \subset U$. This is because the boundaries of these sets are hyperplanes determined by the unit vector (hence parallel). In addition, this property also guarantees that the strong compensation criterion (which in Sections 3.D and 10.E we called simply the compensation criterion) coincides with the choice we would make using a purely utilitarian social welfare function. In this quasilinear case, therefore, the strong compensation criterion does not neglect distributional issues to a larger extent than do purely egalitarian social welfare functions.

Matters are more delicate when we compare two utility possibility sets U and U' which are such that one is not included in the other, that is, whose frontiers cross (see Figure 22.C.6). Suppose that we know that the outcome with utility possibility set U is the vector $u \in U$, and that we are considering a move to U'.[13] If $u \in U'$, and we were to allocate utility optimally in U' according to a social welfare function, then the move to U' would be advisable. More generally, whenever $u \in U'$, the move from (U, u) to U' passes the following *weak compensation test*: There is a $u' \in U'$ such that $u'_i \geq u_i$ for every i. That is, given that we know that the outcome at U is u, we could move to U' and compensate every agent in a manner that makes every agent (weakly) better off. In Figure 22.C.6, U' passes the test with respect to (U, u) but not with respect to $(U, \hat{u})$.

Again, if the compensation is actually paid, then the weak compensation criterion

13. For example, the original U could correspond to some underlying economy and u could be the utility values of a market equilibrium.

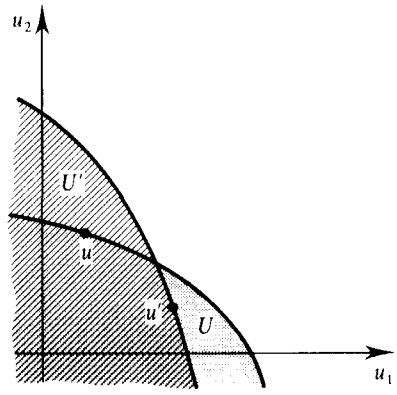


Figure 22.C.7
A paradox: U' passes the weak compensation test over (U, u), and U passes the weak compensation test over (U', u').

carries weight. If it is not paid, then it is subject to two serious criticisms. The first is the same as before (it disregards distributional consequences). The second is that it may lead to paradoxes. As in Figure 22.C.7, it is possible to have two utility possibility sets U and U', with respective outcomes $u \in U$ and $u' \in U'$, such that U' passes the weak compensation test over (U, u) and U passes the weak compensation test over (U', u'). In Exercise 22.C.7 you are asked to provide a more explicit example of this possibility in an economic context. Further elaborations are contained in Exercise 22.C.8.

22.D Invariance Properties of Social Welfare Functions

In this section, we probe deeper into the meaning of the comparisons of individual utilities implicit in the definition of a social welfare function. The significance of the matter derives from the fact that whereas a policy maker may be able to identify individual cardinal utility functions (from revealed risk behavior, say), it may actually do so but only up to a choice of origins and units. Fixing these parameters unavoidably involves making value judgments about the social weight of the different agents. It is therefore worth examining the extent to which such judgments may be avoided. Thus, following an approach to the problem taken by d'Aspremont and Gevers (1977), Roberts (1980), and Sen (1977), we explore such questions as: What are the implications for social decisions of requiring that social preferences be independent of the units, or the origins, of individual utility functions?[14]

To answer these types of questions, we need to contemplate the dependence of social preferences on profiles of individual utility functions. Thus, the social welfare functionals introduced in Chapter 21 provide a natural starting point for our analysis. However, we modify their definition slightly by specifying that individual characteristics are given to us in the form of individual utility functions $\tilde{u}_i(\cdot)$ rather than as individual preference relations.

From now on we are given a set of alternatives X. We denote by $\mathscr{U}$ the set of all possible utility functions on X, and by $\mathscr{R}$ the set of all possible rational (i.e., complete and transitive) preference relations on X.

14. In addition to the previous references, you can consult Moulin (1988) for a succinct presentation of the material of this section.

Definition 22.D.1: Given a set X of alternatives, a *social welfare functional* $F: \mathscr{U}^I \to \mathscr{R}$ is a rule that assigns a rational preference relation $F(\tilde{u}_1, \ldots, \tilde{u}_I)$ among the alternatives in the domain X to every possible profile of individual utility functions $(\tilde{u}_1(\cdot), \ldots, \tilde{u}_I(\cdot))$ defined on X. The strict preference relation derived from $F(\tilde{u}_1, \ldots, \tilde{u}_I)$ is denoted $F_p(\tilde{u}_1, \ldots, \tilde{u}_I)$.[15]

As in Chapter 21, we will concern ourselves only with social welfare functionals that are Paretian.

Definition 22.D.2: The social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ satisfies the (weak) *Pareto property*, or is *Paretian*, if, for any profile $(\tilde{u}_1, \ldots, \tilde{u}_I) \in \mathscr{U}^I$ and any pair $x, y \in X$, we have that $\tilde{u}_i(x) \geq \tilde{u}_i(y)$ for all i implies $x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, y$, and also that $\tilde{u}_i(x) > \tilde{u}_i(y)$ for all i implies $x\, F_p(\tilde{u}_1, \ldots, \tilde{u}_I)\, y$.

The first issue to explore is the relationship between these social welfare functionals and the social welfare functions of Section 22.C. A social welfare function $W(\cdot)$ assigns a social utility *value* to profiles $(u_1, \ldots, u_I) \in \mathbb{R}^I$ of individual utility *values*, whereas a social welfare functional assigns social *preferences* to profiles $(\tilde{u}_1, \ldots, \tilde{u}_I)$ of individual utility *functions* (or, in Section 21.C, of individual preference relations). From a social welfare function $W(\cdot)$ we can generate a social welfare functional simply by letting $F(\tilde{u}_1, \ldots, \tilde{u}_I)$ be the preference relation in X induced by the utility function $\tilde{u}(x) = W(\tilde{u}_1(x), \ldots, \tilde{u}_I(x))$. The converse may not be possible, however. In order to be able to "factor" a social welfare functional through a social welfare function, the following necessary condition must, at the very least, be satisfied. Suppose that the profile of utility functions changes, but that the profiles of utility values for two given alternatives remain unaltered; then the social ordering among these alternatives should not change (since the value given by the social welfare function to each alternative has not changed). That is, the social ordering among two given alternatives should depend only on the profiles of individual utility values for these alternatives. Apart from being formulated in terms of utilities, this property is analogous to the pairwise independence condition for social welfare functionals (Definition 21.C.3). We keep the same term and state the condition formally in Definition 22.D.3.

Definition 22.D.3: The social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ satisfies the *pairwise independence condition* if, whenever $x, y \in X$ are two alternatives and $(\tilde{u}_1, \ldots, \tilde{u}_I) \in \mathscr{U}^I$, $(\tilde{u}'_1, \ldots, \tilde{u}'_I) \in \mathscr{U}^I$ are two utility function profiles with $\tilde{u}_i(x) = \tilde{u}'_i(x)$ and $\tilde{u}_i(y) = \tilde{u}'_i(y)$ for all i, we have

$$x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, y \quad \Leftrightarrow \quad x\, F(\tilde{u}'_1, \ldots, \tilde{u}'_I)\, y.$$

The necessary pairwise independence condition is almost sufficient: In Proposition 22.D.1 we now see that if the number of alternatives is greater than 2, and the Pareto and pairwise independence conditions are satisfied, then we can derive from the social welfare functional a social preference relation defined on profiles $(u_1, \ldots, u_I) \in \mathscr{R}^I$ of utility values.[16] A standard continuity condition then allows us to represent this

15. That is, $x\, F_p(\tilde{u}_1, \ldots, \tilde{u}_I)\, y$ if $x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, y$ but not $y\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, x$.

16. In Exercise 22.D.1 you can find examples showing that the Pareto condition and the restriction on the number of alternatives cannot be dispensed with for the result of Proposition 22.D.1.

preference relation by means of a function $W(u_1, \ldots, u_I)$, thereby yielding a social welfare function.

Proposition 22.D.1: Suppose that there are at least three alternatives in X and that the Paretian social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ satisfies the pairwise independence condition. Then there is a rational preference relation $\succsim$ defined on $\mathbb{R}^I$ [that is, on profiles $(u_1, \ldots, u_I) \in \mathbb{R}^I$ of individual utility values] that generates $F(\cdot)$. In other words, for every profile of utility functions $(\tilde{u}_1, \ldots, \tilde{u}_I) \in \mathscr{U}^I$ and for every pair of alternatives $x, y \in X$ we have

$$x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, y \quad \Leftrightarrow \quad (\tilde{u}_1(x), \ldots, \tilde{u}_I(x)) \succsim (\tilde{u}_1(y), \ldots, \tilde{u}_I(y)).$$

Proof: The desired conclusion dictates directly how $\succsim$ should be constructed. Consider any pair of utility profiles $u = (u_1, \ldots, u_I) \in \mathbb{R}^I$ and $u' = (u'_1, \ldots, u'_I) \in \mathbb{R}^I$. Then we let $u \succsim u'$ if $x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, y$ for some pair $x, y \in X$ and a profile $(\tilde{u}_1, \ldots, \tilde{u}_I) \in \mathscr{U}^I$ with $\tilde{u}_i(x) = u_i$ and $\tilde{u}_i(y) = u'_i$ for every i. We argue first that the conclusion $u \succsim u'$, is independent of the particular two alternatives and the profile of utility functions chosen. Independence of the utility functions chosen is an immediate consequence of the statement of the pairwise independence condition. Proving independence of the pair chosen is a bit more delicate.

It suffices to show that if we have concluded that $u \succsim u'$ by means of a pair x, y then, for any third alternative z (recall that by assumption there are third alternatives), we obtain the same conclusion using the pairs x, z or z, y.[17] We carry out the argument for x, z (in Exercise 22.D.2 you are asked to do the same for z, y). To this effect, take a profile of utility functions $(\tilde{u}_1, \ldots, \tilde{u}_I) \in \mathscr{U}^I$ with $\tilde{u}_i(x) = u_i$, $\tilde{u}_i(y) = u'_i$, and $\tilde{u}_i(z) = u'_i$ for every i. Because we have concluded that $u \succsim u'$ using the pair x, y, we must have $x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, y$. By the Pareto property, we also have $y\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, z$. Hence, by the transitivity of $F(\tilde{u}_1, \ldots, \tilde{u}_I)$, we obtain $x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, z$, which is the property we wanted.

It remains to prove that $\succsim$ is complete and transitive. Completeness follows simply from the fact that the preference relation $F(\tilde{u}_1, \ldots, \tilde{u}_I)$ is complete for any $(\tilde{u}_1, \ldots, \tilde{u}_I) \in \mathscr{U}^I$. As for transitivity, let $u \succsim u' \succsim u''$, where $u, u', u'' \in \mathbb{R}^I$. Take three alternatives $x, y, z \in X$ and a profile of utility functions $(\tilde{u}_1, \ldots, \tilde{u}_I) \in \mathscr{U}^I$ with $\tilde{u}_i(x) = u_i$, $\tilde{u}_i(y) = u'_i$, and $\tilde{u}_i(z) = u''_i$ for every i. Since $u \succsim u'$ and $u' \succsim u''$, it must be that $x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, y$ and $y\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, z$. Because of the transitivity of $F(u_1, \ldots, u_I)$, this implies $x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, z$, and so $u \succsim u''$. Hence, $\succsim$ is transitive. ■

By the Pareto condition, the social preference relation $\succsim$ obtained in Proposition 22.D.1 is monotone. You are asked to show this formally in Exercise 22.D.3.

Exercise 22.D.3: Show that if the social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ satisfies the Pareto property, then a social preference relation $\succsim$ on utility profiles for which the

17. Indeed, suppose that we initially used the pair (x, y). Consider any other pair (v, w). If $v = x$ or $w = y$ then we have just claimed that we get the same ordering between u and u'. Hence, let $v \neq x$ and $w \neq y$. If, in addition, $v \neq y$, then we reach the same ordering by the chain of replacements: $(x, y) \to (v, y) \to (v, w)$. Similarly, if $w \neq x$ we can use $(x, y) \to (x, w) \to (v, w)$. There remains the case $(v, w) = (y, x)$. Here we use a third alternative, z, and the chain $(x, y) \to (x, z) \to (y, z) \to (y, x)$.

conclusion of Proposition 22.D.1 holds must be monotone in the sense that if $u' \geq u$ then $u' \gtrsim u$, and if $u' \gg u$ then $u' \succ u$.

The social preference relation $\gtrsim$ on $\mathbb{R}^I$ obtained in Proposition 22.D.1 need not be continuous or representable by a utility function. Consider, for example, a lexical dictatorship (say that there are two agents and let $u \succ u'$ if $u_1 > u_1'$ or if $u_1 = u_1'$ and $u_2 > u_2'$) and recall from Example 3.C.1 that this type of ordering is not representable by a utility function. Nonetheless, we want to focus on social welfare functions and so from now on we will simply assume that we deal only with social welfare functionals that, in addition to the assumptions of Proposition 22.D.1, yield a continuous social preference relation $\gtrsim$ on $\mathbb{R}^I$. As in Section 3.C, such a social preference relation can then be represented by a utility function: in fact, a continuous one. This is then our social welfare function $W(u_1, \ldots, u_I)$. Note that any increasing, continuous transformation of $W(\cdot)$ is also an admissible social welfare function.

In summary, we have seen that the existence of a social welfare function generating a given social welfare functional amounts, with some minor qualifications, to the satisfaction of the pairwise independence condition by the social welfare functional. Therefore, we will concern ourselves from now on with a social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ that can be generated from an increasing and continuous social welfare function $W: \mathbb{R}^I \to \mathbb{R}$, or equivalently, from a monotone and continuous rational preference relation $\gtrsim$ on $\mathbb{R}^I$. We will discover that, in this context, natural utility invariance requirements on the social welfare functional have quite drastic effects on the form that we can choose for $W(\cdot)$ and, therefore, on the social welfare functional itself.

Definition 22.D.4: We say that the social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ is *invariant to common cardinal transformations* if $F(\tilde{u}_1, \ldots, \tilde{u}_I) = F(\tilde{u}_1', \ldots, \tilde{u}_I')$ whenever the profiles of utility functions $(\tilde{u}_1, \ldots, \tilde{u}_I)$ and $(\tilde{u}_1', \ldots, \tilde{u}_I')$ differ only by a common change of origin and units, that is, whenever there are numbers $\beta > 0$ and α such that $\tilde{u}_i(x) = \beta \tilde{u}_i'(x) + \alpha$ for all i and $x \in X$. If the invariance is only with respect to common changes of origin (i.e., we require $\beta = 1$) or of units (i.e., we require $\alpha = 0$), then we say that $F(\cdot)$ is *invariant to common changes of origin* or *of units*, respectively.

It is hard to quarrel with the requirement of invariance with respect to common cardinal transformations. Even if the policy maker has the ability to compare the utilities of different agents, the notion of an absolute unit or an absolute zero is difficult to comprehend.

We begin by analyzing the implications of invariance with respect to common changes of origin. Suppose that the social welfare functional is generated from the social welfare function $W(\cdot)$. We claim that the invariance with respect to common changes of origin can hold only if $W(u) = W(u')$ implies $W(u + \alpha e) = W(u' + \alpha e)$ for all profiles of utility values $u \in \mathbb{R}^I$, $u' \in \mathbb{R}^I$ and $\alpha \in \mathbb{R}$, where $e = (1, \ldots, 1)$ is the unit vector. Indeed, let $W(u) = W(u')$ and $W(u + \alpha e) < W(u' + \alpha e)$. Consider a pair $x, y \in X$ and profile $(\tilde{u}_1, \ldots, \tilde{u}_I) \in U^I$ with $\tilde{u}_i(x) = u_i$ and $\tilde{u}_i(y) = u_i'$ for every i. Then $x\, F(\tilde{u}_1, \ldots, \tilde{u}_I)\, y$. However, $x\, F(\tilde{u}_1', \ldots, \tilde{u}_I')\, y$ does not hold when $\tilde{u}_i'(\cdot) = \tilde{u}_i(\cdot) + \alpha$, contradicting the invariance to common changes of origin.

Geometrically, the assertion that $W(u) = W(u')$ implies $W(u + \alpha e) = W(u' + \alpha e)$ says that the indifference curves of $W(\cdot)$ are parallel with respect to e—they are

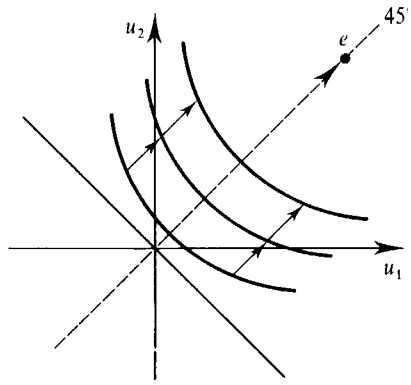


Figure 22.D.1
Indifference map of a social welfare function invariant to identical changes of utility origins.

obtained from each other by translations along the e direction (see Figure 22.D.1). In Proposition 23.D.2 [due to Roberts (1980)], we show that this property has an important implication: up to an increasing transformation, the social welfare function can be written as a sum of a purely utilitarian social welfare function and a dispersion term.

Proposition 22.D.2: Suppose that the social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ is generated from a continuous and increasing social welfare function. Suppose also that $F(\cdot)$ is invariant to common changes of origins. Then the social welfare functional can be generated from a social welfare function of the form

$$W(u_1, \ldots, u_I) = \bar{u} - g(u_1 - \bar{u}, \ldots, u_I - \bar{u}), \tag{22.D.1}$$

where $\bar{u} = (1/I)\sum_i u_i$.

Moreover, if $F(\cdot)$ is also independent of common changes of units, that is, fully invariant to common cardinal transformations, then $g(\cdot)$ is homogeneous of degree one on its domain: $\{s \in \mathbb{R}^I: \sum_i s_i = 0\}$.

Proof: By assumption the social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ can be generated by a continuous and monotone preference relation $\succsim$ on $\mathbb{R}^I$. Moreover the invariance to identical changes of units implies that if $u \sim u'$ then $u + \alpha e \sim u' + \alpha e$ for any $\alpha \in \mathbb{R}$.

We now construct a particular utility function $W(\cdot)$ for $\succsim$. Because of continuity and monotonicity of $\succsim$ there is, for every $u \in \mathbb{R}^I$, a single number α such that $u \sim \alpha e$. Let $W(u)$ denote this number. That is, $W(u)$ is defined by $u \sim W(u)e$. (See Figure 22.D.2 for a depiction.) Because of the monotonicity of preferences, $W(\cdot)$ is a legitimate utility representation for $\succsim$.[18]

The first part of the proof will be concluded if we show that $W(u) - \bar{u}$ depends only on the vector of deviations $(u_1 - \bar{u}, \ldots, u_I - \bar{u}) = u - \bar{u}e$, that is, that if $u - \bar{u}e = u' - \bar{u}'e$ then $W(u) - \bar{u} = W(u') - \bar{u}'$. But this is true because $u \sim W(u)e$ and the invariance to common changes of origin imply that if $u - \bar{u}e = u' - \bar{u}'e$ then

$$u' = u + (\bar{u}' - \bar{u})e \sim W(u)e + (\bar{u}' - \bar{u})e = [W(u) + (\bar{u}' - \bar{u})]e$$

18. Up to here this is identical to the parallel construction in consumption theory carried out in Proposition 3.C.1. We refer to the proof of the latter for details.

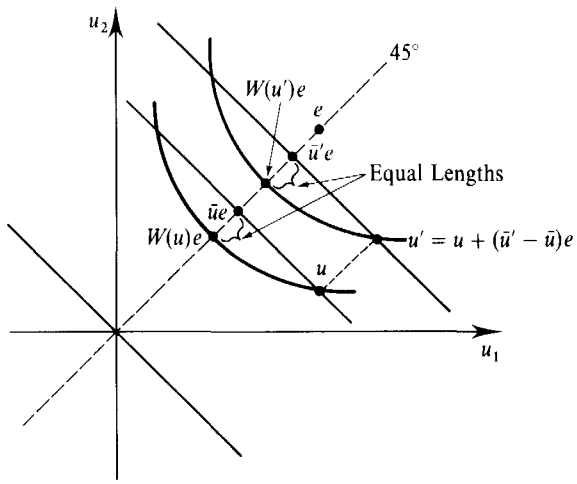


Figure 22.D.2
Construction of the social welfare function of form (22.D.1) for the invariant to identical changes of origin case.

and therefore, $W(u') = W(u) + (\bar{u}' - \bar{u})$ as we wanted. The construction is illustrated in Figure 22.D.2.[19]

To prove the second part, suppose that $F(\cdot)$ is also invariant to common changes of units. Because $F(\cdot)$ is generated from $W(\cdot)$, this can only happen if for every $u \sim u'$ and $\beta > 0$ we have $\beta u \sim \beta u'$. But then $u \sim W(u)e$ implies $\beta u \sim \beta W(u)e$, and so $W(\beta u) = \beta W(u)$ for any $u \in \mathbb{R}^I$ and $\beta > 0$. That is, $W(\cdot)$ is homogeneous of degree one, and since $g(\cdot)$ coincides with $-W(\cdot)$ on the domain where $\bar{u} = 0$, we conclude that $g(\cdot)$ is also homogeneous of degree one. ■

Going further, if the policy maker is not empowered with the ability to compare the absolute levels of utility across consumers, then the social welfare functional must satisfy more demanding invariance notions.

Definition 22.D.5: The social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ *does not allow interpersonal comparisons of utility* if $F(\tilde{u}_1, \ldots, \tilde{u}_I) = F(\tilde{u}'_1, \ldots, \tilde{u}'_I)$ whenever there are numbers $\beta_i > 0$ and α_i such that $\tilde{u}_i(x) = \beta_i \tilde{u}'_i(x) + \alpha_i$ for all i and x. If the invariance is only with respect to independent changes of origin (i.e., we require $\beta_i = 1$ for all i), or only with respect to independent changes of units (i.e., we require that $\alpha_i = 0$ for all i), then we say that $F(\cdot)$ is *invariant to independent changes of origins* or *of units*, respectively.

We have then Proposition 22.D.3.[20]

19. We can gain some intuition on the form of this utility function by noticing its similarity to the quasilinear representations in consumer theory. Here we can write any vector $u \in \mathbb{R}^I$ as $u = \bar{u}e + (u - \bar{u}e)$ and indifference sets can be obtained by parallel displacements in the direction e. In consumer theory we can write any vector $x \in \mathbb{R}^L$ as $x = (x_1, 0, \ldots, 0) + (0, x_2, \ldots, x_L)$ and indifference sets are parallel in the direction $(1, 0, \ldots, 0)$. Similarly, the conclusion in both cases is that there is a utility function that is linearly additive in the first term (i.e., in the direction in which indifference sets are parallel).

20. See d'Aspremont and Gevers (1977) for more results of this type.

Proposition 22.D.3: Suppose that the social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ can be generated from an increasing, continuous social welfare function. If $F(\cdot)$ is invariant to independent changes of origins, then $F(\cdot)$ can be generated from a social welfare function $W(\cdot)$ of the purely utilitarian (but possibly nonsymmetric) form. That is, there are constants $b_i \geq 0$, not all zero, such that

$$W(u_1, \ldots, u_I) = \sum_i b_i u_i \qquad \text{for all } i. \tag{22.D.2}$$

Moreover, if $F(\cdot)$ is also invariant to independent changes of units [i.e., if $F(\cdot)$ does not allow for interpersonal comparisons of utility], then F is dictatorial: There is an agent h such that, for every pair $x, y \in X$, $\tilde{u}_h(x) > \tilde{u}_h(y)$ implies $x F_p(\tilde{u}_1, \ldots, \tilde{u}_I)\, y$.

Proof: Suppose that $\succsim$ is the continuous preference relation on $\mathbb{R}^I$ that generates the given $F(\cdot)$. For a representation of the form (22.D.2) to exist, we require that the indifference sets of $\succsim$ be parallel hyperplanes. Since we already know from Proposition 22.D.2 that those sets are all parallel in the direction e, it suffices to show that they must be hyperplanes, that is, that if we take two $u, u' \in \mathbb{R}^I$ such that $u \sim u'$, then for $u'' = \frac{1}{2}u + \frac{1}{2}u'$ we also have $u'' \sim u \sim u'$.

The invariance of $F(\cdot)$ with respect to independent changes of origins means, in terms of $\succsim$, that for *any* $\alpha \in \mathbb{R}^I$ we have $u + \alpha \succsim u'' + \alpha$ if and only if $u \succsim u''$. Take $\alpha = \frac{1}{2}(u' - u)$. Then $u + \alpha = u''$ and $u'' + \alpha = u'$. Hence, $u \succsim u''$ if and only if $u'' \succsim u'$. If $u \succsim u''$ then $u'' \succsim u'$ and so $u'' \sim u$. If $u'' \succ u$ then $u' \succ u''$ which contradicts $u \sim u'$. We conclude that $u'' \sim u \sim u'$, as we wanted.

Once we know that indifference sets are parallel hyperplanes, the same construction as in the Proof of Proposition 22.D.2 will give us a $W(\cdot)$ of the form (22.D.2). In addition, the Pareto property yields $b_i \geq 0$ for all i.

Finally, suppose that $F(\cdot)$ is also invariant to independent changes of units. Then dictatorship follows simply. Choose an agent h with $b_h > 0$. Take $u, u' \in \mathbb{R}^I$ with $u_h > u'_h$. Then, by invariance to independent changes of units, we have that $\sum_i b_i u_i > \sum_i b_i u'_i$ if and only if $b_h u_h + \varepsilon \sum_{i \neq h} b_i u_i > b_h u'_h + \varepsilon \sum_{i \neq h} b_i u'_i$ for any $\varepsilon > 0$. Therefore, since $b_h u_h > b_h u'_h$ we get, by choosing $\varepsilon > 0$ small enough, that $\sum_i b_i u_i > \sum_i b_i u'_i$. Thus, agent h is a dictator (show, in Exercise 22.D.4, that in fact $b_i = 0$ for all $i \neq h$). ■

We point out that for the dictatorship conclusion of Proposition 22.D.3, it is not necessary that $F(\cdot)$ be generated from a social welfare function. It suffices that it be generated from a social preference relation on $\mathbb{R}^I$.

Proposition 22.D.3 (extended in the manner indicated in the last paragraph) has as a corollary the Arrow impossibility theorem of Chapter 21 (Proposition 21.C.1), which is, in this manner, obtained by a very different methodology. Indeed, suppose that $F(\cdot)$ is a social welfare functional defined, as was done in Chapter 21, on profiles of preference relations $(\succsim_1, \ldots, \succsim_I) \in \mathscr{R}^I$. Then we can construct a social welfare functional $F'(\cdot)$ defined on profiles of utility functions $(\tilde{u}_1, \ldots, \tilde{u}_I) \in \mathscr{U}^I$ by letting $F'(\tilde{u}_1, \ldots, \tilde{u}_I) = F(\succsim_1, \ldots, \succsim_I)$, where $\succsim_i$ is the preference relation induced by the utility function $\tilde{u}_i(\cdot)$. In Exercise 22.D.5 you are asked to verify, first, that $F'(\cdot)$ inherits the Paretian and pairwise independence conditions from $F(\cdot)$, second, that $F'(\cdot)$ does not allow for interpersonal comparisons of utility and, third, that a dictator for $F'(\cdot)$ is a dictator for $F(\cdot)$.

Other invariance properties of social welfare functionals have been found to be of interest. We mention two.

We say that the social welfare functional $F: \mathscr{U}^I \to \mathscr{R}$ is *invariant to common ordinal transformations* if $F(\tilde{u}_1, \ldots, \tilde{u}_I) = F(\tilde{u}'_1, \ldots, \tilde{u}'_I)$ whenever there is an increasing function $\gamma(\cdot)$ such that $\tilde{u}_i(x) = \gamma(\tilde{u}'_i(x))$ for every $x \in X$ and all i. The interpretation of this invariance is that although the social planner has no notion of individual utility scales she can, nonetheless, recognize that one individual is better off than another (but the question "by how much?" is meaningless). An example is provided by the social welfare functional induced by the symmetric Rawlsian social welfare function $W(u) = \text{Min}\{u_1, \ldots, u_I\}$. With this SWF, the ordering over policies depends only on the ability to determine the worse-off individual (see Exercise 22.D.8 for further elaboration).

We say that a social welfare function $W(\cdot)$ generating a given social welfare functional $F: \mathscr{U}^I \to \mathbb{R}$ is *independent of irrelevant individuals* if, when we split the set of agents into any two groups, the social preference among utility vectors in one of the groups is independent of the level at which we fix the utilities of the agents in the other group (we should add that, if so desired, the condition can be formulated directly in terms of the social welfare functional). This is a sensible requirement: tt says that the distributional judgments concerning the inhabitants of, say, California, should be independent of the individual welfare levels of the inhabitants of, say, Massachusetts.

As in the formally similar situation in consumer theory (Exercise 3.G.4), a social welfare function for $I > 2$ agents that is continuous, increasing, and independent of irrelevant individuals has, up to an increasing transformation, the *additively separable form* $W(u) = \sum_i g_i(u_i)$; that is, $W(u)$ is generalized utilitarian, possible nonsymmetric. Moreover, under weak conditions it is also true that the only social welfare functions that, up to increasing transformations, both admit an additively separable form and are invariant to common changes of origin are the utilitarian $W(u) = \sum_i b_i u_i$. Thus, from an invariance viewpoint we can arrive at the utilitarian form for a social welfare function by two roads: one, Proposition 22.D.3, is based on invariance to independent changes of origins; the other, just mentioned, is based on independence of irrelevant individuals and invariance to common changes of origins. See Maskin (1978) for more on this.

Example 22.D.1: Fix an alternative x^* and define a social welfare functional $F(\cdot)$ by associating to every profile of individual utility functions $(\tilde{u}_1, \ldots, \tilde{u}_I)$ the social preference relation generated by a utility function $V(x) = \sum_i g_i(\tilde{u}_i(x) - \tilde{u}_i(x^*))$. Then, informally, this social welfare functional is both invariant to independent changes of origins and independent of irrelevant individuals, but it is neither utilitarian nor dictatorial. Note, however, that this functional cannot be generated from a social welfare function because it is not pairwise independent: the social preference among two alternatives *may depend on the utility of the third alternative* x^*. ■

22.E The Axiomatic Bargaining Approach

In this section, we briefly review an alternative approach to the determination of *reasonable* social compromises. The role of a planner endowed with her own preferences is now replaced by that of an (implicit) *arbitrator* who tries to distribute the gains from trade or, more generally, from cooperation in a manner that reflects "fairly" the bargaining strength of the different agents. The origin of the theory is game-theoretic. However, it sidesteps the construction of explicit noncooperative

bargaining games (such as those considered in Appendix A of Chapter 9) by adopting an *axiomatic* point of view. Thus, the approach is more related to ideas of cooperative game theory (as reviewed in Appendix A of Chapter 18).[21]

For current purposes, the description of a *bargaining problem* among I agents is composed of two elements: a *utility possibility* set $U \subset \mathbb{R}^I$ and a *threat*, or *status-quo*, *point* $u^* \in U$. The set U represents the allocations of utility that can be settled on if there is cooperation among the different agents. The point u^* is the outcome that will occur if there is a breakdown of cooperation. Note that cooperation requires the unanimous participation of all agents, in which case, to repeat, the available utility options are given by $U \subset \mathbb{R}^I$. If one agent does not participate, then the only possible outcome is the vector u^*. This setup is completely general with two agents and, because of this, the two-agent case is our central reference case in this section. With more than two agents, the assumption is a bit extreme, since we may want to allow for the possibility of partial cooperation. We take up this possibility in Section 22.F.

Throughout this section we assume that $U \subset \mathbb{R}^I$ is convex and closed and that it satisfies the free disposal property $U - R^I_+ \subset U$ (i.e. if $u' \leq u$ and $u \in U$ then $u' \in U$). As in Definition 22.B.1, $U \subset \mathbb{R}^I$ could be generated from a set of underlying alternatives X, which could well include lotteries over deterministic outcomes.[22] For simplicity we also assume that u^* is interior to U and that $\{u \in U: u \geq u^*\}$ is bounded.

Definition 22.E.1: A *bargaining solution* is a rule that assigns a solution vector $f(U, u^*) \in U$ to every bargaining problem (U, u^*).[23]

We devote the rest of this section to a discussion of some of the properties one may want to impose on $f(\cdot)$ and to a presentation of four examples of bargaining solutions: the *egalitarian*, the *utilitarian*, the *Nash* and the *Kalai–Smorodinsky solutions*. We should emphasize, however, that a strong assumption has already been built into the formalization of our problem: we are implicitly assuming that the solution depends on the set X of feasible alternatives only through the resulting utility values.

Definition 22.E.2: The bargaining solution $f(\cdot)$ is *independent of utility origins* (IUO), or *invariant to independent changes of origins*, if for any $\alpha = (\alpha_1 \ldots, \alpha_I) \in \mathbb{R}^I$ we have

$$f_i(U', u^* + \alpha) = f_i(U, u^*) + \alpha_i \qquad \text{for every } i$$

whenever $U' = \{(u_1 + \alpha_1, \ldots, u_I + \alpha_I): u \in U\}$.

The IUO property says that the bargaining solution does not depend on absolute scales of utility. From now on we assume that this property holds. Note that we therefore always have $f(U, u^*) = f(U - \{u^*\}, 0) + u^*$. This allows us to normalize our problems to $u^* = 0$. From now on we do so and simply write $f(U)$ for $f(U, 0)$.

21. For general introductions to the material of this section, see Roth (1979), Moulin (1988), and Thomson (1995).

22. In principle, the underlying set X and the corresponding utility functions on X could be different for different $U \subset \mathbb{R}^I$. For the theory that follows all that matters is the utility set U.

23. Thus, a bargaining solution is a choice rule in the sense of Chapter 1. If an underlying alternative set X is kept fixed and, therefore, the form of U, generated as in Definition 22.B.1, depends only on the utility functions, we can also regard the bargaining solution as a choice of function in the sense of Definition 21.E.1.

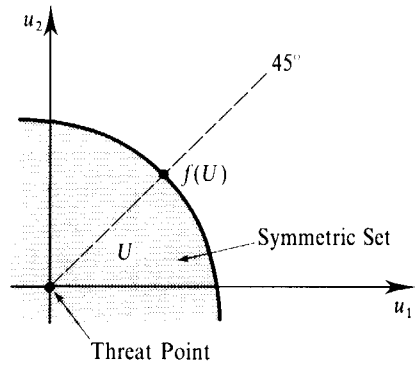


Figure 22.E.1
The symmetry property for bargaining solutions.

It should not be forgotten, however, that a change in the threat point (which will now show up as a change in U) will affect the point settled on.

Definition 22.E.3: The bargaining solution $f(\cdot)$ is *independent of utility units* (IUU), or *invariant to independent changes of units*, if for any $\beta = (\beta_1, \ldots, \beta_I) \in \mathbb{R}^I$ with $\beta_i > 0$ for all i, we have

$$f_i(U') = \beta_i f_i(U) \qquad \text{for every } i$$

whenever $U' = \{(\beta_1 u_1, \ldots, \beta_I u_I) : u \in U\}$.[24]

With independence of utility origins (implicitly assumed in Definition 22.E.3), independence of utility units tells us that, although the bargaining solution uses cardinal information on preferences, it does not in any way involve interpersonal comparisons of utilities.

Definition 22.E.4: The bargaining solution $f(\cdot)$ satisfies the *Pareto* property (P), or is *Paretian*, if, for every U, $f(U)$ is a (weak) Pareto optimum, that is, there is no $u \in U$ such that $u_i > f_i(U)$ for every i.

Definition 22.E.5: The bargaining solution $f(\cdot)$ satisfies the property of *symmetry* (S) if whenever $U \subset \mathbb{R}^I$ is a symmetric set (i.e., U remains unaltered under permutations of the axes;[25] see Figure 22.E.1), we have that all the entries of $f(U)$ are equal.

The interpretation of the symmetry property is straightforward: if, as reflected in U, all agents are identical, then the gains from cooperation are split equally.

Definition 22.E.6: The bargaining solution $f(\cdot)$ satisfies the property of *individual rationality* (IR) if $f(U) \geq 0$.

In words: the cooperative solution does not give any agent less than the threat point (recall also that, after normalization, we consider only sets U with $0 \in U$). It is a sensible property: if some agent got less than zero, then she would do better by opting out and bringing about the breakdown of negotiation.

The next property is more substantial.

24. Geometrically, U' is obtained from U by stretching the different axes by the rescaling factors $(\beta_1, \ldots, \beta_I)$.

25. More precisely, if $u \in U$ then $u' \in U$ for any u' differing from u only by a permutation of its entries.

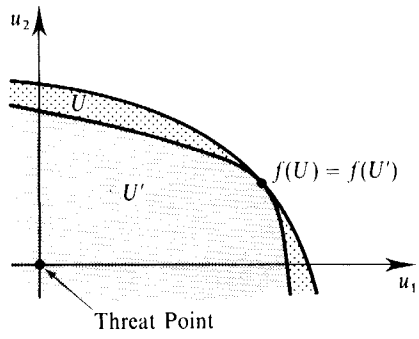


Figure 22.E.2
The property of independence of irrelevant alternatives for bargaining solutions.

Definition 22.E.7: The bargaining solution satisfies the property of *independence of irrelevant alternatives* (IIA) if, whenever $U' \subset U$ and $f(U) \in U'$, it follows that $f(U') = f(U)$.

The IIA condition says that if $f(U)$ is the "reasonable" outcome in U and we consider a U' that is smaller than U but retains the feasibility of $f(U)$, that is, we only eliminate from U "irrelevant alternatives," then $f(U)$ remains the reasonable outcome (see Figure 22.E.2). This line of justification would be quite persuasive if we could replace "reasonable" by "best." Indeed, if $f(U)$ has been obtained as the unique maximizer on U of some social welfare function $W(u)$, then the IIA condition is clearly satisfied [if $f(U)$ maximizes $W(\cdot)$ on U then it also maximizes $W(\cdot)$ on $U' \subset U$]. We note that while the converse is not true, it is nonetheless the case that, in practice, the interesting examples where IIA is satisfied involve the maximization of some SWF.

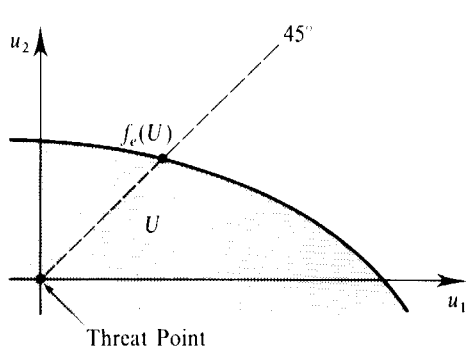
We proceed to present four examples of bargaining solutions. To avoid repetition, we put on record that all of them satisfy the Paretian, symmetry, and individual rationality properties (as well as, by the formulation itself, the independence of utility origins). You are asked to verify this in Exercise 22.E.1. In Exercise 22.E.2 you are asked to construct examples violating some of these conditions.

Example 22.E.1: *Egalitarian Solution.* At the egalitarian solution $f_e(\cdot)$, the gains from cooperation are split equally among the agents. That is, for every bargaining problem $U \subset \mathbb{R}^I$, $f_e(U)$ is the vector in the frontier of U with all its coordinates equal. Figure 22.E.3 depicts the case $I = 2$. Note also that, as illustrated in the figure, every $f_e(U)$ maximizes the Rawlsian social welfare function Min $\{u_1, \ldots, u_I\}$ on U.

The egalitarian solution satisfies the IIA property (verify this). Clearly, for this olution, utility units are comparable across agents, and so the IUU property is not satisfied.[26] ■

Example 22.E.2: *Utilitarian Solution.* For every U we now let $f_u(U)$ be a maximizer of $\sum_i u_i$ on $U \cap \mathbb{R}^I_+$. If U is strictly convex, then this point is uniquely defined and, therefore, on the domain of strictly convex bargaining problems the IIA property is satisfied. As with the previous example, the solution violates the IUU condition. Figure 22.E.4 illustrates the utilitarian solution in the case $I = 2$. ■

26. Do not forget that the utility values are not absolute values but rather utility differences from the threat point. It is because of this that changes of origins do not matter.



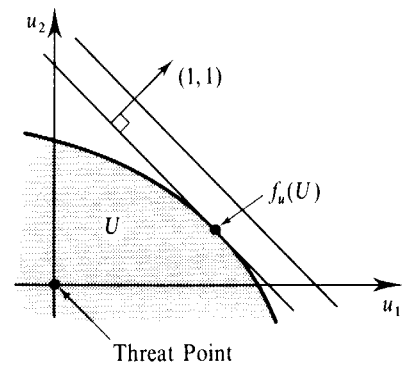


Figure 22.E.3 (left)
The egalitarian solution for bargaining problems.

Figure 22.E.4 (right)
The utilitarian solution for bargaining problems.

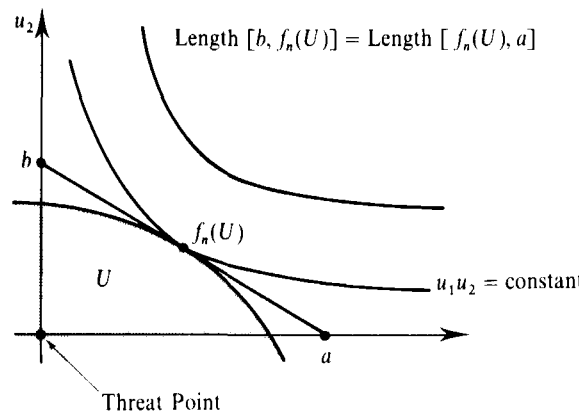


Figure 22.E.5
The Nash solution for bargaining problems.

Example 22.E.3: *Nash Solution.* For this solution, we take a position intermediate between the two previous examples by requiring that $f_n(U)$ be the point in $U \cap \mathbb{R}^I_+$ that maximizes the product of utilities $u_1 \times \cdots \times u_I$, or, equivalently, that maximizes $\sum_i \ln u_i$ (this corresponds to the case $\rho = 1$ in Example 22.C.4). In Figure 22.E.5, we provide an illustration for $I = 2$. In this case, the Nash solution has a simple geometry: $f_n(U)$ is the boundary point of U through which we can draw a tangent line with the property that its midpoint in the positive orthant is precisely the given boundary point $f_n(U)$; see Exercise 22.E.3.

As with the egalitarian and the utilitarian examples, the Nash solution satisfies the IIA property (because it is defined by the maximization of a strictly concave function). Interestingly, and in contrast to those solutions, *the condition of independence of utility units* (IUU) *holds for the Nash solution.* To see this, note that $\sum_i \ln u_i \geq \sum_i \ln u'_i$ is equivalent to $\sum_i \ln \beta_i u_i = \sum_i \ln u_i + \sum_i \ln \beta_i \geq \sum_i \ln u'_i + \sum_i \ln \beta_i = \sum_i \ln \beta_i u'_i$ for any constants $\beta_i > 0$. The Nash solution is therefore invariant to whatever origins or units we wish to fix. It depends only on the cardinal characteristics of the utility functions of the agents over the underlying set of alternatives.

There is a way to view the Nash solution as a synthesis of the egalitarian and the utilitarian solutions designed to accomplish the invariance to units: *Given a bargaining problem U, the Nash solution is the only utility outcome that, for some rescaling of units of utility, coincides simultaneously with the utilitarian and the egalitarian solutions.* More formally, suppose that

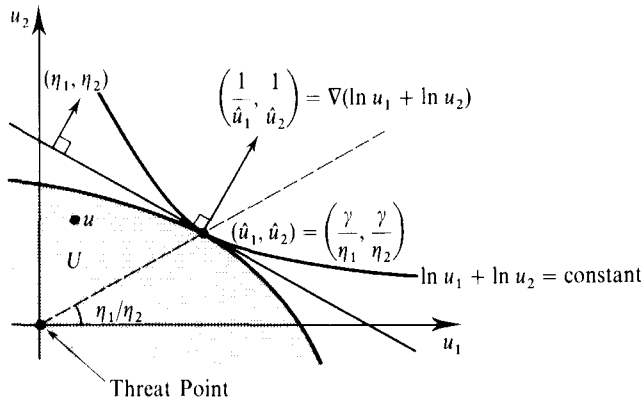


Figure 22.E.6
For some rescaling factors (η_1, η_2) the Nash solution is simultaneously egalitarian and utilitarian.

$\eta_i > 0$ are transformation rates of the given units into new units comparable across agents. If the utilitarian and the egalitarian solutions coincide when applied to the rescaled U' $(= \{(\eta_1 u_1, \ldots, \eta_I u_I)\colon (u_1, \ldots, u_I) \in U\})$ then the chosen point $\hat{u} \in U$ must be such that, first, it maximizes $\sum_i \eta_i u_i$ on U and, second, for some $\gamma > 0$ it satisfies $\eta_1 \hat{u}_1 = \cdots = \eta_I \hat{u}_I = \gamma$, that is, $\eta_i = \gamma(1/\hat{u}_i)$ for every i. Consider now any $u' \in U$. We have $\sum_i \eta_i u_i' \leq \sum_i \eta_i \hat{u}_i$ and therefore $\sum_i (1/\hat{u}_i) u_i' \leq \sum_i (1/\hat{u}_i)\hat{u}_i$. Since $(1/\hat{u}_1, \ldots, 1/\hat{u}_I)$ is the gradient of the concave function $\sum_i \ln u_i$ at $(\hat{u}_1, \ldots, \hat{u}_I)$, this implies $\sum_i \ln u_i' \leq \sum_i \ln \hat{u}_i$ (see Section M.C of the Mathematical Appendix). Hence $\hat{u}$ maximizes $\sum_i \ln u_i$ on U, that is, $\hat{u} = f_n(U)$.[27] See Figure 22.E.6 for an illustration of the argument. In Exercise 22.E.3 you should show the converse—that the Nash solution is simultaneously utilitarian and egalitarian for appropriate choice of units.

■

The Nash solution was proposed by Nash (1950), who also established the notable fact that it is the only solution that satisfies all the conditions so far.

Proposition 22.E.1: The Nash solution is the only bargaining solution that is independent of utility origins and units, Paretian, symmetric, and independent of irrelevant alternatives.[28]

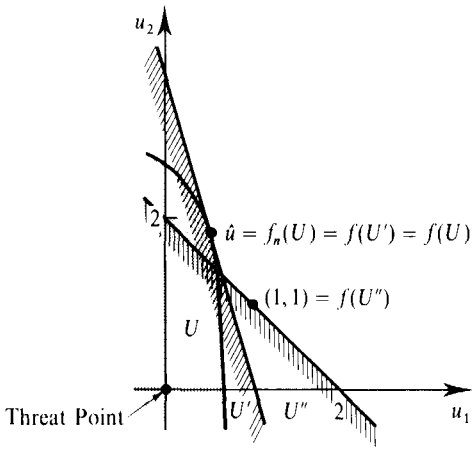
Proof: We have already shown in the discussion of Example 22.E.4 that the Nash solution satisfies the properties claimed.

To establish the converse, suppose we have a candidate solution $f(\cdot)$ satisfying all the properties. By the independence of utility origins, we can assume, as we have done so far, that $f(\cdot)$ is defined on sets where the threat point has been normalized to the origin. Given now an arbitrary U, let $\hat{u} = f_n(U)$ and consider the sets

$$U' = \{u \in \mathbb{R}^I\colon \sum_i u_i/\hat{u}_i \leq I\} \qquad \text{and} \qquad U'' = \{u \in \mathbb{R}^I\colon \sum_i u_i \leq I\}.$$

27. To repeat in more geometric terms: the hyperplane with normal $(\eta_1, \ldots, \eta_I)$ passing through $\hat{u}$ leaves U below it (because of the utilitarian property). Thus, it suffices to show that the set $\{u : \sum_i \ln u_i \leq \sum_i \ln \hat{u}_i\}$ lies above the hyperplane. But note that this follows from the fact that, because of the egalitarian property, $(\eta_1, \ldots, \eta_I)$ is proportional to $(1/\hat{u}_1, \ldots, 1/\hat{u}_I)$, which is the gradient of the concave function $\sum_i \ln u_i$ at $\hat{u}$.

28. Note that we do not assume individual rationality explicitly: tt turns out to be implied by the other conditions.



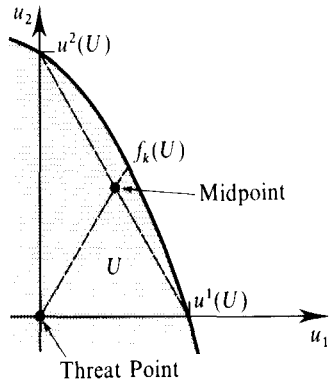


Figure 22.E.7 (left)
The Nash solution is determined uniquely from the independence of utility origins and units, symmetry, Pareto, and independence of irrelevant alternatives properties (Proposition 22.E.1).

Figure 22.E.8 (right)
The Kalai–Smorodinsky solution for bargaining problems.

See Figure 22.E.7 for an illustration in the case $I = 2$. Note that $U \subset U'$ because the concave function $\sum_i \ln u_i$ has gradient $(1/\hat{u}_1, \ldots, 1/\hat{u}_I)$ at $\hat{u}$, the point where it reaches its maximum value in the convex set U. The set U'' is symmetric and, therefore, by the symmetry and Paretian properties, we conclude that $f(U'') = (1, \ldots, 1)$. By the independence of utility units, it then follows that $f(U') = (\hat{u}_1, \ldots, \hat{u}_I) = \hat{u}$ [observe that $u \in U''$ if and only if $(\hat{u}_1 u_1, \ldots, \hat{u}_I u_I) \in U'$]. Finally, since $\hat{u} \in U$ and $U \subset U'$, the independence of irrelevant alternatives property yields $f(U) = \hat{u} = f_n(U)$, which is the result we wanted. ■

Example 22.E.4: *Kalai–Smorodinsky Solution.* This will be an example of a solution that does not satisfy the independence of irrelevant alternatives property. It was proposed by Kalai and Smorodinsky (1975). Given a bargaining problem $U \subset \mathbb{R}^I$, denote by $u^i(U) \in \mathbb{R}$ the maximum utility value that agent i could attain by means of some vector in $U \cap \mathbb{R}^I_+$. See Figure 22.E.8 for the case $I = 2$. To motivate the solution suppose that agent i has all the bargaining power (i.e., she can make a take-or-leave offer to the remaining agents). Then the outcome would give $u^i(U)$ to agent i and nothing to the remaining ones.[29] Hence, we could regard the numbers $u^i(U)$ as rough measures of the contributions of the respective agents to the cooperative endeavor and argue, perhaps, that if cooperation takes place then the solution should be the Pareto optimal allocation where the utilities of the different agents are proportional to $(u^1(U), \ldots, u^I(U))$; in other words, where utilities are proportional to the expected utilities that would obtain if we chose with equal probability the agent making a take-or-leave offer. This is precisely the Kalai–Smorodinsky solution $f_k(U)$. Its construction is indicated in Figure 22.E.8.

The Kalai Smorodinsky solution satisfies the Paretian and the symmetry properties. As with the Nash solution, *it does not involve interpersonal comparisons of utilities.* However, it is different from the Nash solution and, therefore (by Proposition 22.E.1), it cannot satisfy the IIA property. In Exercise 22.E.4 you are asked to verify all this. ■

29. We neglect cases where the utility vector that gives $u^i(U)$ to agent i and nothing to the remaining ones is Pareto dominated.

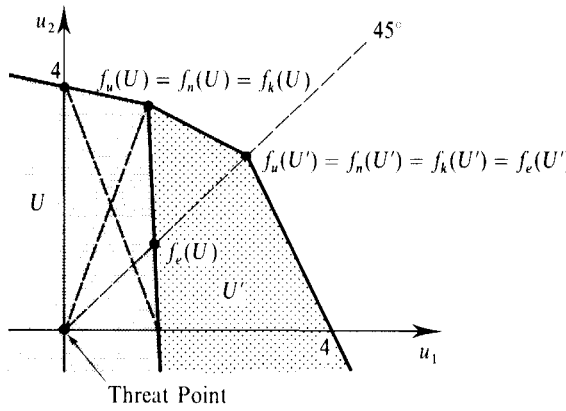


Figure 22.E.9
Lack of monotonicity of the utilitarian, Nash, and Kalai–Smorodinsky bargaining solutions.

The conditions summarized so far are by no means an exhaustive list of the properties that have been found to be of interest in the study of the bargaining problem. We conclude with a few informal comments about some others.

(i) *Linearity, or decomposability, properties.* Suppose that given two bargaining problems, $U \subset \mathbb{R}^I$ and $U' \subset \mathbb{R}^I$, we consider $\alpha U + (1 - \alpha)U' \subset \mathbb{R}^I$, for $\alpha \in [0, 1]$; we may think of this as, for example, a randomization between the two problems. Then we may want $f(\alpha U + (1 - \alpha)U') = \alpha f(U) + (1 - \alpha) f(U')$; that is, we may wish that all agents be indifferent between coming to a settlement before or after the resolution of uncertainty. This is a strong requirement, and none of the solutions we have studied satisfies it. In fact, it can be shown that, essentially (i.e., with a few other weak conditions), it is only satisfied by the modified version of the utilitarian solution that does not impose individual rationality. The same conclusion is reached if we consider $U + U'$ and ask that the overall settlement $f(U + U')$ equals the sequential settlement $f(U' + \{f(U)\}, f(U))$. Recall that by the IUO condition the last expression equals $f(U') + f(U)$.

(ii) *Monotonicity properties.* A bargaining solution $f(\cdot)$ is *monotone* if $f(U) \leq f(U')$ whenever $U \subset U'$, that is, if whenever the utility possibility set expands (keeping the threat point fixed at the origin), it is to everyone's advantage. The monotonicity requirement is stronger than may appear at first sight because the utility possibility set may expand in ways that are very asymmetric across agents. In fact, you can verify in Figure 22.E.9 that neither the utilitarian, nor the Nash, nor the Kalai–Smorodinsky solutions satisfy monotonicity. On the other hand, the egalitarian solution clearly does. In Exercise 22.E.5 you are invited to verify that the egalitarian solution is essentially the only symmetric and Paretian bargaining solution that satisfies the monotonicity condition. In Exercise 22.E.6 you can also check that the Kalai–Smorodinsky solution for $I = 2$ is characterized, with the IUO, IUU, P, and S properties, by a certain condition of partial monotonicity.

(iii) *Consistency properties.* This type of property concerns the mutual fit of the bargaining solutions when we apply them to problems with different numbers of agents. Let $f^I(\cdot)$ be a family of bargaining solutions (one for every set of agents I). Suppose, to be specific, that we start with $I = \{1, 2, 3\}$. Take any i, say $i = 1$, and imagine that, conditional on final cooperation, we give agent 1 the utility level $f_1^I(U)$, but that after making this commitment we reopen the negotiation between the two remaining agents. These two agents then have to find a settlement between themselves in the set $U' = \{(u_2, u_3)\colon (f_1^I(U), u_2, u_3) \in U\} \subset \mathbb{R}^2$. It is then natural to apply the solution in our family, that is, $f^{I\setminus\{1\}}(U')$. Our family is consistent if

$f^{I\setminus\{1\}}(U') = (f_2^I(U), f_3^I(U))$. In words: the renegotiation leads exactly to the outcome of the initial negotiation. The utilitarian and the Nash solution are consistent (in general, any solution obtained by maximizing a generalized utilitarian SWF is consistent; see Exercise 22.E.7). The Kalai–Smorodinski solution is not (verify this in Exercise 22.E.8). It is an interesting, and nontrivial, fact that the consistency axiom can be used instead of IIA in the characterization of the Nash solution [see Lensberg (1987) and Thomson and Lensberg (1992)].

22.F Coalitional Bargaining: The Shapley Value

The analysis of Section 22.E is limited in one important respect: it does not contemplate the possibility of situations intermediate between the full cooperation of all agents and the complete breakdown represented by the threat-point outcome. When there are more than two agents this is definitely restrictive. In this section, we make allowance for the possibility of *partial cooperation* and discuss how this may influence the eventual distribution of the gains from cooperation.

We are given a set I of agents and we proceed by specifying a possible set of utility outcomes to every potential subset of cooperators $S \subset I$. When $S \neq I$ these utility outcomes are interpreted as a description of what may occur if bargaining breaks down and the members of S end up cooperating among themselves. For the purpose of simplicity, we limit ourselves to the case where utility is comparable across agents (we fix individual units to have the same social utility value) and freely transferable among them. We can then represent the total amount of utility available to the members of $S \subset I$, if they cooperate, by a number $v(S)$ or, equivalently, by the utility possibility set $\{u \in \mathbb{R}^S: \sum_{i \in S} u_i \leq v(S)\}$. In cooperative game theory, which we reviewed in Appendix A of Chapter 18, the rule that assigns $v(S)$ to every $S \subset I$ is known as a *game in characteristic form*, a subset $S \subset I$ is usually referred to as a *coalition*, and the number $v(S)$ is known as the *worth* of the coalition S.

The situation in which there are no gains from partial cooperation is captured by a characteristic form where $v(S) = \sum_{i \in S} v(i)$ for every $S \neq I$. When we interpret the vector $(v(1), \ldots, v(I))$ as the threat point, this situation reduces to the bargaining problem studied in Section 22.E. In the current world of transferable utility, the egalitarian, Nash, and Kalai–Smorodinsky bargaining solutions[30] lead to the same proposal: The gains from cooperation should be split equally among the agents; that is, agent $i \in I$ should get

$$v(i) + \frac{1}{I}\left(v(I) - \sum_{h \in I} v(h)\right).$$

In fact, any bargaining solution that is independent of utility origins, Paretian, and symmetric makes this recommendation (Exercise 22.F.1).

The question we will try to answer in this section is the following: *Assume that, in an environment of games in characteristic form, all the members of I decide on cooperation, and therefore on distributing $v(I)$ among themselves. What is then the proper generalization of the equal split solution?* It stands to reason that the solution will have to reflect, in some manner, the numbers $v(S)$, $S \subset I$, since these incorporate

30. The utilitarian solution is not uniquely defined in the transferable utility case.

the information on how valuable the contribution of a particular agent is compared to that of another.

Definition 22.F.1: Given the set of agents I, a *cooperative solution* $f(\cdot)$ is a rule that assigns to every game $v(\cdot)$ in characteristic form a utility allocation $f(v) \in \mathbb{R}^I$ that is feasible for the entire group, that is, such that $\sum_i f_i(v) \leq v(I)$.

Reiterating the analytical strategy of Section 22.E, we continue by stating a number of desirable properties for a solution. The first three are merely variations of properties already encountered above.

Definition 22.F.2: The cooperative solution $f(\cdot)$ is *independent of utility origins and of common changes of utility units* if, whenever we have two characteristic forms $v(\cdot)$ and $v'(\cdot)$ such that $v(S) = \beta v'(S) + \sum_{i \in S} \alpha_i$ for every $S \subset I$ and some numbers $\alpha_1, \ldots, \alpha_I$, and $\beta > 0$, it follows that $f(v) = \beta f(v') + (\alpha_1, \ldots, \alpha_I)$.

From now on we assume the property of Definition 22.F.2. Because of it we can, in particular, normalize $v(\cdot)$ to $v(i) = 0$ for all i.

Definition 22.F.3: The cooperative solution $f(\cdot)$ is *Paretian* if $\sum_i f_i(v) = v(I)$, for every characteristic form $v(\cdot)$.

Definition 22.F.4: The cooperative solution $f(\cdot)$ is *symmetric* if the following property holds: Suppose that two characteristic forms, $v(\cdot)$ and $v'(\cdot)$ differ only by a permutation $\pi: I \to I$ of the names of the agents; that is, $v'(S) = v(\pi(S))$ for all $S \subset I$. Then the solution also differs only by this permutation; that is, $f_i(v') = f_{\pi(i)}(v)$ for all i.

The property defined next, in Definition 22.F.5, underlines the fact that we are trying to solve not the welfare-theoretic problem of distributing total utility equitably but rather the more limited problem of distributing equitably the *surplus* that can be attributed to the cooperation among agents, given the realities of the particular bargaining situation captured by the characteristic form.

Definition 22.F.5: The cooperative solution $f(\cdot)$ satisfies the *dummy axiom* if, for all games $v(\cdot)$ and all agents i such that $v(S \cup \{i\}) = v(S)$ for all $S \subset I$, we have $f_i(v) = v(i)$ $(= 0)$. In words: If agent i is a *dummy* (i.e., does not contribute anything to any coalition), then agent i does not receive any share of the surplus.

There are a number of cooperative solutions that satisfy the above properties. The most important is the *Shapley value* [proposed by Shapley (1953)]. We refer to Appendix A of Chapter 18 on cooperative game theory for examples, motivation, and extended discussion of this concept. Here we limit ourselves to offering a definition.

Suppose that we consider an arbitrary ordering of the agents or, formally, an arbitrary permutation π of the names $\{1, \ldots, I\}$. Then

$$g_{v,\pi}(i) = v(\{h: \pi(h) \leq \pi(i)\}) - v(\{h: \pi(h) < \pi(i)\})$$

represents how much agent i contributes when she joins the group of her predecessors in the ordering. This is the amount the predecessors would agree to pay i if she had

all the negotiating power, that is, if she could make a take-it-or-leave-it offer.[31] Note that $\sum_i g_{v,\pi}(i) = v(I)$ for all permutations π.

The agents do not come to us ordered. They all stand on the same footing. We may account for this by giving every agent the same chance of being in any position, thereby making all positions equally likely. Equivalently, we could take the (equal weighting) average of agent i contributions over all permutations π (there are $I!$ of these). This is precisely the Shapley value solution.

Definition 22.F.6: The Shapley value solution $f_s(\cdot)$ is defined by

$$f_{si}(v) = \frac{1}{I!}\sum_{\pi} g_{v,\pi}(i) \qquad \text{for every } i. \tag{22.F.1}$$

It is simple to verify that $f_s(\cdot)$ satisfies all the properties listed so far (see Exercise 22.F.2). For further discussion, see Exercises 22.F.3 and 22.F.4 (and, to repeat, Appendix A of Chapter 18). In Exercise 22.F.5 we describe another cooperative solution (the nucleolus).

The Cost-Allocation Problem

The following discussion is in the nature of an appendix as there is no immediate conceptual connection with the previous material. The link is that we again make use of the Shapley value.[32]

Suppose that we have a set I of *projects* and that a policy decision to carry them forward has already been taken. For some reason (e.g., accounting or financing rules), the total cost $C(I)$ must be allocated exactly to the different projects; that is, we must specify $(c_1, \ldots, c_I)$ such that $\sum_i c_i = C(I)$. What is a reasonable way to determine such *cost allocation rules*? This is the cost-allocation problem.

We must first of all emphasize that, from the point of view of first-best optimality, the cost allocation rules *should not* be used to guide investment, that is, to decide which projects to carry out. Loosely speaking, we have seen in Section 16.G that the correct efficiency prices for inputs (note that we can think of projects as inputs to the production function for welfare) do not need to cover total cost exactly (see Exercises 22.F.6 and 22.F.7). Because we wish to avoid the temptation to use cost allocation rules in this way we insist that the set of projects to be implemented be exogenously given.

An alternative would be to recognize that the cost-covering constraint confers to the welfare problem a second-best nature. That is, we could attempt to maximize social welfare subject to the condition that input (i.e., project) prices must be fixed at levels that exactly cover costs. This approach was taken by Boiteux (1956) in the context of the theory of the regulated firm, and the solution is closely related to Ramsey pricing (see Example 22.B.2 and Exercise 22.F.6).

Any welfare-theoretic approach, however, would need to use information on individual preferences. If this cannot, or should not, be done, we are still left with an unresolved the problem. A suggested approach proceeds then as follows: suppose that we have information on the cost of every subset of projects (this is far from an innocuous demand); that is, we know $C(T)$ for every $T \subset I$. Then, formally, $C(\cdot)$ is a cooperative game in characteristic form

31. In other words, an offer that, if rejected by some predecessor, would permanently preclude the possibility of agent i or any successor from joining the coalition of the predecessors. To verify the informal claim that the offer will be $g_{v,\pi}(i)$, determine how much the last agent in the ordering will get and proceed by backward induction.

32. See Young (1994) for an introductory account to cost-allocation and related problems.

and we could resort to the Shapley value as a way to split $C(I)$ among the different projects. An example may help to clarify the point.

Example 22.F.1: This is a favorite example for academics. A professor of economics based in the United States is planning a grand tour of Europe that will take her to three universities, one in Britain (B), one in Spain (S), and one in Germany (G). The total air fare comes to 1600 dollars. How is this to be reimbursed by the three institutions? Suppose that after some research it turns out that $C(B) = C(S) = C(G) = 800$, $C(BS) = C(BG) = 1000$, and $C(SG) = 1400$. The Shapley value calculation (carry it out!) then gives $c_B = 400$ and $c_S = c_G = 600$. This split does indeed seem to reflect well the comparative ease of managing a side trip to Britain if already going to some of the other destinations. ■

REFERENCES

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

Atkinson, A. (1973). How progressive should income-tax be? In *Economic Justice, Selected Readings*, edited by E. Phelps. London: Penguin Books.

Atkinson, A., and J. Stiglitz. (1980). *Lectures on Public Economics*. New York: McGraw-Hill.

Bergson, A. (1938). A reformulation of certain aspects of welfare economics. *Quarterly Journal of Economics* **52**: 310–34.

Boiteux, M. (1956). Sur la gestion des monopoles publiques astreints à l'équilibre budgétaire. *Econometrica* **24**: 22–40. [Translated in *Journal of Economic Theory* (1991) **3**: 219–40.]

Harsanyi, J. (1955). Cardinal welfare, individual ethics, and interpersonal comparability of utility. *Journal of Political Economy* 61: 309–21. [Also in Phelps (1973).]

Guesnerie, R. (1995). *A Contribution to the Pure Theory of Taxation*, Cambridge, U.K.: Cambridge University Press.

Kalai, E., and M. Smorodinsky. (1975). Other solutions to Nash's bargaining problem. *Econometrica* **43**: 513–18.

Laffont, J.-J. (1988). *Fundamentals of Public Economics*. Cambridge, Mass.: MIT Press.

Lipsey, R. C., and K. Lancaster. (1956). The general theory of the second best. *Review of Economic Studies* **24**: 11–32.

Lensberg, T. (1987). Stability and collective rationality. *Econometrica* **55**: 935–62.

Maskin, E. (1978). A theorem on utilitarianism. *Review of Economic Studies* **42**: 93–96.

Moulin, H. (1988). *Axioms of Cooperative Decision Making*. Cambridge, U.K.: Cambridge University Press.

Nash, J. F. (1950). The bargaining problem. *Econometrica* **28**: 155–62.

Phelps, E., ed. (1973). *Economic Justice, Selected Reading*. London: Penguin Books.

Ramsey, F. (1927). A contribution to the theory of taxation. *Economic Journal* **37**: 47–61.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Roberts, K. (1980). Possibility theorems with interpersonally comparable welfare levels. *Review of Economic Studies* **47**: 409–20.

Roth, A. (1979). *Axiomatic Models of Bargaining*. New York: Springer-Verlag.

Samuelson, P. (1947). *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.

Sen, A. (1977). On weights and measures: informational constraints in social welfare analysis. *Econometrica* **45**: 1539–72.

Shapley, L. (1953). A value for n-person games. In *Contributions to the Theory of Games II. Annals of Mathematics Studies*, 28, edited by H. Kuhn, and A. Tucker. Princeton, N.J.: Princeton University Press.

Starrett, D. A. (1988). *Foundations of Public Economics*. Cambridge, U.K.: Cambridge University Press.

Thomson, W., and T. Lensberg. (1992). *The Theory of Bargaining with a Variable Number of Agents*. Cambridge, U.K.: Cambridge University Press.

Thomson, W. (1995). Cooperative models of bargaining. In *Handbook of Game Theory*, Vol. II, edited by R. Aumann, and S. Hart. Amsterdam: North-Holland.

Young, H. P. (1994). *Equity. In Theory and Practice.* Princeton, N.J.: Princeton University Press.

EXERCISES

22.B.1^{A} Give sufficient conditions for the convexity of the first-best utility possibility set in the context of the exchange economies of Example 22.B.1.

22.B.2^{A} Derive the first-order conditions stated in Example 22.B.2.

22.B.3^{A} Derive the first-order conditions (22.B.2) of Example 22.B.3.

22.B.4^{B} Show as explicitly as you can that the utility possibility set of Example 22.B.4 may not be convex.

22.C.1^{A} Suppose that the utility possibility set $U \subset \mathbb{R}^I$ is symmetric and convex. Show that the social optimum of an increasing, symmetric, strictly concave social welfare function $W(\cdot)$ assigns the same utility values to every agent. [Note: A set U is symmetric if $u \in U$ implies $u' \in U$ for any u' obtained from u by a permutation of its entries.] Observe that the same conclusion obtains if $W(\cdot)$ is allowed to be just concave, as in the utilitarian case, but U is required to be strictly convex.

22.C.2^{A} Suppose that we contemplate a decision maker in an original position (or *ex-ante*, or *behind the veil of ignorance*) before the occurrence of a state of the world that will determine which of I possible identities the decision maker will have. There is a finite set X_i of possible final outcomes in identity i. Denote $X = X_1 \times \cdots \times X_I$.

(a) Appeal to the theory of state-dependent utility presented in Section 6.E to justify a utility function on X of the form

$$U(x_1, \ldots, x_I) = u_1(x_1) + \cdots + u_I(x_I).$$

Interpret and discuss the implications of this utility function for the usage of a purely utilitarian social welfare function.

(b) Suppose that $X_1 = \cdots = X_I$ and the preference relation on X defined by the utility function in (a) is symmetric. What does this imply for the form of the utility function? Discuss and interpret.

22.C.3^{B} We have N final social outcomes and we consider a set of alternatives X that is the set of lotteries over these outcomes. An alternative can be represented by the list of probabilities assigned to the different final outcomes, that is, $p = (p_1, \ldots, p_N)$ where $p_n \geq 0$ for every n and $p_1 + \cdots + p_N = 1$.

We assume that we are given a social preference relation $\succsim$ on X that is continuous and conforms to the independence axiom. Thus, it can be represented by a utility function of the expected utility form

$$U(p) = u_1 p_1 + \cdots + u_N p_N.$$

From now on we assume that this social utility function $U(\cdot)$ defined on X is given.

(a) Suppose that there are two final outcomes and that they are specified by which of two individuals will receive a certain indivisible object. Suppose also that social preferences are symmetric in the sense that there is social indifference between the lottery that gives the object to individual 1 for sure and the lottery that gives the object to individual 2 for sure. Show that all the lotteries must then be socially indifferent. Discuss and interpret this conclusion.

Is it plausible to you? If you wanted to escape from it, how would you do it? What does this all say about the independence axiom as applied to social decisions?

Suppose now that there are I agents and that in addition to the social utility function $U(\cdot)$ we are also given I individual preference relations $\succsim_i$ defined on the same set of lotteries X. We assume that they are also represented by utility functions of the expected utility form

$$U_i(p) = u_{1i}p_1 + \cdots + u_{Ni}p_N \qquad \text{for } i = 1, \ldots, I.$$

We say that the social utility function $U(\cdot)$ is *Paretian* if we have $U(p) > U(p')$ whenever $U_i(p) > U_i(p')$ for every i.

(b) Consider a case with $N = 3$ and $I = 2$ and illustrate, in the 2-dimensional simplex of lotteries, how the indifference map of the utility functions of the two agents and of the social utility function fit together when the social utility function is Paretian.

(c) Exhibit a case where the Paretian condition determines uniquely the social indifference map (recall that we are always assuming the independence axiom for social preferences!). Argue, however, that in general the Paretian condition does not determine uniquely the social indifference map. In fact, exhibit an example where any social utility function is Paretian.

(d) Argue (you can restrict yourself to $N = 3$ and $I = 2$) that if the social utility function $U(p)$ is Paretian then it can be written in the form

$$U(p) = \beta_1 U_1(p) + \cdots + \beta_I U_I(p)$$

where $\beta_i \geq 0$ for every i and $\beta_i \neq 0$ for some i. What does this conclusion say for the usage of a purely utilitarian social welfare function? Interpret the β_i weights, as well as the fact that they need not be equal across individuals.

22.C.4[A] The *leximin* ordering, or preference relation, on $\mathbb{R}^I$ has been mentioned in footnote 11 of this chapter when discussing the Rawlsian SWF. It is formally defined as follows. Given a vector $u = (u_1, \ldots, u_I)$ let $u^r \in \mathbb{R}^I$ be the vector that is the *nondecreasing rearrangement* of u. That is, the entries of u^r are in nondecreasing order and its numerical values (multiplicities included) are the same as for u. We then say that the vector u is at least as good as the vector $\hat{u}$ in the leximin order if u^r is at least as good as $\hat{u}^r$ in the lexicographic ordering introduced in Example 3.C.1.

(a) Interpet the definition of the leximin as a refinement of the Rawlsian preference relation.

(b) Show that the leximin ordering cannot be represented by a utility function. It is enough to show this for $I = 2$.

(c) (Harder) Show that the social optimum of a leximin ordering is a Pareto optimum. You can limit yourself to the case $I = 3$.

22.C.5[B] Consider the constant elasticity family of social welfare functions (Example 22.C.4). Argue that $W_\rho(u) \to \text{Min}\{u_1, \ldots, u_I\}$ as $\rho \to \infty$.

22.C.6[A] Suppose that U and U' are utility possibility sets and that we associate with them Pareto optimal utility outcomes $\bar{u} \in U$ and $\bar{u}' \in U'$, respectively. Show graphically that:

(a) It is possible for U' to pass the strong compensation test over U and yet for the outcome with U' to be worse than the outcome with U, as measured by the purely utilitarian SWF.

(b) If the utility possibility sets are derived from a quasilinear economy and U' passes the weak compensation test over U, then it also passes the strong compensation test and, moreover, the outcome for U' is a utilitarian improvement over the outcome for U. Is this conclusion valid if we evaluate social welfare by a nonutilitarian SWF?

22.C.7[B] Construct an explicit example of two Edgeworth box economies, differing only in their distributions of the initial endowments, such that the utility possibility set of each one

passes the weak compensation test over the utility possibility set of the other, when the utility outcome in the latter is chosen to correspond to one of its competitive equilibria.

22.C.8[A] Suppose we have two utility possibility sets U, U' with respective outcomes $u \in U$ and $u' \in U'$. We say that (U', u') passes the *Kaldor compensation test* over (U, u) if U' passes the weak compensation test over (U, u) and U *does not* pass the weak compensation test over (U', u').

(a) For $I = 2$, represent graphically a situation where Kaldor comparability is possible and one where it is not.

(b) Observe that Kaldor comparability is asymmetric. Define your terms.

(c) Show that Kaldor comparability may not be transitive.

22.D.1[B] In this exercise we verify the indispensability of the assumptions of Proposition 22.D.1.

(a) Suppose there are three agents and only two alternatives, that is, $X = \{x, y\}$. The social welfare functional is given by

$$x\ F(\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)\ y \qquad \text{if and only if} \qquad \tilde{u}_i(x) \geq \tilde{u}_i(y) \text{ for every } i$$

and

$$y\ F(\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)\ x \qquad \text{if and only if} \qquad \tilde{u}_i(y) \geq \tilde{u}_i(x) \text{ for at least one } i.$$

Check that the social preference relation is always complete, that the social welfare functional cannot be represented by means of a social welfare function, and that only the condition on the number of alternatives fails from Proposition 22.D.1.

(b) Now we have three agents and three alternatives, that is, $X = \{x, y, z\}$. The social welfare functional is given by

$$x\ F_p(\tilde{u}_1, \ldots, \tilde{u}_I)\ y\ F_p(\tilde{u}_1, \ldots, \tilde{u}_I)\ z$$

for every $(\tilde{u}_1, \ldots, \tilde{u}_I) \in \mathscr{U}^I$. Show that, again, no representation by means of a social welfare function is possible and that, of the assumptions of Proposition 22.D.1, only the Paretian property fails to be satisfied.

(c) Exhibit an example in which the only condition of Proposition 22.D.1 that fails to be satisfied is pairwise independence.

22.D.2[A] Carry out the verification requested in the second paragraph of the proof of Proposition 22.D.1.

22.D.3[A] In text.

22.D.4[A] A social welfare functional F is *lexically dictatorial* if there is a list of $n > 0$ agents $h_1, \ldots, h_n$ such that the strict preference of h_1 prevails socially, the strict preference of h_2 prevails among the alternatives for which h_1 is indifferent, and so on.

(a) Show that if F is lexically dictatorial then F is Paretian, is pairwise independent, and does not allow for interpersonal comparisons of utility.

(b) Under what conditions can a social welfare functional that is lexically dictatorial be generated from a social welfare function?

(c) Show that if a dictatorial social welfare functional is generated from a social welfare function $W(u) = \sum_i b_i u_i$, then $b_i = 0$ for every i distinct from the dictator.

22.D.5[C] Complete the proof of Arrow's impossibility theorem along the lines suggested in the last paragraph prior to the small-type text at the end of Section 22.D. (Assume that Proposition

22.D.3 is valid under the weakened assumption that F is generated from a social preference relation on $\mathbb{R}^I$.)

22.D.6[B] This exercise is concerned with social welfare functions satisfying expression (22.D.1).

(a) Show that the nonsymmetric utilitarian function $W(u) = \sum_i b_i u_i$ can be written in the form (22.D.1).

(b) Show that if $W(\cdot)$ is symmetric and $g(0) = 0$ then $g(\cdot) \geq 0$.

(c) Show that the symmetric Rawlsian social welfare function $W(u) = \text{Min}\{u_1, \ldots, u_I\}$ can be written in the form (22.D.1). What about nonsymmetric Rawlsian social welfare functions? [*Hint*: Check the condition of invariance to common changes of origins.]

(d) Give other examples satisfying (22.D.1), in particular, examples with $g(\cdot) \geq 0$ and intermediate between the utilitarian and the Rawlsian cases. Interpret them.

(e) Argue that if in (22.D.1) the function $g(\cdot)$ is homogeneous of degree one and differentiable, then it must be linear (and so we are back to the utilitarian case).

22.D.7[B] Consider the constant elasticity family of social welfare functions studied in Example 22.C.4.

(a) Show that the social welfare functionals derived from SWFs in this family are invariant to common changes of units.

(b) Show that the only members of this family which are also invariant to common changes of origins, and therefore admit a representation in the form (22.D.1), are the purely utilitarian (i.e., $\rho = 0$) and the Rawlsian (i.e., $\rho = \infty$).

22.D.8[B] This is an exercise on the property of invariance to common ordinal transformation.

(a) Show that the symmetric, Rawlsian social welfare function satisfies the property.

(b) Show that the anti-Rawlsian function $W(u) = \text{Max}\{u_1, \ldots, u_I\}$ also satisfies it.

(c) Show that the property is satisfied for dictatorial social welfare functionals.

(d) (Harder) Suppose that $I = 2$ and $W(u) = W(u')$ for two vectors $u, u' \in \mathbb{R}^2$, with $u'_1 < u_1 < u_2 < u'_2$. Assume also that $W(\cdot)$ is increasing. Show that the induced social welfare functional cannot be invariant to identical ordinal transformations. From this, argue informally (you can do it graphically) that for $I = 2$ a continuous, increasing social welfare function that is also invariant to identical ordinal transformations must be either dictatorial, Rawlsian, or anti-Rawlsian.

22.E.1[A] Verify that the bargaining solutions in Examples 22.E.1 to 22.E.4 are independent of utility origins, Paretian, symmetric, and individually rational. It is enough if you do so for $I = 2$.

22.E.2[A] State nonsymmetric versions of the four bargaining solutions studied in Section 22.E (egalitarian, utilitarian, Nash, and Kalai–Smorodinsky). Motivate them.

22.E.3[B] This is an exercise on the Nash solution.

(a) Verify that for $I = 2$, $f_n(U)$ is the boundary point of U through which we can draw a tangent line with the property that its midpoint in the positive orthant is precisely the given boundary point $f_n(U)$.

(b) Verify that if $U \subset \mathbb{R}^I$ is a bargaining problem then there are rescaling units for the individual utilities with the property that the Nash solution becomes simultaneously egalitarian and utilitarian.

22.E.4[A] Verify that the Kalai–Smorodinsky solution satisfies the property of independence of utility units but violates the property of independence of irrelevant alternatives. You can restrict yourself to $I = 2$.

22.E.5[B] This is an exercise on the monotonicity property.

(a) Show that the egalitarian solution is the only bargaining solution that is independent of utility origins, Paretian, symmetric and monotonic. [*Hint*: Consider first a family of symmetric utility possibility sets with linear boundaries. Notice then that for any two sets U, U' we always have $U \cap U' \subset U$ and $U \cap U' \subset U'$.]

(b) (Harder) Suppose that $f(\cdot)$ is a bargaining solution that is independent of utility origins, Paretian, and strongly monotonic [if $U \subset U'$ then $f(U) \leq f(U')$ and, in addition, if $f(U)$ is interior to U' then $f(U) \ll f(U')$]. Show that there is a curve in $\mathbb{R}^I$ starting at the origin and strictly increasing such that, for every U, $f(U)$ is the intersection point of the boundary of U with this curve. You can restrict yourself to the case $I = 2$.

22.E.6[C] Let $I = 2$. A bargaining solution $f(\cdot)$ is *partially monotone* if when $U \subset U'$ and $u^i(U) = u^i(U')$, that is, U' expands U only in the direction of agent $j \neq i$, we have $f_j(U') \geq f_j(U)$ for $j \neq i$. Argue that the Kalai–Smorodinsky solution is characterized by the following properties: independence of utility origins and units, Pareto, symmetry, and partial monotonicity. [*Hint*: use sets U such that $U' \subset U$ and $u^1(U) = u^1(U')$, $u^2(U) = u^2(U')$].

22.E.7[A] Consider a family of bargaining solutions $f^I(\cdot)$ such that, for every set of agents I, $f^I(\cdot)$ is independent of utility origins and is generated by maximizing the social welfare function $\sum_i g(u_i)$ on normalized bargaining problems $U \subset \mathbb{R}^I$, where $g(\cdot)$ is increasing, strictly concave, and independent of the particular I considered. Show that the family f^I is consistent.

22.E.8[C] Show by example that the Kalai–Smorodinsky solution is not consistent. It is enough to consider three agents and its subgroups of two agents.

22.E.9[A] This exercise is aimed at showing the independence of the assumptions of Proposition 22.E.1. To this effect, give five examples such that for each of the five assumptions of Proposition 22.E.1 there is one of the examples that violates this assumption but satisfies the remaining four.

22.E.10[A] Give an example of a utilitarian bargaining solution (Example 22.E.2) that violates the property of independence of irrelevant alternatives. [*Hint*: It suffices to consider $I = 2$. Also, the violation should involve sets U that are convex but not strictly convex.]

22.E.11[C] Go back to the infinite horizon Rubinstein's bargaining model discussed in the Appendix A to Chapter 9 (specifically, Example 9.AA.2). The only modification is that the two agents are risk averse on the amount of money they get. That is, each has an increasing, concave, differentiable utility function $u_i(m_i)$ on the nonnegative amounts of money that they receive. The factor of discount $\delta < 1$ is the same for the two agents. Also $u_i(0) = 0$. The total amount of money is m.

(a) Write down the equations for a subgame perfect Nash equilibrium (SPNE) *in stationary strategies*. Argue that there is a single configuration of utility payoffs that can be obtained as payoffs of a SPNE in stationary strategies.

(b) Consider the utility possibility set

$$U = \{(u_1(m_1), u_2(m_2)) \in \mathbb{R}^2 : m_1 + m_2 = m\} - \mathbb{R}^2_+ .$$

Show that if δ is close to 1 then the payoffs of a SPNE in stationary strategies are nearly equal to the Nash bargaining solution payoffs.

(c) (Harder) Argue that every payoff configuration of a SPNE can be obtained as the payoff configuration of a SPNE in stationary strategies. Thus, the uniqueness result presented in Example 9.AA.2 extends to the case in which the agents have strictly concave, possibly different, utility functions for money.

22.F.1[A] Show that in the transferable utility case any bargaining solution that is invariant to independent changes of origin, symmetric, and Paretian divides the gains from cooperation equally among the agents.

22.F.2[A] Show that the Shapley value cooperative solution presented in Section 22.F satisfies the following properties: invariance to independent changes of utility origins, invariance to common changes of utility units, Paretian, symmetry, and the dummy axiom.

22.F.3[A] Suppose that for a given set of agents I we take two characteristic forms v and v' and consider their sum $v + v'$; that is, $v + v'$ is the characteristic form where $(v + v')(S) = v(S) + v'(S)$ for every $S \subset I$.

(a) Verify that the Shapley value is *linear* in the characteristic form; that is, $f_{Si}(v + v') = f_{Si}(v) + f_{Si}(v')$ for all v, v' and i.

(b) Interpret the linearity property as a postulate that agents are indifferent to the timing of resolution of uncertainty when we randomize among bargaining situations.

22.F.4[C] The linearity property of the previous exercise can be restated in a perhaps more intuitive form. We say that a characteristic form $v(\cdot)$ is a *unanimity game* if for some $T \subset I$ we have that $v(S) = v(T)$ if $T \subset S$, and $v(S) = 0$ otherwise (thus, the bargaining situations of Section 22.E correspond to $T = I$).

(a) Show that the independence of utility origins and invariance to common changes of utility units, Pareto, symmetry, and dummy axiom properties imply that, for a unanimity game $v(\cdot)$, any cooperative solution $f(\cdot)$ assigns the values $f_i(v) = (1/T)v(I)$ if $i \in T$, and $f_i(v) = 0$ otherwise.

(b) We say that the cooperative solution $f(\cdot)$ is *weakly linear* if for any v and v' differing only by a unanimity game [i.e., there is $T \subset I$ and $\alpha \in \mathbb{R}$ such that $v'(S) = v(S) + \alpha$ if $T \subset S$, and $v'(S) = v(S)$ otherwise] we have that $f_i(v') = f_i(v) + \alpha/T$ if $i \in T$, and $f_i(v') = f_i(v)$ otherwise. Show that if, in addition to the properties listed in (a), the cooperative solution $f(\cdot)$ is weakly linear, then it is fully linear, that is, $f(v + v') = f(v) + f(v')$ for any two characteristic forms v and v'.

(c) Show that the Shapley value is the only cooperative solution that satisfies the following properties: independence of utility origins and invariance to common changes of utility units, Paretian, symmetry, dummy axiom, and linearity.

22.F.5[C] In this exercise we describe another cooperative solution for a game in characteristic form: the *nucleolus*. For simplicity we do it for the particular case in which $I = 3$, $v(1) = v(2) = v(3) = 0$, and $0 \leq v(S) \leq v(I)$, for any group S of two agents.

Given a utility vector $u = (u_1, u_2, u_3) \geq 0$ and an $S \subset I$ the *excess* of S at u is $e(u, S) = v(S) - \sum_{i \in S} u_i$. We define the *first maximum excess* as $m_1(u) = \text{Max}\,\{e(u, S)\colon 1 < \#S < 3\}$. Choose a two-agent coalition S such that $m_1(u) = e(u, S)$. Then we define the *second maximum excess* as $m_2(u) = \text{Max}\,\{e(u, S')\colon 1 < \#S' < 3 \text{ and } S' \neq S\}$.

We say that an exactly feasible [i.e., $\sum_{i \in I} u_i = v(I)$] utility profile $u = (u_1, u_2, u_3) \geq 0$ is in the nucleolus if for any other such profile u' we have either $m_1(u) < m_1(u')$ or $m_1(u) = m_1(u')$ and $m_2(u) \leq m_2(u')$.

(a) Show that if $u = (u_1, u_2, u_3)$ is in the nucleolus then either the three excesses for two-agent coalitions are identical or two are identical and the third is larger.

(b) Show that there is one and only one utility profile in the nucleolus. [*Hint*: Argue first that there is a two-agent coalition S such that $e(u, S) = m_1(u)$ for every profile in the nucleolus.] From now on we refer to this profile as the *nucleolus solution*.

(c) Argue that the nucleolus solution is symmetric.

(d) Suppose that agent 1 is a dummy. Then $u_1 = 0$ at the nucleolus solution.

(e) Suppose that $\frac{1}{2}v(I) \leq v(S)$ for any coalition S of two agents. Show then that at the nucleolus profile the three excesses for two-agent coalitions are identical.

(f) Compute and compare the Shapley value and the nucleolus for the characteristic form: $v(1) = v(2) = v(3) = 0$, $v(\{1, 2\}) = v\{(1, 3\}) = 4$, $v(\{2, 3\}) = 5$, $v(I) = 6$.

(g) Show that if the core is nonempty (see Appendix A to Chapter 18 for the definition of the core in this context) then the nucleolus utility profile belongs to the core.

22.F.6[B] Consider a regulated firm that produces an output by means of a cost function $c(q)$. Assuming a quasilinear economy, the consumer surplus generated by q is $S(q)$.

(a) Suppose that $c(q)$ is strictly concave (i.e., strictly increasing returns to scale). Show that at the first-best price the firm will not cover costs. Conversely, for any q suppose that the price $p(q)$ is determined so that the cost is covered; that is, $p(q) = c(q)/q$. Show that if q is then determined so as to have $p(q) = S'(q)$, we will not reach the first-best optimum. Illustrate graphically.

(b) Suppose that the quantity produced, q, has to be determined under the constraint that with $p = S'(q)$ we have $pq \geq c(q)$. Solve this second-best welfare problem. Illustrate graphically.

(c) Interpret the units of output as "projects." For any production decision q, what is the cost allocation suggested by the Shapley value?

22.F.7[C] This exercise is similar to Exercise 22.F.6, except that the firm now produces two outputs under the separable cost functions $c_1(q_1), c_2(q_2)$. The surplus $S_1(q_1) + S_2(q_2)$ is also separable.

(a) The second-best problem [first studied by Boiteux (1956)] is now richer than in Exercise 22.F.6. Suppose that the quantities q_1, q_2 have to be determined so that with $p_1 = S'(q_1)$ and $p_2 = S'(q_2)$ we have $p_1q_1 + p_2q_2 \geq c_1(q_1) + c_2(q_2)$ (equivalently, at the chosen prices demand must be served and cost covered). Derive first-order conditions for this problem. Make them as similar as possible to the Ramsey formula of Example 22.B.2.

(b) (Harder) Interpret the units of outputs as projects. Suppose that these units are very small, so that a given production decision $(q_1, q_2) \gg 0$ represents the implementation of many projects of each of the two types. Can you guess, given (q_1, q_2), what is an approximate value for the cost allocation suggested by the Shapley value? [*Hint*: For most orderings of projects, any particular project will have preceding it an almost perfect sample of all the projects.]

(c) Suppose that for the productions $(\bar{q}_1, \bar{q}_2)$, the Shapley value cost allocation assigns cost per unit of c_1 and c_2 (note that "projects" of the same type receive the same cost imputation). Suppose also that $c_1 = \partial S_1(\bar{q}_1)/\partial q_1$ and $c_2 = \partial S_2(\bar{q}_2)/\partial q_2$. Interpret. Argue that, in general, these productions will not correspond to either the first-best or the second-best optima of the problem.

CHAPTER

23

Incentives and Mechanism Design

23.A Introduction

In Chapter 21, we studied how individual preferences might be aggregated into social preferences and ultimately into a collective decision. However, an important feature of many settings in which collective decisions must be made is that individuals' actual preferences are not publicly observable. As a result, in one way or another, individuals must be relied upon to reveal this information.

In this chapter, we study how this information can be elicited, and the extent to which the information revelation problem constrains the ways in which social decisions can respond to individual preferences. This topic is known as the *mechanism design problem.*

Mechanism design has many important applications throughout economics. The design of voting procedures, the writing of contracts among parties who will come to have private information, and the construction of procedures for deciding upon public projects or environmental standards are all examples.[1]

The chapter is organized as follows. In Section 23.B, we introduce the mechanism design problem. We begin by illustrating the difficulties introduced by the need to elicit agents' preferences. We also define and discuss the concepts of *social choice functions* (already introduced in Section 21.E), *ex post efficiency*, *mechanisms*, *implementation*, *direct revelation mechanisms*, and *truthful implementation.*

In Section 23.C, we identify the circumstances under which a social choice function can be implemented in *dominant strategy equilibria* when agents' preferences are private information. Our analysis begins with a formal statement and proof of the *revelation principle*, a result that tells us that we can restrict attention to direct revelation mechanisms that induce agents to truthfully reveal their preferences. Using this fact, we then study the constraints that the information revelation

1. Simple examples of the last two applications were encountered in Sections 14.C and 11.E, respectively.

problem puts on the set of implementable social choice functions. We first present the important *Gibbard–Satterthwaite theorem*, which provides a very negative conclusion for cases in which individual preferences can take unrestricted forms. In the rest of the section, we go on to study the special case of *quasilinear environments*, discussing in detail *Groves–Clarke mechanisms*.

In Section 23.D, we study implementation in *Bayesian Nash equilibria*. We begin by discussing the *expected externality mechanism* as an example of how the weaker Bayesian implementation concept can allow us to implement a wider range of social choice functions than is possible with dominant strategy implementation. We go on to provide a characterization of Bayesian implementable social choice functions for the case in which agents have quasilinear preferences that are linear in their type. As an application of this result, we prove the remarkable *revenue equivalence theorem* for auctions.

In Section 23.E, we consider the possibility that participation in a mechanism may be voluntary and study how the need to satisfy the resulting *participation constraints* limits the set of implementable social choice functions. Here we prove the important *Myerson–Satterthwaite theorem*, which shows that, under very general conditions, it is impossible to achieve ex post efficiency in bilateral trade settings when agents have private information and trade is voluntary.

In Section 23.F, we discuss the welfare comparison of mechanisms, defining the notions of *ex ante* and *interim incentive efficiency*, and providing several illustrations of the computation of welfare optimal Bayesian mechanisms.

Appendices A and B are devoted to, first, a discussion of the issue of multiple equilibria in mechanism design and, second, the issue of mechanism design when agents know each others' types but the mechanism designer does not (so-called *complete information environments*).

References for further reading are provided at the start of the various sections. We would be remiss, however, not to mention here two early seminal articles: Mirrlees (1971) and Hurwicz (1972).

23.B The Mechanism Design Problem

In this section, we provide an introduction to the *mechanism design problem* that we study in detail in the rest of the chapter.

To begin, consider a setting with I agents, indexed by $i = 1, \ldots, I$. These agents must make a collective choice from some set X of possible alternatives. Prior to the choice, however, each agent i privately observes his preferences over the alternatives in X. Formally, we model this by supposing that agent i privately observes a parameter, or signal, θ_i that determines his preferences. We will often refer to θ_i as agent i's *type*. The set of possible types for agent i is denoted Θ_i. Each agent i is assumed to be an expected utility maximizer, whose Bernoulli utility function when he is of type θ_i is $u_i(x, \theta_i)$. The ordinal preference relation over pairs of alternatives in X that is associated with utility function $u_i(x, \theta_i)$ is denoted $\succsim_i(\theta_i)$. Agent i's set of possible preference relations over X is therefore given by

$$\mathcal{R}_i = \{\succsim_i : \succsim_i = \succsim_i(\theta_i) \text{ for some } \theta_i \in \Theta_i\}.$$

Note that because θ_i is observed only by agent i, in the language of Section 8.E

we are in a setting characterized by *incomplete information*. As in Section 8.E, we suppose that agents' types are drawn from a commonly known prior distribution. In particular, denoting a profile of the agents' types by $\theta = (\theta_1, \ldots, \theta_I)$, the probability density over the possible realizations of $\theta \in \Theta_1 \times \cdots \times \Theta_I$ is $\phi(\cdot)$. The probability density $\phi(\cdot)$ as well as the sets $\Theta_1, \ldots, \Theta_I$ and the utility functions $u_i(\cdot, \theta_i)$ are assumed to be common knowledge among the agents, but the specific value of each agent i's type is observed only by i.[2]

Because the agents' preferences depend on the realizations of $\theta = (\theta_1, \ldots, \theta_I)$, the agents may want the collective decision to depend on θ. To capture this dependence formally, we introduce in Definition 23.B.1 the notion of a *social choice function*, a concept already discussed in Section 21.E.[3]

Definition 23.B.1: A *social choice function* is a function $f: \Theta_1 \times \cdots \times \Theta_I \to X$ that, for each possible profile of the agents' types $(\theta_1, \ldots, \theta_I)$, assigns a collective choice $f(\theta_1, \ldots, \theta_I) \in X$.[4]

One desirable feature for a social choice function to satisfy is the property of *ex post efficiency* described in Definition 23.B.2.

Definition 23.B.2: The social choice function $f: \Theta_1 \times \cdots \times \Theta_I \to X$ is *ex post efficient* (or *Paretian*) if for no profile $\theta = (\theta_1, \ldots, \theta_I)$ is there an $x \in X$ such that $u_i(x, \theta_i) \geq u_i(f(\theta), \theta_i)$ for every i, and $u_i(x, \theta_i) > u_i(f(\theta), \theta_i)$ for some i.

Definition 23.B.2 says that a social welfare function is ex post efficient if it selects, for every profile $\theta = (\theta_1, \ldots, \theta_I)$, an alternative $f(\theta) \in X$ that is Pareto optimal given the agents' utility functions $u_1(\cdot, \theta_1), \ldots, u_I(\cdot, \theta_I)$.

The problem faced by the agents is that the θ_i's are not publicly observable, and so for the social choice $f(\theta_1, \ldots, \theta_I)$ to be chosen when the agents' types are $(\theta_1, \ldots, \theta_I)$, each agent i must be relied upon to disclose his type θ_i. However, for a given social choice function $f(\cdot)$, an agent may not find it to be in his best interest to reveal this information truthfully. We illustrate this information revelation problem in Examples 23.B.1 through 23.B.4, which range from very abstract to more applied settings.

2. The formulation here is restrictive in one sense: in some settings of interest, agents' preferences over outcomes depend not only on their own observed signals but also on signals observed by others (e.g., agent i's preferences over whether to hold a picnic indoors may depend on agent j's knowledge of likely weather conditions). Through most of this chapter, we focus on the case in which an agent's preferences depend only on his own signal, known as the *private values* case. We generalize our analysis in Section 23.F.

3. In Section 21.E an agent's type was equivalent to his ordinal preferences over X, and so a social choice function was defined there simply as a mapping from $\mathscr{R}_1 \times \cdots \times \mathscr{R}_I$ to X. Moreover, it was assumed there that for all i we have $\mathscr{R}_i = \mathscr{R}$, the set of all possible ordinal preference orderings on X.

4. Two points should be noted about this definition. First, it restricts attention to *deterministic* social choice functions. This is largely for expositional purposes; although much of the chapter considers deterministic social choice functions, in Sections 23.D to 23.F we allow social choice functions that assign *lotteries* over X. Second, as in Section 21.E, we limit our attention to single-valued choice functions.

Example 23.B.1: *An Abstract Social Choice Setting.* In the most abstract case, we are given a set X and, for each agent i, a set $\mathcal{R}_i$ of possible rational preference orderings on X. To consider a very simple example, suppose that $X = \{x, y, z\}$ and that $I = 2$. Suppose also that agent 1 has one possible type, so that $\Theta_1 = \{\bar{\theta}_1\}$, and that agent 2 has two possible types, so that $\Theta_2 = \{\theta_2', \theta_2''\}$. The agents' possible preference orderings $\mathcal{R}_1 = \{\succsim_1(\bar{\theta}_1)\}$ and $\mathcal{R}_2 = \{\succsim_2(\theta_2'), \succsim_2(\theta_2'')\}$ are given by

$\succsim_1(\bar{\theta}_1)$	$\succsim_2(\theta_2')$	$\succsim_2(\theta_2'')$
x	z	y
y	y	x
z	x	z

[A higher positioned alternative is strictly preferred to a lower positioned one; so, for example, $x \succ_1(\bar{\theta}_1)\ y \succ_1(\bar{\theta}_1)\ z$.]

Now suppose that the agents wish to implement the ex post efficient social choice function $f(\cdot)$ with

$$f(\bar{\theta}_1, \theta_2') = y \qquad \text{and} \qquad f(\bar{\theta}_1, \theta_2'') = x.$$

If so, then agent 2 must be relied upon to truthfully reveal his preferences. But it is apparent that he will not find it in his interest to do so: When $\theta_2 = \theta_2''$, agent 2 will wish to lie and claim that his type is θ_2'.

In abstract social choice settings, a case of central interest arises when $\mathcal{R}_i$ is, for each agent i, equal to $\mathcal{R}$, the set of all possible rational preference relations on X. In this case, an agent has many possible false claims that he can make and, intuitively, it may be very difficult for a social choice function always to induce the agents to reveal their preferences truthfully. We will see a formal illustration of this point in Section 23.C when we present the Gibbard–Satterthwaite theorem. ■

Example 23.B.2: *A Pure Exchange Economy.* Consider a pure exchange economy with L goods and I consumers in which agent i has consumption set $\mathbb{R}^L_+$ and endowment vector $\omega_i = (\omega_{1i}, \ldots, \omega_{Li}) \gg 0$ (see Chapter 15). The set of alternatives is

$$X = \{(x_1, \ldots, x_I)\colon x_i \in \mathbb{R}^L_+ \text{ and } \sum_i x_{\ell i} \leq \sum_i \omega_{\ell i} \text{ for } \ell = 1, \ldots L\}.$$

In this setting it may be natural to suppose that $\mathcal{R}_i$, each consumer i's set of possible preference relations over alternatives in X, is a subset of $\mathcal{R}_E$, the set of individualistic (i.e., depending on x_i only), monotone, and convex preference relations on X.

To consider a simple example, suppose that $I = 2$, that consumer 1 has only one possible type, so that $\Theta_1 = \{\bar{\theta}_1\}$ and $\mathcal{R}_1 = \{\succsim_1(\bar{\theta}_1)\}$, and that for consumer 2 we have $\mathcal{R}_2 = \mathcal{R}_E$. Imagine then that we try to implement a social choice function that, for each pair $(\succsim_1(\theta_1), \succsim_2(\theta_2))$, chooses a Walrasian equilibrium allocation (note that this social choice function is ex post efficient). As Figure 23.B.1 illustrates, consumer 2 will not generally find it optimal to reveal his preferences truthfully. In the figure, $f(\bar{\theta}_1, \theta_2')$ is the unique Walrasian equilibrium when preferences are $(\succsim_1(\bar{\theta}_1), \succsim_2(\theta_2'))$ [it is the unique intersection of the consumers' offer curves OC_1 and OC_2' occurring at a point other than the endowment point]. However, by claiming that he has type θ_2'', which has as its offer curve OC_2'', consumer 2 can obtain the allocation $f(\bar{\theta}_1, \theta_2'')$ [the unique Walrasian equilibrium allocation when preferences

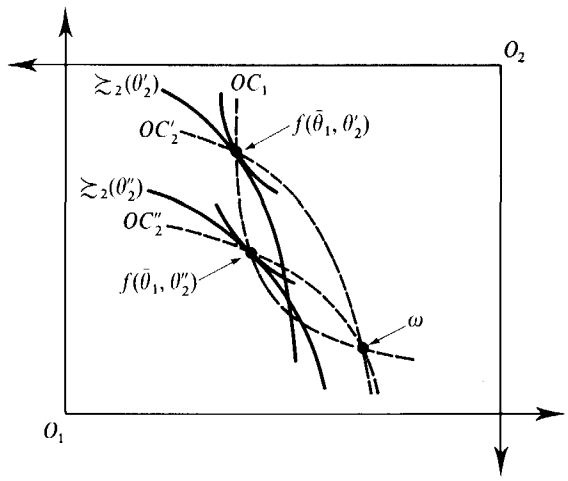


Figure 23.B.1
In the social choice function that selects a Walrasian equilibrium for each preference profile, agent 2 has an incentive to claim to be type θ_2'' when he is really type θ_2'.

are $(\succsim_1(\bar{\theta}_1), \succsim_2(\theta_2''))$], an allocation that he strictly prefers to $f(\bar{\theta}_1, \theta_2')$ when his preferences are $\succsim_2(\theta_2')$. ■

Example 23.B.3: *A Public Project.* Consider a situation in which I agents must decide whether to undertake a public project, such as building a bridge, whose cost must be funded by the agents themselves. An outcome is a vector $x = (k, t_1, \ldots, t_I)$, where $k \in \{0, 1\}$ is the decision whether to build the bridge ($k = 1$ if the bridge is built, and $k = 0$ if not), and $t_i \in \mathbb{R}$ is a monetary transfer to (or from, if $t_i < 0$) agent i. The cost of the project is $c \geq 0$ and so the set of feasible alternatives for the I agents is

$$X = \{(k, t_1, \ldots, t_I): k \in \{0, 1\}, t_i \in \mathbb{R} \text{ for all } i, \text{ and } \sum_i t_i \leq -ck\}.$$

The constraint $\sum_i t_i \leq -ck$ reflects the fact that there is no source of outside funding for the agents (so that we must have $c + \sum_i t_i \leq 0$ if $k = 1$, and $\sum_i t_i \leq 0$ if $k = 0$). We assume that type θ_i's Bernoulli utility function has the quasilinear form

$$u_i(x, \theta_i) = \theta_i k + (\bar{m}_i + t_i),$$

where $\bar{m}_i$ is agent i's initial endowment of the numeraire ("money") and $\theta_i \in \mathbb{R}$. We can then interpret θ_i as agent i's willingness to pay for the bridge.

In this context, the social choice function $f(\theta) = (k(\theta), t_1(\theta), \ldots, t_I(\theta))$ is ex post efficient if, for all θ,

$$k(\theta) = \begin{cases} 1 & \text{if } \sum_i \theta_i \geq c, \\ 0 & \text{otherwise,} \end{cases} \tag{23.B.1}$$

and

$$\sum_i t_i(\theta) = -ck(\theta). \tag{23.B.2}$$

Suppose that the agents wish to implement a social choice function that satisfies (23.B.1) and (23.B.2) and in which an egalitarian contribution rule is followed, that is, in which $t_i(\theta) = -(c/I)k(\theta)$. To consider a simple example, suppose that $\Theta_i = \{\bar{\theta}_i\}$ for $i \neq 1$ (so that all agents other than agent 1 have preferences that are known) and

$\Theta_1 = [0, \infty)$. Suppose also that $c > \sum_{i \neq 1} \bar{\theta}_i > c(I-1)/I$. These inequalities imply, first, that with this social choice function agent 1's type is critical for whether the bridge is built (if $\theta_1 \geq c - \sum_{i \neq 1} \bar{\theta}_i$ it is; if $\theta_1 < c - \sum_{i \neq 1} \bar{\theta}_i$ it is not), and that the sum of the utilities of agents $2, \ldots, I$ is strictly greater if the bridge is built under this egalitarian contribution rule than if it is not built [since $\sum_{i \neq 1} \bar{\theta}_i - c(I-1)/I > 0$].

Let us examine agent 1's incentives for truthfully revealing his type when $\theta_1 = c - \sum_{i \neq 1} \theta_i + \varepsilon$ for $\varepsilon > 0$. If agent 1 reveals his true preferences, the bridge will be built because

$$\left(c - \sum_{i \neq 1} \bar{\theta}_i + \varepsilon\right) + \sum_{i \neq 1} \bar{\theta}_i > c.$$

Agent 1's utility in this case is

$$\begin{aligned} \theta_1 + \bar{m}_1 - \frac{c}{I} &= \left(c - \sum_{i \neq 1} \bar{\theta}_i + \varepsilon\right) + \bar{m}_1 - \frac{c}{I} \\ &= \left(\frac{c(I-1)}{I} - \sum_{i \neq 1} \bar{\theta}_i + \varepsilon\right) + \bar{m}_1. \end{aligned}$$

But, for $\varepsilon > 0$ small enough, this is less than $\bar{m}_1$, which is agent 1's utility if he instead claims that $\theta_1 = 0$, a claim that results in the bridge not being built. Thus, agent 1 will prefer not to tell the truth. Intuitively, under this allocation rule, when agent 1 causes the bridge to be built he has a positive externality on the other agents (in the aggregate). Because he fails to internalize this effect, he has an incentive to understate his benefit from the project. ■

Example 23.B.4: *Allocation of a Single Unit of an Indivisible Private Good.* Consider a setting in which there is a single unit of an indivisible private good to be allocated to one of I agents. Monetary transfers can also be made. An outcome here may be represented by a vector $x = (y_1, \ldots, y_I, t_1, \ldots, t_I)$, where $y_i = 1$ if agent i gets the good, $y_i = 0$ if agent i does not get the good, and t_i is the monetary transfer received by agent i. The set of feasible alternatives is then

$$X = \{(y_1, \ldots, y_I, t_1, \ldots, t_I) : y_i \in \{0, 1\} \text{ and } t_i \in \mathbb{R} \text{ for all } i, \sum_i y_i = 1, \text{ and } \sum_i t_i \leq 0\}.$$

We suppose that type θ_i's Bernoulli utility function takes the quasilinear form

$$u_i(x, \theta_i) = \theta_i y_i + (\bar{m}_i + t_i),$$

where $\bar{m}_i$ is once again agent i's initial endowment of the numeraire ("money"). Here $\theta_i \in \mathbb{R}$ can be viewed as agent i's valuation of the good, and we take the set of possible valuations for agent i to be $\Theta_i = [\underline{\theta}_i, \bar{\theta}_i] \subset \mathbb{R}$.

In this situation, a social choice function $f(\theta) = (y_1(\theta), \ldots, y_I(\theta), t_1(\theta), \ldots, t_I(\theta))$ is ex post efficient if it always allocates the good to the agent who has the highest valuation (or to one of them if there are several) and if it involves no waste of the numeraire; that is, if for all $\theta = (\theta_1, \ldots, \theta_I) \in \Theta_1 \times \cdots \times \Theta_I$,

$$y_i(\theta)(\theta_i - \text{Max}\{\theta_1, \ldots, \theta_I\}) = 0 \qquad \text{for all } i$$

and

$$\sum_i t_i(\theta) = 0.$$

Two special cases that have received a great deal of attention in the literature deserve mention. The first is the case of *bilateral trade*. In this case we have $I = 2$; agent 1 is interpreted as the initial owner of the good (the "seller"), and agent 2 is the potential purchaser of the good (the "buyer"). When $\underline{\theta}_2 > \bar{\theta}_1$ there are certain to be gains from trade regardless of the realizations of θ_1 and θ_2; when $\underline{\theta}_1 > \bar{\theta}_2$ there are certain to be no gains from trade; finally, if $\underline{\theta}_2 < \bar{\theta}_1$ and $\underline{\theta}_1 < \bar{\theta}_2$ then there may or may not be gains from trade, depending on the realization of θ.

The second special case is the *auction* setting. Here, one agent, whom we shall designate as agent 0, is interpreted as the seller of the good (the "auctioneer") and is assumed to derive no value from it (more generally, the seller might have a known value $\theta_0 = \bar{\theta}_0$ different from zero). The other agents, $1, \ldots, I$, are potential buyers (the "bidders").[5]

To illustrate the problem with information revelation in this example, consider an auction setting with two buyers ($I = 2$). In the previous examples, we simplified the discussion of information revelation by assuming that only one agent has more than one possible type. We now suppose instead that both buyers' (privately observed) valuations θ_i are drawn independently from the uniform distribution on $[0, 1]$ and that this fact is common knowledge among the agents. Consider the social choice function $f(\theta) = (y_0(\theta), y_1(\theta), y_2(\theta), t_0(\theta), t_1(\theta), t_2(\theta))$ in which

$$y_1(\theta) = 1 \quad \text{if } \theta_1 \geq \theta_2; \quad = 0 \text{ if } \theta_1 < \theta_2 \tag{23.B.3}$$

$$y_2(\theta) = 1 \quad \text{if } \theta_1 < \theta_2; \quad = 0 \text{ if } \theta_1 \geq \theta_2 \tag{23.B.4}$$

$$y_0(\theta) = 0 \quad \text{for all } \theta \tag{23.B.5}$$

$$t_1(\theta) = -\theta_1 y_1(\theta) \tag{23.B.6}$$

$$t_2(\theta) = -\theta_2 y_2(\theta) \tag{23.B.7}$$

$$t_0(\theta) = -(t_1(\theta) + t_2(\theta)). \tag{23.B.8}$$

In this social choice function, the seller gives the good to the buyer with the highest valuation (to buyer 1 if there is a tie) and this buyer gives the seller a payment equal to his valuation (the other, low-valuation buyer makes no transfer payment to the seller). Note that $f(\cdot)$ is not only ex post efficient but also is very attractive for the seller: if $f(\cdot)$ can be implemented, the seller will capture all of the consumption benefits that are generated by the good.

Suppose we try to implement this social choice function. Assume that the buyers are expected utility maximizers. We now ask: If buyer 2 always announces his true value, will buyer 1 find it optimal to do the same? For each value of θ_1, buyer 1's problem is to choose the valuation to announce, say $\hat{\theta}_1$, so as to solve

$$\max_{\hat{\theta}_1} \quad (\theta_1 - \hat{\theta}_1) \text{ Prob } (\theta_2 \leq \hat{\theta}_1)$$

or

$$\max_{\hat{\theta}_1} \quad (\theta_1 - \hat{\theta}_1)\hat{\theta}_1.$$

5. Note that, for ease of notation, we take there to be $I + 1$ agents in the auction setting.

The solution to this problem has buyer 1 set $\hat{\theta}_1 = \theta_1/2$. We see then that if buyer 2 always tells the truth, truth telling is *not* optimal for buyer 1. A similar point applies to buyer 2. Intuitively, for this social choice function, a buyer has an incentive to understate his valuation so as to lower the transfer he must make in the event that he has the highest announced valuation and gets the good. The cost to him of doing this is that he gets the good less often, but this is a cost worth incurring to at least some degree.[6] Thus, we again see that there may be a problem in implementing certain social choice functions in settings in which information is privately held. (For a similar point in the bilateral trade context, see Exercise 23.B.2.)

Although buyers have an incentive to lie given the social choice function described in (23.B.3) to (23.B.8), this is *not* true of *all* social choice functions in this auction setting. To see this point, suppose we try to implement the social choice function $\tilde{f}(\cdot)$ that has the same allocation rule as that above [i.e., in which the functions $y_i(\cdot)$ for $i = 0, 1, 2$ are the same as those described in (23.B.3) to (23.B.5)] but instead has transfer functions

$$t_1(\theta) = -\theta_2 y_1(\theta)$$

$$t_2(\theta) = -\theta_1 y_2(\theta)$$

$$t_0(\theta) = -(t_1(\theta) + t_2(\theta)).$$

In this social choice function, instead of buyer i paying the seller an amount equal to his own valuation θ_i if he wins the object, he now pays θ_j, where $j \neq i$; that is, he pays an amount equal to the *second-highest valuation.* Consider buyer 1's incentives for truth telling now. If buyer 2 announces his valuation to be $\hat{\theta}_2 \leq \theta_1$, buyer 1 can receive a utility of $(\theta_1 - \hat{\theta}_2) \geq 0$ by truthfully announcing that his valuation is θ_1. For any other announcement, buyer 1's resulting utility is either the same (if he announces a valuation of at least $\hat{\theta}_2$) or zero (if he announces a valuation below $\hat{\theta}_2$). So if $\hat{\theta}_2 \leq \theta_1$, announcing the truth is weakly best for buyer 1. On the other hand, if buyer 2's announced valuation is $\hat{\theta}_2 > \theta_1$, then buyer 1's utility is 0 if he reveals his true valuation. However, buyer 1 can receive only a negative utility by making a false claim that gets him the good (a claim that his valuation is at least $\hat{\theta}_2$). We conclude that truth telling is optimal for buyer 1 regardless of what buyer 2 announces. Formally, in the language of the theory of games, truth telling is a weakly dominant strategy for buyer 1 (see Section 8.B). A similar conclusion follows for buyer 2. Thus, this social choice function *is* implementable even though the buyers' valuations are private information: it suffices to simply ask each buyer to report his type, and then to choose $\tilde{f}(\theta)$.[7] ■

Examples 23.B.1 to 23.B.4 suggest that when agents' types are privately observed the information revelation problem may constrain the set of social choice functions that can be successfully implemented. With these examples as motivation, we can now pose the central question that is our focus in this chapter: *What social choice functions can be implemented when agents' types are private information?*

6. This trade-off is similar to that faced by a monopolist (see Section 12.B): when the monopolist raises his price, he lowers his sales but makes more on his remaining sales.

7. For other examples of implementable social choice functions, see Exercise 23.B.1.

To answer this question, we need in principle to begin by thinking of all the possible ways in which a social choice function might be implemented. In the above examples we have implicitly imagined a very simple scenario in which each agent i is asked to directly reveal θ_i and then, given the announcements $(\hat{\theta}_1, \ldots, \hat{\theta}_I)$, the alternative $f(\hat{\theta}_1, \ldots, \hat{\theta}_I) \in X$ is chosen. But this is not the only way a social choice function might be implemented. In particular, a given social choice function might be *indirectly* implemented by having the agents interact through some type of institution in which there are rules governing the actions the agents may take and how these actions translate into a social outcome. To illustrate this point, Examples 23.B.5 and 23.B.6 study two commonly used auction institutions.

Example 23.B.5: *First-Price Sealed-Bid Auction.* Consider again the auction setting introduced in Example 23.B.4. In a *first-price sealed-bid auction* each potential buyer i is allowed to submit a sealed bid, $b_i \geq 0$. The bids are then opened and the buyer with the highest bid gets the good and pays an amount equal to his bid to the seller.[8]

To be specific, consider again the case where there are two potential buyers ($I = 2$) and each θ_i is independently drawn from the uniform distribution on [0, 1]. We will look for an equilibrium in which each buyer's strategy $b_i(\cdot)$ takes the form $b_i(\theta_i) = \alpha_i \theta_i$ for $\alpha_i \in [0, 1]$. Suppose that buyer 2's strategy has this form, and consider buyer 1's problem. For each θ_1 he wants to solve

$$\underset{b_1 \geq 0}{\text{Max}} \quad (\theta_1 - b_1)\,\text{Prob}(b_2(\theta_2) \leq b_1).$$

Since buyer 2's highest possible bid is α_2 (he submits a bid of α_2 when $\theta_2 = 1$), it is evident that buyer 1 should never bid more than α_2. Moreover, since θ_2 is uniformly distributed on [0, 1] and $b_2(\theta_2) \leq b_1$ if and only if $\theta_2 \leq (b_1/\alpha_2)$, we can write buyer 1's problem as

$$\underset{b_1 \in [0, \alpha_2]}{\text{Max}} \quad (\theta_1 - b_1)(b_1/\alpha_2).$$

The solution to this problem is

$$b_1(\theta_1) = \begin{cases} \frac{1}{2}\theta_1 & \text{if } \frac{1}{2}\theta_1 \leq \alpha_2, \\ \alpha_2 & \text{if } \frac{1}{2}\theta_1 > \alpha_2. \end{cases}$$

By similar reasoning,

$$b_2(\theta_2) = \begin{cases} \frac{1}{2}\theta_2 & \text{if } \frac{1}{2}\theta_2 \leq \alpha_1, \\ \alpha_1 & \text{if } \frac{1}{2}\theta_2 > \alpha_1. \end{cases}$$

Letting $\alpha_1 = \alpha_2 = \frac{1}{2}$, we see that the strategies $b_i(\theta_i) = \frac{1}{2}\theta_i$ for $i = 1, 2$ constitute a Bayesian Nash equilibrium for this auction. Thus, there is a Bayesian Nash equilibrium of this first-price sealed-bid auction that indirectly yields the outcomes specified by the social choice function $f(\theta) = (y_0(\theta), y_1(\theta), y_2(\theta), t_0(\theta), t_1(\theta), t_2(\theta))$

8. If there are several highest bids, we suppose that the lowest numbered of these bidders gets the good. We could equally well randomize among the highest bidders if there are more than one, but this would require that we expand the set of alternatives to $\Delta(X)$, the set of all lotteries over X. In fact, we do precisely this when we study auctions in Sections 23.D and 23.F.

in which

$$y_1(\theta) = 1 \quad \text{if } \theta_1 \geq \theta_2; \quad = 0 \text{ if } \theta_1 < \theta_2 \tag{23.B.9}$$

$$y_2(\theta) = 1 \quad \text{if } \theta_1 < \theta_2; \quad = 0 \text{ if } \theta_1 \geq \theta_2 \tag{23.B.10}$$

$$y_0(\theta) = 0 \quad \text{for all } \theta \tag{23.B.11}$$

$$t_1(\theta) = -\tfrac{1}{2}\theta_1 y_1(\theta) \tag{23.B.12}$$

$$t_2(\theta) = -\tfrac{1}{2}\theta_2 y_2(\theta) \tag{23.B.13}$$

$$t_0(\theta) = -(t_1(\theta) + t_2(\theta)). \tag{23.B.14}$$

■

Example 23.B.6: *Second-Price Sealed-Bid Auction.*[9] Once again, consider the auction setting described in Example 23.B.4. In a *second-price sealed-bid auction*, each potential buyer i is allowed to submit a sealed bid, $b_i \geq 0$. The bids are then opened and the buyer with the highest bid gets the good, but now he pays the seller an amount equal to the *second-highest* bid.[10]

By reasoning that parallels that at the end of Example 23.B.4, the strategy $b_i(\theta_i) = \theta_i$ for all $\theta_i \in [0, 1]$ is a weakly dominant strategy for each buyer i (see Exercise 23.B.3). Thus, when $I = 2$ the second-price sealed-bid auction implements the social choice function $f(\theta) = (y_0(\theta), y_1(\theta), y_2(\theta), t_0(\theta), t_1(\theta), t_2(\theta))$ in which

$$y_1(\theta) = 1 \quad \text{if } \theta_1 \geq \theta_2; \quad = 0 \text{ if } \theta_1 < \theta_2$$

$$y_2(\theta) = 1 \quad \text{if } \theta_1 < \theta_2; \quad = 0 \text{ if } \theta_1 \geq \theta_2$$

$$y_0(\theta) = 0 \quad \text{for all } \theta$$

$$t_1(\theta) = -\theta_2 y_1(\theta)$$

$$t_2(\theta) = -\theta_1 y_2(\theta)$$

$$t_0(\theta) = -(t_1(\theta) + t_2(\theta)). \quad ■$$

Examples 23.B.5 and 23.B.6 illustrate that, as a general matter, we need to consider not only the possibility of directly implementing social choice functions by asking agents to reveal their types but also their indirect implementation through the design of institutions in which the agents interact. The formal representation of such an institution is known as a *mechanism*.

Definition 23.B.3: A *mechanism* $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ is a collection of I strategy sets $(S_1, \ldots, S_I)$ and an outcome function $g: S_1 \times \cdots \times S_I \to X$.

A mechanism can be viewed as an institution with rules governing the procedure for making the collective choice. The allowed actions of each agent i are summarized by the strategy set S_i, and the rule for how agents' actions get turned into a social choice is given by the outcome function $g(\cdot)$.

Formally, the mechanism Γ combined with possible types $(\Theta_1, \ldots, \Theta_I)$, probability density $\phi(\cdot)$, and Bernoulli utility functions $(u_1(\cdot), \ldots, u_I(\cdot))$ defines

9. This auction is also called a *Vickrey auction*, after Vickrey (1961).

10. If there is more than one highest bid, we again select the lowest-numbered of these bidders.

a Bayesian game of incomplete information. That is, letting $\tilde{u}_i(s_1, \ldots, s_I, \theta_i) = u_i(g(s_1, \ldots, s_I), \theta_i)$, the game

$$[I, \{S_i\}, \{\tilde{u}_i(\cdot)\}, \Theta_1 \times \cdots \times \Theta_I, \phi(\cdot)]$$

is exactly the type of Bayesian game studied in Section 8.E. Note that a mechanism could in principle be a complex dynamic procedure, in which case the elements of the strategy sets S_i would consist of contingent plans of action (see Chapter 7).[11]

For the auction setting, the first-price sealed-bid auction is the mechanism in which $S_i = \mathbb{R}_+$ for all i and, given the bids $(b_1, \ldots, b_I) \in \mathbb{R}_+^I$, the outcome function $g(b_1, \ldots, b_I) = (\{y_i(b_1, \ldots, b_I)\}_{i=1}^I, \{t_i(b_1, \ldots, b_I)\}_{i=1}^I)$ is such that

$$y_i(b_1, \ldots, b_I) = 1 \quad \text{if and only if} \quad i = \text{Min}\{j: b_j = \text{Max}\{b_1, \ldots, b_I\}\},$$

$$t_i(b_1, \ldots, b_I) = -b_i y_i(b_1, \ldots, b_I).$$

In the second-price sealed-bid auction, on the other hand, we have the same strategy sets and functions $y_i(\cdot)$, but instead $t_i(b_1, \ldots, b_I) = -\text{Max}\{b_j: j \neq i\} y_i(b_1, \ldots, b_I)$.

A strategy for agent i in the game of incomplete information created by a mechanism Γ is a function $s_i: \Theta_i \to S_i$ giving agent i's choice from S_i for each possible type in Θ_i that he might have. Loosely put, we say that a mechanism *implements* social choice function $f(\cdot)$ if there is an equilibrium of the game induced by the mechanism that yields the same outcomes as $f(\cdot)$ for each possible profile of types $\theta = (\theta_1, \ldots, \theta_I)$. This is stated formally in Definition 23.B.4.

Definition 23.B.4: The mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ *implements* social choice function $f(\cdot)$ if there is an equilibrium strategy profile $(s_1^*(\cdot), \ldots, s_I^*(\cdot))$ of the game induced by Γ such that $g(s_1^*(\theta_1), \ldots, s_I^*(\theta_I)) = f(\theta_1, \ldots, \theta_I)$ for all $(\theta_1, \ldots, \theta_I) \in \Theta_1 \times \cdots \times \Theta_I$.

Note, however, that we have not specified in Definition 23.B.4 exactly what we mean by an "equilibrium". This is because, as we have seen in Part II, there is no single equilibrium concept that is universally agreed upon as *the* appropriate solution concept for games. As a result, the mechanism design literature has investigated the implementation question for a variety of solution concepts. In Sections 23.C and 23.D we focus on two central solution concepts: dominant strategy equilibrium and Bayesian Nash equilibrium.[12]

Note also that the notion of implementation that we have adopted in Definition 23.B.4 is in one sense a weak one: in particular, the mechanism Γ may have *more than one equilibrium*, but Definition 23.B.4 requires only that *one of them* induce outcomes in accord with $f(\cdot)$. Implicitly, then, Definition 23.B.4 assumes that, if multiple equilibria exist, the agents will play the equilibrium that the mechanism designer wants. Throughout the chapter we shall stick to this notion of implementation. Appendix A is devoted to a further discussion of this issue.

11. Note also that we are representing the game created by a mechanism using its normal form. For all the analysis that follows in the text this will be sufficient. In Appendix B, however, we consider a case where the extensive form representation is used.

12. Appendix B considers several other equilibrium concepts in the special context of *complete information* settings in which the players observe each others' types.

The identification of all social choice functions that are implementable may seem like a daunting task because, in principle, it appears that we need to consider all possible mechanisms—a very large set. Fortunately, an important result known as the *revelation principle* (to be formally stated and proven in Sections 23.C and 23.D) tells us that we can often restrict attention to the very simple type of mechanisms that we were implicitly considering at the outset, that is, mechanisms in which each agent is asked to reveal his type, and given the announcements $(\hat{\theta}_1, \ldots, \hat{\theta}_I)$, the alternative chosen is $f(\hat{\theta}_1, \ldots, \hat{\theta}_I) \in X$.[13] These are known as *direct revelation mechanisms*, and formally constitute a special case of the mechanisms of Definition 23.B.3.

Definition 23.B.5: A *direct revelation mechanism* is a mechanism in which $S_i = \Theta_i$ for all i and $g(\theta) = f(\theta)$ for all $\theta \in \Theta_1 \times \cdots \times \Theta_I$.

Moreover, as we shall see, the revelation principle also tells us that we can further restrict our attention to direct revelation mechanisms in which *truth telling is an optimal strategy for each agent*. This fact motivates the notion of *truthful implementation* that we introduce in Definition 23.B.6 (we are again purposely vague in the definition about the equilibrium concept we wish to employ).

Definition 23.B.6: The social choice function $f(\cdot)$ is *truthfully implementable* (or *incentive compatible*) if the direct revelation mechanism $\Gamma = (\Theta_1, \ldots, \Theta_I, f(\cdot))$ has an equilibrium $(s_1^*(\cdot), \ldots, s_I^*(\cdot))$ in which $s_i^*(\theta_i) = \theta_i$ for all $\theta_i \in \Theta_i$ and all $i = 1, \ldots, I$; that is, if truth telling by each agent i constitutes an equilibrium of $\Gamma = (\Theta_1, \ldots, \Theta_I, f(\cdot))$.

To offer a hint as to why we may be able to restrict attention to direct revelation mechanisms that induce truth telling, we briefly verify that the social choice functions that are implemented indirectly through the first-price and second-price sealed-bid auctions of Examples 23.B.5 and 23.B.6 can also be truthfully implemented using a direct revelation mechanism. In fact, for the second-price sealed-bid auction of Example 23.B.6 we have already seen this fact, because the social choice function implemented by the second-price auction is exactly the social choice function that we studied at the end of Example 23.B.4 in which truth telling is a weakly dominant strategy for both buyers. Example 23.B.7 considers the first-price sealed-bid auction.

Example 23.B.7: *Truthful Implementation of the Social Choice Function Implemented by the First-Price Sealed-Bid Auction.* When facing the direct revelation mechanism $(\Theta_1, \ldots, \Theta_I, f(\cdot))$ with $f(\theta) = (y_0(\theta), y_1(\theta), y_2(\theta), t_0(\theta), t_1(\theta), t_2(\theta))$ satisfying (23.B.9) to (23.B.14), buyer 1's optimal announcement $\hat{\theta}_1$ when he has type θ_1 solves

$$\max_{\hat{\theta}_1} \quad (\theta_1 - \tfrac{1}{2}\hat{\theta}_1)\,\text{Prob}(\theta_2 \leq \hat{\theta}_1)$$

or

$$\max_{\hat{\theta}_1} \quad (\theta_1 - \tfrac{1}{2}\hat{\theta}_1)\hat{\theta}_1.$$

13. Some early versions of the revelation principle were derived by Gibbard (1973), Green and Laffont (1977), Myerson (1979), and Dasgupta, Hammond and Maskin (1979).

The first-order condition for this problem gives $\hat{\theta}_1 = \theta_1$. So truth telling is buyer 1's optimal strategy given that buyer 2 always tells the truth. A similar conclusion follows for buyer 2. Thus, the social choice function implemented by the first-price sealed-bid auction (in a Bayesian Nash equilibrium) can also be truthfully implemented (in a Bayesian Nash equilibrium) through a direct revelation mechanism. That is, the social choice function (23.B.9) to (23.B.14) is incentive compatible. ■

Because of the revelation principle, when we explore in Sections 23.C and 23.D the constraints that incomplete information about types puts on the set of implementable social choice functions, we will be able to restrict our analysis to identifying those social choice functions that can be truthfully implemented.

Finally, we note that, in some applications, participation in the mechanism may be *voluntary*, and so a social choice function must not only induce truthful revelation of information but must also satisfy certain *participation* (or *individual rationality*) *constraints* if it is to be successfully implemented. In Sections 23.C and 23.D, however, we shall abstract from issues of participation to focus exclusively on the information revelation problem. We introduce participation constraints in Section 23.E.

23.C Dominant Strategy Implementation

In this section, we study implementation in *dominant strategies.*[14] Throughout we follow the notation introduced in Section 23.B: The vector of agents' types $\theta = (\theta_1, \ldots, \theta_I)$ is drawn from the set $\Theta = \Theta_1 \times \cdots \times \Theta_I$ according to a probability density $\phi(\cdot)$, and agent i's Bernoulli utility function over the alternatives in X given his type θ_i is $u_i(x, \theta_i)$. We also adopt the notational convention of writing $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_I)$, $\theta = (\theta_i, \theta_{-i})$, and $\Theta_{-i} = \Theta_1 \times \cdots \times \Theta_{i-1} \times \Theta_{i+1} \times \cdots \times \Theta_I$. A mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ is a collection of I sets $S_1, \ldots, S_I$, each S_i containing agent i's possible actions (or plans of action), and an outcome function $g: S \to X$, where $S = S_1 \times \cdots \times S_I$. As discussed in Section 23.B, a mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ combined with possible types $(\Theta_1, \ldots, \Theta_I)$, density $\phi(\cdot)$, and Bernoulli utility functions $(u_1(\cdot), \ldots, u_I(\cdot))$ defines a Bayesian game of incomplete information (see Section 8.E). We will also often write $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_I)$, $s = (s_i, s_{-i})$, and $S_{-i} = S_1 \times \cdots \times S_{i-1} \times S_{i+1} \times \cdots \times S_I$.

Recall from Section 8.B that a strategy is a weakly dominant strategy for a player in a game if it gives him at least as large a payoff as any of his other possible strategies for every possible strategy that his rivals might play. In the present incomplete information environment, strategy $s_i: \Theta_i \to S_i$ is a weakly dominant strategy for agent i in mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ if, for all $\theta_i \in \Theta_i$ and all possible strategies for agents $j \neq i$, $s_{-i}(\cdot) = [s_1(\cdot), \ldots, s_{i-1}(\cdot), s_{i+1}(\cdot), \ldots, s_I(\cdot)]$,[15]

$$E_{\theta_{-i}}[u_i(g(s_i(\theta_i), s_{-i}(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(\hat{s}_i, s_{-i}(\theta_{-i})), \theta_i)|\theta_i] \quad \text{for all } \hat{s}_i \in S_i. \tag{23.C.1}$$

Condition (23.C.1) holding for all $s_{-i}(\cdot)$ and θ_i is equivalent to the condition that,

14. Good sources for further reading on the subject of this section are Dasgupta, Hammond and Maskin (1979) and Green and Laffont (1979).

15. The expectation in (23.C.1) is taken over realizations of $\theta_{-i} \in \Theta_{-i}$.

for all $\theta_i \in \Theta_i$,

$$u_i(g(s_i(\theta_i), s_{-i}), \theta_i) \geq u_i(g(\hat{s}_i, s_{-i}), \theta_i) \tag{23.C.2}$$

for all $\hat{s}_i \in S_i$ and all $s_{-i} \in S_{-i}$.[16] This leads to Definition 23.C.1.

Definition 23.C.1: The strategy profile $s^*(\cdot) = (s_1^*(\cdot), \ldots, s_I^*(\cdot))$ is a *dominant strategy equilibrium* of mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ if, for all i and all $\theta_i \in \Theta_i$,

$$u_i(g(s_i^*(\theta_i), s_{-i}), \theta_i) \geq u_i(g(s_i', s_{-i}), \theta_i)$$

for all $s_i' \in S_i$ and all $s_{-i} \in S_{-i}$.

We now specialize Definition 23.B.4 to the notion of dominant strategy equilibrium.

Definition 23.C.2: The mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ *implements the social choice function* $f(\cdot)$ *in dominant strategies* if there exists a dominant strategy equilibrium of Γ, $s^*(\cdot) = (s_1^*(\cdot), \ldots, s_I^*(\cdot))$, such that $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

The concept of dominant strategy implementation is of special interest because if we can find a mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ that implements $f(\cdot)$ in dominant strategies, then this mechanism implements $f(\cdot)$ in a very strong and robust way. This is true in several senses. First, we can feel fairly confident that a rational agent who has a (weakly) dominant strategy will indeed play it.[17] Unlike the equilibrium strategies in Nash-related equilibrium concepts, a player need not correctly forecast his opponents' play to justify his play of a dominant strategy. Second, although we have assumed that the agents know the probability density $\phi(\cdot)$ over realizations of the types $(\theta_1, \ldots, \theta_I)$, and hence can deduce the correct conditional probability distribution over realizations of θ_{-i}, if Γ implements $f(\cdot)$ in dominant strategies this implementation will be robust even if agents have incorrect, and perhaps even contradictory, beliefs about this distribution. In particular, agent i's beliefs regarding the distribution of θ_{-i} do not affect the dominance of his strategy $s_i^*(\cdot)$.[18] Third, it follows that if Γ implements $f(\cdot)$ in dominant strategies then it does so regardless of the probability density $\phi(\cdot)$. Thus, the same mechanism can be used to implement $f(\cdot)$ for any $\phi(\cdot)$. One advantage of this is that if the mechanism designer is an outsider (say, the "government"), he need not know $\phi(\cdot)$ to successfully implement $f(\cdot)$.

As we noted in Section 23.B, to identify whether a particular social choice function $f(\cdot)$ is implementable, we need, in principle, to consider all possible mechanisms. Fortunately, it turns out that for dominant strategy implementation it suffices to ask

16. Condition (23.C.2) follows from (23.C.1) simply by setting $s_{-i}(\theta_{-i}) = s_{-i}$ for all $\theta_{-i} \in \Theta_{-i}$. To see that (23.C.2) implies (23.C.1), consider the case where S_{-i} is a finite set. Then, for any s_i,

$$E_{\theta_{-i}}[u_i(g(s_i, s_{-i}(\theta_{-i})), \theta_i)|\theta_i] = \sum_{s_{-i} \in S_{-i}} \text{Prob}(s_{-i}(\theta_{-i}) = s_{-i})\, u_i(g(s_i, s_{-i}), \theta_i).$$

Thus, (23.C.2) implies (23.C.1).

17. We leave aside the question of what might happen if an agent has *several* weakly dominant strategies. This is the issue of multiple equilibria that we discuss in Appendix A. Even so, we at least mention one conclusion from that discussion: The problem of multiple equilibria is relatively small when we are dealing with dominant strategy equilibrium.

18. In fact, the implementation of $f(\cdot)$ using Γ is also robust to substantial relaxations of the hypothesis that agents maximize expected utility.

whether a particular $f(\cdot)$ is truthfully implementable in the sense introduced in Definition 23.C.3.

Definition 23.C.3: The social choice function $f(\cdot)$ is *truthfully implementable in dominant strategies* (or *dominant strategy incentive compatible*, or *strategy-proof*, or *straightforward*) if $s_i^*(\theta_i) = \theta_i$ for all $\theta_i \in \Theta_i$ and $i = 1, \ldots, I$ is a dominant strategy equilibrium of the direct revelation mechanism $\Gamma = (\Theta_1, \ldots, \Theta_I, f(\cdot))$. That is, if for all i and all $\theta_i \in \Theta_i$,

$$u_i(f(\theta_i, \theta_{-i}), \theta_i) \geq u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i) \tag{23.C.3}$$

for all $\hat{\theta}_i \in \Theta_i$ and all $\theta_{-i} \in \Theta_{-i}$.

The ability to restrict our inquiry, without loss of generality, to the question of whether $f(\cdot)$ is truthfully implementable is a consequence of what is known as the *revelation principle for dominant strategies.*

Proposition 23.C.1: (*The Revelation Principle for Dominant Strategies*) Suppose that there exists a mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ that implements the social choice function $f(\cdot)$ in dominant strategies. Then $f(\cdot)$ is truthfully implementable in dominant strategies.

Proof: If $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ implements $f(\cdot)$ in dominant strategies, then there exists a profile of strategies $s^*(\cdot) = (s_1^*(\cdot), \ldots, s_I^*(\cdot))$ such that $g(s^*(\theta)) = f(\theta)$ for all θ and, for all i and all $\theta_i \in \Theta_i$,

$$u_i(g(s_i^*(\theta_i), s_{-i}), \theta_i) \geq u_i(g(\hat{s}_i, s_{-i}), \theta_i) \tag{23.C.4}$$

for all $\hat{s}_i \in S_i$ and all $s_{-i} \in S_{-i}$. Condition (23.C.4) implies, in particular, that for all i and all $\theta_i \in \Theta_i$,

$$u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i) \geq u_i(g(s_i^*(\hat{\theta}_i), s_{-i}^*(\theta_{-i})), \theta_i) \tag{23.C.5}$$

for all $\hat{\theta}_i \in \Theta_i$ and all $\theta_{-i} \in \Theta_{-i}$. Since $g(s^*(\theta)) = f(\theta)$ for all θ, (23.C.5) means that, for all i and all $\theta_i \in \Theta_i$,

$$u_i(f(\theta_i, \theta_{-i}), \theta_i) \geq u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i)$$

for all $\hat{\theta}_i \in \Theta_i$ and all $\theta_{-i} \in \Theta_{-i}$. But, this is precisely condition (23.C.3), the condition for $f(\cdot)$ to be truthfully implementable in dominant strategies. ■

The intuitive idea behind the revelation principle for dominant strategies can be put as follows: Suppose that the indirect mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ implements $f(\cdot)$ in dominant strategies, and that in this indirect mechanism each agent i finds playing $s_i^*(\theta_i)$ when his type is θ_i better than playing any other $s_i \in S_i$ for any choices $s_{-i} \in S_{-i}$ by agents $j \neq i$. Now consider altering this mechanism simply by introducing a mediator who says to each agent i: "You tell me your type, and when you say your type is θ_i, I will play $s_i^*(\theta_i)$ for you." Clearly, if $s_i^*(\theta_i)$ is agent i's optimal choice for each $\theta_i \in \Theta_i$ in the initial mechanism Γ for any strategies chosen by the other agents, then agent i will find telling the truth to be a dominant strategy in this new scheme. But this means that we have found a way to truthfully implement $f(\cdot)$.

The implication of the revelation principle is that to identify the set of social choice functions that are implementable in dominant strategies, we need only identify those that are truthfully implementable. In principle, for any $f(\cdot)$, this is just a matter of checking the inequalities (23.C.3).

The inequalities (23.C.3), which are necessary and sufficient for a social choice function $f(\cdot)$ to be truthfully implementable in dominant strategies, can be usefully thought of in terms of a certain *weak preference reversal property*. In particular, consider any agent i and any pair of possible types for i, θ_i' and θ_i''. If truth telling is a dominant strategy for agent i, then for any $\theta_{-i} \in \Theta_{-i}$ we must have

$$u_i(f(\theta_i', \theta_{-i}), \theta_i') \geq u_i(f(\theta_i'', \theta_{-i}), \theta_i')$$

and

$$u_i(f(\theta_i'', \theta_{-i}), \theta_i'') \geq u_i(f(\theta_i', \theta_{-i}), \theta_i'').$$

That is, agent i's preference ranking of $f(\theta_i', \theta_{-i})$ and $f(\theta_i'', \theta_{-i})$ must *weakly reverse* when his type changes from θ_i' to θ_i'', with him weakly preferring alternative $f(\theta_i', \theta_{-i})$ to $f(\theta_i'', \theta_{-i})$ when his type is θ_i', but weakly preferring alternative $f(\theta_i'', \theta_{-i})$ to $f(\theta_i', \theta_{-i})$ when his type is θ_i''. In the reverse direction, if this weak preference reversal property holds for *all* $\theta_{-i} \in \theta_{-i}$ and *all* pairs θ_i', $\theta_i'' \in \Theta_i$, then truth telling is indeed a dominant strategy for agent i (check this in Exercise 23.C.1).

This weak preference reversal property can be succinctly stated using agent i's *lower contour sets*. Define the lower contour set of alternative x when agent i has type θ_i by (see Section 3.B):

$$L_i(x, \theta_i) = \{z \in X: u_i(x, \theta_i) \geq u_i(z, \theta_i)\}.$$

Using this lower contour set we get the characterization of the set of social choice functions that can be truthfully implemented in dominant strategies that is given in Proposition 23.C.2.

Proposition 23.C.2: The social choice function $f(\cdot)$ is truthfully implementable in dominant strategies if and only if for all i, all $\theta_{-i} \in \Theta_{-i}$, and all pairs of types for agent i, θ_i' and $\theta_i'' \in \Theta_i$, we have

$$f(\theta_i'', \theta_{-i}) \in L_i(f(\theta_i', \theta_{-i}), \theta_i') \quad \text{and} \quad f(\theta_i', \theta_{-i}) \in L_i(f(\theta_i'', \theta_{-i}), \theta_i''). \quad (23.C.6)$$

The idea behind this preference reversal characterization of the social choice functions that can be truthfully implemented in dominant strategies is illustrated in Figures 23.C.1 and 23.C.2. In Figure 23.C.1, we represent the social choice function $f(\cdot)$ for each possible configuration of types (θ_1, θ_2) in a situation in which there are two agents ($I = 2$), two possible values of θ_1, and three possible values of θ_2. Consider agent 1's incentives to tell the truth. If truth telling is a weakly dominant strategy for agent 1, then when his type changes from θ_1' to θ_1'', he must experience a weak preference reversal between outcomes $f(\theta_1', \theta_2)$ and $f(\theta_1'', \theta_2)$ for each possible value of θ_2. A similar point applies for agent 2.

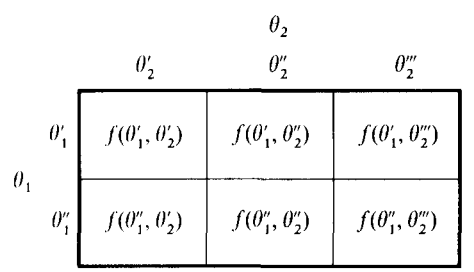


Figure 23.C.1
For agent 1 to find truth telling to be his dominant strategy, he must experience a weak preference reversal between outcomes $f(\theta_1', \theta_2)$ and $f(\theta_1'', \theta_2)$ when his type changes from θ_1' to θ_1'', for each possible θ_2.

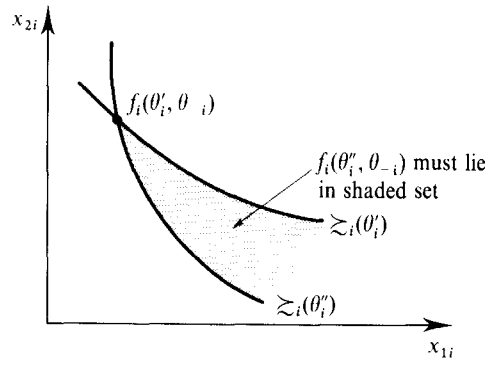


Figure 23.C.2
The weak preference reversal property of Proposition 23.C.2 when preferences satisfy the single-crossing property.

Figure 23.C.2 depicts a change in some agent i's type from θ_i' to θ_i'' in an exchange setting in which agent i's preferences satisfy the *single-crossing property* that we discussed in Sections 13.C and 14.C. In the figure, we denote agent i's allocation in outcome $f(\theta_1, \theta_2)$ by $f_i(\theta_1, \theta_2)$. According to Proposition 23.C.2, $f_i(\theta_i'', \theta_{-i})$ must lie in the shaded region of the figure if truth telling is to be a dominant strategy for agent i. Thus, the characterization in Proposition 23.C.2 can be seen as a multiperson extension of the truth-telling constraints that we encountered in Section 14.C (here they must hold for every possible $\theta_{-i} \in \Theta_{-i}$).

In the remainder of this section we explore in more detail the characteristics of social choice functions that can be truthfully implemented in dominant strategies.

The Gibbard–Satterthwaite Theorem

The Gibbard–Satterwaite theorem was discovered independently in the early 1970s by the two named authors [Gibbard (1973), Satterthwaite (1975)]. It is an impossibility result similar in spirit to Arrow's theorem (Proposition 21.C.1), and has shaped the course of research on incentives and implementation to a great extent. It shows that for a very general class of problems there is no hope of implementing satisfactory social choice functions in dominant strategies.

In what follows, we let $\mathscr{P}$ denote the set of all rational preference relations $\succsim$ on X having the property that no two alternatives are indifferent, and we recall that $\mathscr{R}_i = \{\succsim_i : \succsim_i = \succsim_i(\theta_i)$ for some $\theta_i \in \Theta_i\}$ is agent i's set of possible ordinal preference relations over X. We denote by $f(\Theta)$ the image of $f(\cdot)$; that is, $f(\Theta) = \{x \in X : f(\theta) = x$ for some $\theta \in \Theta\}$. In Definitions 23.C.4 and 23.C.5 we also recall two properties of social choice functions introduced and discussed in Section 21.E.

Definition 23.C.4: The social choice function $f(\cdot)$ is *dictatorial* if there is an agent i such that, for all $\theta = (\theta_1, \ldots, \theta_I) \in \Theta$,

$$f(\theta) \in \{x \in X : u_i(x, \theta_i) \geq u_i(y, \theta_i) \text{ for all } y \in X\}.$$

In words: A social choice function is dictatorial if there is an agent i such that $f(\cdot)$ always chooses one of i's top-ranked alternatives.

Definition 23.C.5: The social choice function $f(\cdot)$ is *monotonic* if, for any θ, if θ' is such that $L_i(f(\theta), \theta_i) \subset L_i(f(\theta), \theta_i')$ for all i [i.e., if $L_i(f(\theta), \theta_i)$ is weakly included in $L_i(f(\theta), \theta_i')$ for all i], then $f(\theta') = f(\theta)$.

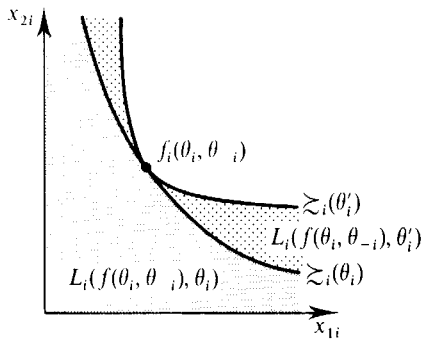


Figure 23.C.3
If $f(\cdot)$ is monotonic, then $f(\theta_i', \theta_{-i}) = f(\theta_i, \theta_{-i})$.

Monotonicity requires the following: Suppose that $f(\theta) = x$, and that the I agents' types change to a $\theta' = (\theta_1', \ldots, \theta_I')$ with the property that no agent finds that some alternative that was weakly worse for him than x when his type was θ_i becomes strictly preferred to x when his type is θ_i'. Then x must still be the social choice. The monotonicity property is illustrated in Figure 23.C.3 for an exchange setting. In the figure, $f_i(\theta_i, \theta_{-i})$ represents agent i's allocation in outcome $f(\theta_i, \theta_{-i})$. The figure depicts a change in agent i's type from θ_i to a θ_i' having the property that $L_i(f(\theta_i, \theta_{-i}), \theta_i) \subset L_i(f(\theta_i, \theta_{-i}), \theta_i')$. If $f(\cdot)$ is monotonic, then $f(\theta_i', \theta_{-i}) = f(\theta_i, \theta_{-i})$.

With these definitions we now state and prove the Gibbard–Satterthwaite theorem.

Proposition 23.C.3: (*The Gibbard Satterthwaite Theorem*) Suppose that X is finite and contains at least three elements, that $\mathcal{R}_i = \mathcal{P}$ for all i, and that $f(\Theta) = X$.[19] Then the social choice function $f(\cdot)$ is truthfully implementable in dominant strategies if and only if it is dictatorial.

Proof: It is immediate that a dictatorial $f(\cdot)$ is truthfully implementable (check for yourself that every agent will tell the truth). We now show that if $f(\cdot)$ is truthfully implementable in dominant strategies then it must be dictatorial. The argument consists of three steps.

Step 1: If $\mathcal{R}_i = \mathcal{P}$ for all i, and $f(\cdot)$ is truthfully implementable in dominant strategies, then $f(\cdot)$ is monotonic.

Consider two profiles of types θ and θ' such that $L_i(f(\theta), \theta_i) \subset L_i(f(\theta), \theta_i')$ for all i. We want to show that $f(\theta') = f(\theta)$. Let us begin by determining $f(\theta_1', \theta_2, \ldots, \theta_I)$. By Proposition 23.C.2 we know that we must have $f(\theta_1', \theta_2, \ldots, \theta_I) \in L_1(f(\theta), \theta_1)$. Hence, $f(\theta_1', \theta_2, \ldots, \theta_I) \in L_1(f(\theta), \theta_1')$. But Proposition 23.C.2 also implies that $f(\theta) \in L_1(f(\theta_1', \theta_2, \ldots, \theta_I), \theta_1')$. Since, by hypothesis, no two alternatives can be indifferent in preference relation $\succsim_1(\theta_1')$, this must imply that $f(\theta_1', \theta_2, \ldots, \theta_I) = f(\theta)$. The same line of argument can be used to show next that $f(\theta_1', \theta_2', \theta_3, \ldots, \theta_I) = f(\theta)$. Indeed, proceeding iteratively, we establish that $f(\theta') = f(\theta)$. Thus, $f(\cdot)$ is monotonic.

Step 2: If $\mathcal{R}_i = \mathcal{P}$ for all i, $f(\cdot)$ is monotonic, and $f(\Theta) = X$, then $f(\cdot)$ is ex post efficient.

19. Strictly speaking, finiteness of the set X is not required for the result. But in the absence of finiteness, our assumption that agents are expected utility maximizers may not be compatible with the condition that $\mathcal{R}_i = \mathcal{P}$ (e.g., if $X = \mathbb{R}_+^L$, the lexicographic preference relation studied in Example 3.C.1 is a strict preference relation that is not representable by a utility function). For a proof that Proposition 23.C.3 continues to be true if we let X be an arbitrary set and $\mathcal{R}_i$ be the set of all *continuous* preferences on X, see Barberà and Peleg (1990).

To verify this, suppose not. Then there is a $\theta \in \Theta$ and a $y \in X$ such that $u_i(y, \theta_i) > u_i(f(\theta), \theta_i)$ for all i (recall that no two alternatives can be indifferent). Because $f(\Theta) = X$, there exists a $\theta' \in \Theta$ such that $f(\theta') = y$. Now choose a vector of types $\theta'' \in \Theta$ such that, for all i, $u_i(y, \theta_i'') > u_i(f(\theta), \theta_i'') > u_i(z, \theta_i'')$ for all $z \neq f(\theta), y$. (Remember that all preferences in $\mathscr{P}$ are possible.) Since $L_i(y, \theta_i') \subset L_i(y, \theta_i'')$ for all i, monotonicity implies that $f(\theta'') = y$. But, since $L_i(f(\theta), \theta_i) \subset L_i(f(\theta), \theta_i'')$ for all i, monotonicity also implies that $f(\theta'') = f(\theta)$: a contradiction because $y \neq f(\theta)$. Hence, $f(\cdot)$ must be ex post efficient.

Step 3: A social choice function $f(\cdot)$ that is monotonic and ex post efficient is necessarily dictatorial.

Step 3 follows directly from Proposition 21.E.1.

Together, steps 1 to 3 establish the result. ■

It should be noted that the conclusion of Proposition 23.C.3 does *not* follow if X contains two elements. For example, in this case, a majority voting social choice function (see Section 21.E) is both nondictatorial and truthfully implementable in dominant strategies (Exercise 23.C.2).

Note also that when $\mathscr{R}_i = \mathscr{P}$ for all i, any ex post efficient social choice function *must* have $f(\Theta) = X$ (verify this in Exercise 23.C.3). Thus, the Gibbard–Satterthwaite theorem tells us that when $\mathscr{R}_i = \mathscr{P}$ for all i, and X contains more than two elements, the only ex post efficient social choice functions that are truthfully implementable in dominant strategies are dictatorial social choice functions.

Given this negative conclusion, if we are to have any hope of implementing desirable social choice functions, we must either weaken the demands of our implementation concept by accepting implementation by means of less robust equilibrium notions (such as Bayesian Nash equilibria) or we must focus on more restricted environments. In the remainder of this section, we follow the latter course, studying the possibilities for implementing desirable social choice functions in dominant strategies when preferences take a quasilinear form. Section 23.D explores the former possibility: It studies implementation in Bayesian Nash equilibria.

Proposition 23.C.3 is readily extended in two ways. First, the result's conclusion still follows whenever $\mathscr{R}_i$ contains $\mathscr{P}$ (the set of all rational preference relations having the property that no two alternatives are indifferent), and so it extends to environments in which individual indifference is possible. This is stated formally in Corollary 23.C.1.

Corollary 23.C.1: Suppose that X is finite and contains at least three elements, that $\mathscr{P} \subset \mathscr{R}_i$ for all i, and that $f(\Theta) = X$. Then the social choice function $f(\cdot)$ is truthfully implementable in dominant strategies if and only if it is dictatorial.

Proof: It is again immediate that a dictatorial social choice function is truthfully implementable. We now show that under the stated hypotheses $f(\cdot)$ must be dictatorial if it is truthfully implementable.

An implication of Proposition 23.C.3 is that there must be an agent h such that $f(\theta) \in \{x \in X : u_h(x, \theta_h) \geq u_h(y, \theta_h)$ for all $y \in X\}$ whenever $\succsim_i(\theta_i) \in \mathscr{P}$ for all i (see Exercise 23.C.4). Without loss of generality, let this be agent I. Suppose now that the result is not true. Then there is a profile of types $\theta' \in \Theta$ such that $f(\theta') \notin \{x \in X : u_I(x, \theta_I') \geq u_I(y, \theta_I')$ for all

$y \in X\}$. Let $z \in \{x \in X: u_I(x, \theta_I') \geq u_I(y, \theta_I') \text{ for all } y \in X\}$. Now consider a profile of types $\theta'' \in \Theta$ such that (i) $\succsim_i(\theta_i'') \in \mathscr{P}$ for all $i = 1, \ldots, I$; (ii) for all agents $i \neq I$, $u_i(f(\theta'), \theta_i'') > u_i(z, \theta_i'') > u_i(x, \theta_i'')$ for all $x \notin \{f(\theta'), z\}$; and (iii) $u_I(z, \theta_I'') > u_I(f(\theta'), \theta_I'') > u_I(x, \theta_I'')$ for all $x \notin \{f(\theta'), z\}$. Consider the profile of types $(\theta_1'', \theta_2', \ldots, \theta_I')$. By Proposition 23.C.2, we must have $f(\theta') \in L_1(f(\theta_1'', \theta_2', \ldots, \theta_I'), \theta_1'')$, and so it must be that $f(\theta_1'', \theta_2', \ldots, \theta_I') = f(\theta')$. The same argument can be applied iteratively for all $i \neq I$ to show that $f(\theta_1'', \ldots, \theta_{I-1}'', \theta_I') = f(\theta')$. Next, note that (by Proposition 23.C.2) we must have $f(\theta_1'', \ldots, \theta_{I-1}'', \theta_I') \in L_I(f(\theta''), \theta_I')$. Hence, $f(\theta'') \in \{z, f(\theta')\}$. But (by Proposition 23.C.2) we must also have $f(\theta'') \in L_I(f(\theta_1'', \ldots, \theta_{I-1}'', \theta_I'), \theta_I')$, and since $u_I(z, \theta_I') > u_I(f(\theta'), \theta_I')$ this means we cannot have $f(\theta'') = z$. Hence, $f(\theta'') = f(\theta')$. But, since $u_I(z, \theta_I'') > u_I(f(\theta'), \theta_I'')$, this contradicts agent I being a dictator whenever $\succsim_i(\theta_i) \in \mathscr{P}$ for all i. ■

As our second extension, we can derive a related dictatorship result for social choice functions whose image $f(\Theta)$ is smaller than X. We first offer Definition 23.C.6.

Definition 23.C.6: The social choice function $f(\cdot)$ is *dictatorial on set* $\hat{X} \subset X$ if there exists an agent i such that, for all $\theta = (\theta_1, \ldots, \theta_I) \in \Theta$, $f(\theta) \in \{x \in \hat{X}: u_i(x, \theta_i) \geq u_i(y, \theta_i) \text{ for all } y \in \hat{X}\}$.

This weaker notion of dictatorship requires only that $f(\cdot)$ select one of the dictator's most preferred alternatives in $\hat{X}$, rather than in X.

Corollary 23.C.2: Suppose that X is finite, that the number of elements in $f(\Theta)$ is at least three, and that $\mathscr{P} \subset \mathscr{R}_i$ for all $i = 1, \ldots, I$. Then $f(\cdot)$ is truthfully implementable in dominant strategies if and only if it is dictatorial on the set $f(\Theta)$.

Proof: It is immediate that $f(\cdot)$ is truthfully implementable if it is dictatorial on the set $f(\Theta)$, and so we now show that under the stated hypotheses $f(\cdot)$ must be dictatorial on set $f(\Theta)$. If $f: \Theta \to X$ is truthfully implementable in dominant strategies when the set of alternatives is X, then the social choice function $\hat{f}: \Theta \to f(\Theta)$ which has $\hat{f}(\theta) = f(\theta)$ for all $\theta \in \Theta$ is truthfully implementable in dominant strategies when the set of alternatives is $f(\Theta)$. By Corollary 23.C.1, $\hat{f}(\cdot)$ must be dictatorial. Hence, $f(\cdot)$ is dictatorial on the set $f(\Theta)$. ■

The implication flowing from Corollary 23.C.2 is therefore this: When $\mathscr{R}_i \subset \mathscr{P}$ for all i, the set of social choice functions which have an image that contains at least three elements and which are truthfully implementable in dominant strategies is exactly the set of social choice functions that can be implemented (indirectly) by restricting the set of possible choices to some subset $\hat{X} \subset X$ and assigning a single agent i to choose from within this set.

Quasilinear Environments: Groves–Clarke Mechanisms

In this subsection we focus on the special, but much studied, class of environments in which agents have quasilinear preferences. In particular, an alternative is now a vector $x = (k, t_1, \ldots, t_I)$, where k is an element of a finite set K, to be called the "project choice," and $t_i \in \mathbb{R}$ is a transfer of a numeraire commodity ("money") to agent i. Agent i's utility function takes the quasilinear form

$$u_i(x, \theta_i) = v_i(k, \theta_i) + (\bar{m}_i + t_i),$$

where $\bar{m}_i$ is agent i's endowment of the numeraire. We assume that we are dealing with a closed system in which the I agents have no outside source of financing. The set

of alternatives is therefore[20]

$$X = \{(k, t_1, \ldots, t_I): k \in K, t_i \in \mathbb{R} \text{ for all } i, \text{ and } \sum_i t_i \leq 0\}.$$

Note that this environment encompasses the cases studied in Examples 23.B.3 and 23.B.4:

Example 23.C.1: *A Public Project.* We can fit a generalized version of the public project setting of Example 23.B.3 into the framework outlined above. To do so, let K contain the possible levels of a public project (e.g., if $K = \{0, 1\}$, then either the project is "not done" or "done") and denote by $c(k)$ the cost of project level $k \in K$. Suppose that $\tilde{v}_i(k, \theta_i)$ is agent i's *gross* benefit from project level k and that, in the absence of any other transfers, projects will be financed through equal contribution [i.e., each agent i will pay the amount $c(k)/I$].[21] Then, we can write agent i's *net* benefit from project level k when his type is θ_i as $v_i(k, \theta_i) = \tilde{v}_i(k, \theta_i) - (c(k)/I)$. The t_i's are now transfers over and above the payments $c(k)/I$. ■

Example 23.C.2: *Allocation of a Single Unit of an Indivisible Private Good.* Consider the environment described in Example 23.B.4 in which an indivisible unit of a private good is to be allocated to one of I agents. Here the "project choice" $k = (y_1, \ldots, y_I)$ represents the allocation of the private good and $K = \{(y_1, \ldots, y_I): y_i \in \{0, 1\}$ for all i and $\sum_i y_i = 1\}$. Agent i's valuation function takes the form $v_i(k, \theta_i) = \theta_i y_i$. ■

A social choice function in this quasilinear environment takes the form $f(\cdot) = (k(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ where, for all $\theta \in \Theta$, $k(\theta) \in K$ and $\sum_i t_i(\theta) \leq 0$. Note that if the social choice function $f(\cdot)$ is ex post efficient then, for all $\theta \in \Theta$, $k(\theta)$ must satisfy

$$\sum_{i=1}^{I} v_i(k(\theta), \theta_i) \geq \sum_{i=1}^{I} v_i(k, \theta_i) \qquad \text{for all } k \in K. \tag{23.C.7}$$

We begin with a result that identifies a class of social choice functions that satisfy (23.C.7) and that are truthfully implementable in dominant strategies.

Proposition 23.C.4: Let $k^*(\cdot)$ be a function satisfying (23.C.7). The social choice function $f(\cdot) = (k^*(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ is truthfully implementable in dominant strategies if, for all $i = 1, \ldots, I$,

$$t_i(\theta) = \left[\sum_{j \neq i} v_j(k^*(\theta), \theta_j)\right] + h_i(\theta_{-i}), \tag{23.C.8}$$

where $h_i(\cdot)$ is an arbitrary function of θ_{-i}.

Proof: If truth is not a dominant strategy for some agent i, then there exist θ_i, $\hat{\theta}_i$, and θ_{-i} such that

$$v_i(k^*(\hat{\theta}_i, \theta_{-i}), \theta_i) + t_i(\hat{\theta}_i, \theta_{-i}) > v_i(k^*(\theta_i, \theta_{-i}), \theta_i) + t_i(\theta_i, \theta_{-i}).$$

20. Observe that X is not a compact set. This explains what might appear as a small paradox: in this setting, there are no dictatorial social choice functions because any agent i, when allowed to pick his best alternative in X, faces no bound on how much money he can extract from the other agents.

21. Nothing we do depends on this choice for the "base" method of contribution.

Substituting from (23.C.8) for $t_i(\hat{\theta}_i, \theta_{-i})$ and $t_i(\theta_i, \theta_{-i})$, this implies that

$$\sum_{j=1}^{I} v_j(k^*(\hat{\theta}_i, \theta_{-i}), \theta_j) > \sum_{j=1}^{I} v_j(k^*(\theta), \theta_j),$$

which contradicts $k^*(\cdot)$ satisfying (23.C.7). Thus, $f(\cdot)$ must be truthfully implementable in dominant strategies. ■

A direct revelation mechanism $\Gamma = (\Theta_1, \ldots, \Theta_I, f(\cdot))$ in which $f(\cdot) = (k(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ satisfies (23.C.7) and (23.C.8) is known as a *Groves mechanism* or *Groves scheme* [after Groves (1973)].[22] In a Groves mechanism, given the announcements θ_{-i} of agents $j \neq i$, agent i's transfer depends on his announced type only through his announcement's effect on the project choice $k^*(\theta)$. Moreover, the change in agent i's transfer that results when his announcement changes the project decision k is exactly equal to the effect of this change in k on agents $j \neq i$. Put differently, the change in agent i's transfer reflects exactly the *externality* that he is imposing on the other agents. As a result, agent i is led to internalize this externality and make an announcement – namely, truth—that leads to a level of k that maximizes the I agents' joint payoff from the project, $\sum_i v_i(k, \theta_i)$.

A special case of a Groves mechanism was discovered independently by Clarke (1971) and is known as the *Clarke*, or *pivotal, mechanism*. This mechanism corresponds to the case in which $h_i(\theta_{-i}) = -\sum_{j \neq i} v_j(k^*_{-i}(\theta_{-i}), \theta_j)$ where, for all $\theta_{-i} \in \Theta_{-i}$, $k^*_{-i}(\theta_{-i})$ satisfies

$$\sum_{j \neq i} v_j(k^*_{-i}(\theta_{-i}), \theta_j) \geq \sum_{j \neq i} v_j(k, \theta_j) \qquad \text{for all } k \in K.$$

That is, $k^*_{-i}(\theta_{-i})$ is the project level that would be ex post efficient if there were only the $I - 1$ agents $j \neq i$. Agent i's transfer in the Clarke mechanism is then given by

$$t_i(\theta) = \left[\sum_{j \neq i} v_j(k^*(\theta), \theta_j)\right] - \left[\sum_{j \neq i} v_j(k^*_{-i}(\theta_{-i}), \theta_j)\right].$$

Note that agent i's transfer is 0 if his announcement does not change the project decision relative to what would be ex post efficient for agents $j \neq i$ in isolation [i.e., if $k^*(\theta) = k^*_{-i}(\theta_{-i})$], and is negative if it does change the project decision [i.e., if $k^*(\theta) \neq k^*_{-i}(\theta_{-i})$], that is, if agent i is "pivotal" to the efficient project choice. Thus, in the Clarke mechanism agent i pays a tax equal to his effect on the other agents if he is pivotal to the project decision, and he pays nothing otherwise.[23]

22. We will sometimes be a little loose in our terminology and simply refer to a social choice function $f(\cdot)$ satisfying (23.C.7) and (23.C.8) as a Groves mechanism.

23. Note that the social choice function of the Clarke mechanism satisfies the feasibility condition that $\sum_i t_i(\theta) \leq 0$ for all θ. Indeed, examining (23.C.8), we see that a sufficient (but *not* necessary) condition for a Groves scheme to satisfy the condition that $\sum_i t_i(\theta) \leq 0$ for all θ is that

$$h_i(\theta_{-i}) \leq -\sum_{j \neq i} v_j(k^*_{-i}(\theta_{-i}), \theta_j) \qquad \text{for all } \theta_{-i} \in \Theta_{-i}.$$

It is interesting to note that in the case of allocation of a single indivisible unit of a private good, the Clarke mechanism is precisely the social choice function implemented by the second-price sealed-bid auction (see Example 23.B.6). In particular: (*i*) $k^*(\theta)$ is the allocation rule that gives the item to the agent with the highest valuation; (ii) agent i is pivotal precisely when he is the buyer with the highest valuation; and (iii) when he is pivotal his "tax" is exactly equal to the second-highest valuation (in particular, in this case $\sum_{j \neq i} v_j(k^*(\theta), \theta_j) = 0$, and $\sum_{j \neq i} v_j(k^*_{-i}(\theta_{-i}), \theta_j)$ is equal to the amount of the second-highest valuation).

We have seen so far that social choice functions satisfying (23.C.7) and (23.C.8) are truthfully implementable in dominant strategies. Are these the only social choice functions satisfying (23.C.7) that are truthfully implementable? The result given in Proposition 23.C.5 [due to Green and Laffont (1979)] provides one set of conditions under which the answer is "yes."[24] In it, we let $\mathscr{V}$ denote the set of all possible functions $v\colon K \to \mathbb{R}$.

Proposition 23.C.5: Suppose that for each agent $i = 1, \ldots, I$, $\{v_i(\cdot, \theta_i)\colon \theta_i \in \Theta_i\} = \mathscr{V}$; that is, every possible valuation function from K to $\mathbb{R}$ arises for some $\theta_i \in \Theta_i$. Then a social choice function $f(\cdot) = (k^*(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ in which $k^*(\cdot)$ satisfies (23.C.7) is truthfully implementable in dominant strategies only if $t_i(\cdot)$ satisfies (23.C.8) for all $i = 1, \ldots, I$.

Proof: Note first that we can always write

$$t_i(\theta_i, \theta_{-i}) = \sum_{j \neq i} v_j(k^*(\theta_i, \theta_{-i}), \theta_j) + h_i(\theta_i, \theta_{-i}). \tag{23.C.9}$$

What we want to show, then, is that the function $h_i(\cdot)$ must in fact be independent of θ_i if $f(\cdot)$ is truthfully implementable in dominant strategies. Suppose that it is not; that is, that $f(\cdot)$ is truthfully implementable in dominant strategies but that for some θ_i, $\hat{\theta}_i$, and θ_{-i}, we have $h_i(\theta_i, \theta_{-i}) \neq h_i(\hat{\theta}_i, \theta_{-i})$. We now consider two distinct cases.

(i) $k^*(\theta_i, \theta_{-i}) = k^*(\hat{\theta}_i, \theta_{-i})$: If $f(\cdot)$ is truthfully implementable in dominant strategies, then by (23.C.3) we have

$$v_i(k^*(\theta_i, \theta_{-i}), \theta_i) + t_i(\theta_i, \theta_{-i}) \geq v_i(k^*(\hat{\theta}_i, \theta_{-i}), \theta_i) + t_i(\hat{\theta}_i, \theta_{-i})$$

and

$$v_i(k^*(\hat{\theta}_i, \theta_{-i}), \hat{\theta}_i) + t_i(\hat{\theta}_i, \theta_{-i}) \geq v_i(k^*(\theta_i, \theta_{-i}), \hat{\theta}_i) + t_i(\theta_i, \theta_{-i}).$$

Since, $k^*(\theta_i, \theta_{-i}) = k^*(\hat{\theta}_i, \theta_{-i})$, these two inequalities imply that $t_i(\theta_i, \theta_{-i}) = t_i(\hat{\theta}_i, \theta_{-i})$, and so by (23.C.9) we have $h_i(\theta_i, \theta_{-i}) = h_i(\hat{\theta}_i, \theta_{-i})$: a contradiction.

(ii) $k^*(\theta_i, \theta_{-i}) \neq k^*(\hat{\theta}_i, \theta_{-i})$: Suppose, without loss of generality, that $h_i(\theta_i, \theta_{-i}) > h_i(\hat{\theta}_i, \theta_{-i})$. Consider the type $\theta_i^\varepsilon \in \Theta_i$ such that

$$v_i(k, \theta_i^\varepsilon) = \begin{cases} -\sum_{j \neq i} v_j(k^*(\theta_i, \theta_{-i}), \theta_j) & \text{if } k = k^*(\theta_i, \theta_{-i}) \\ -\sum_{j \neq i} v_j(k^*(\hat{\theta}_i, \theta_{-i}), \theta_j) + \varepsilon & \text{if } k = k^*(\hat{\theta}_i, \theta_{-i}) \\ -\infty & \text{otherwise.} \end{cases} \tag{23.C.10}$$

24. For another, see the small-type discussion at the end of this section and Exercise 23.C.10.

We will argue that for a sufficiently small $\varepsilon > 0$, type θ_i^ε will strictly prefer to falsely report that he is type θ_i when the other agents' types are θ_{-i}. To see this, note first that $k^*(\theta_i^\varepsilon, \theta_{-i}) = k^*(\hat{\theta}_i, \theta_{-i})$ since setting $k = k^*(\hat{\theta}_i, \theta_{-i})$ maximizes $v_i(k, \theta_i^\varepsilon) + \sum_{j \neq i} v_j(k, \theta_j)$. Thus, truth telling being a dominant strategy requires that

$$v_i(k^*(\hat{\theta}_i, \theta_{-i}), \theta_i^\varepsilon) + t_i(\theta_i^\varepsilon, \theta_{-i}) \geq v_i(k^*(\theta_i, \theta_{-i}), \theta_i) + t_i(\theta_i, \theta_{-i}),$$

or, substituting, from (23.C.9) and (23.C.10),

$$\varepsilon + h_i(\theta_i^\varepsilon, \theta_{-i}) \geq h_i(\theta_i, \theta_{-i}).$$

But by the logic of part (i), $h_i(\theta_i^\varepsilon, \theta_{-i}) = h_i(\hat{\theta}_i, \theta_{-i})$ because $k^*(\theta_i^\varepsilon, \theta_{-i}) = k^*(\hat{\theta}_i, \theta_{-i})$. This gives

$$\varepsilon + h_i(\hat{\theta}_i, \theta_{-i}) \geq h_i(\theta_i, \theta_{-i}). \tag{23.C.11}$$

By hypothesis we have $h(\theta_i, \theta_{-i}) > h(\hat{\theta}_i, \theta_{-i})$, and so (23.C.11) must be violated for small enough $\varepsilon > 0$. This completes the proof. ■

Thus, when all possible functions $v_i(\cdot)$ can arise for some $\theta_i \in \Theta_i$, the only social choice functions satisfying (23.C.7) that are truthfully implementable in dominant strategies are those in the Groves class.

Groves mechanisms and budget balance

Up to this point, we have studied whether we can implement in dominant strategies a social choice function that always results in an efficient choice of k [one satisfying (23.C.7)]. But ex post efficiency also requires that none of the numeraire be wasted, that is, that we satisfy the *budget balance condition*:

$$\sum_i t_i(\theta) = 0 \quad \text{for all } \theta \in \Theta. \tag{23.C.12}$$

We now briefly explore when fully ex post efficient social choice functions [those satisfying *both* (23.C.7) *and* (23.C.12)] can be truthfully implemented in dominant strategies.

Unfortunately, in many cases it is impossible to truthfully implement fully ex post efficient social choice functions in dominant strategies. For example, the result [due to Green and Laffont (1979)] in Proposition 23.C.6, whose proof we omit, shows that if the set of possible types for each agent is sufficiently rich, then no social choice functions that are truthfully implementable in dominant strategies are ex post efficient.[25]

Proposition 23.C.6: Suppose that for each agent $i = 1, \ldots, I$, $\{v_i(\cdot, \theta_i): \theta_i \in \Theta_i\} = \mathscr{V}$; that is, every possible valuation function from K to $\mathbb{R}$ arises for some $\theta_i \in \Theta_i$. Then there is no social choice function $f(\cdot) = (k^*(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ that is truthfully implementable in dominant strategies and is ex post efficient, that is, that satisfies (23.C.7) and (23.C.12).

Thus, under the hypotheses of Proposition 23.C.6, the presence of private information means that the I agents must either accept some waste of the numeraire

25. For another negative result, see the small-type discussion at the end of this section and Exercise 23.C.10.

[i.e., have $\sum_i t_i(\theta) < 0$ for some θ, as in the Clarke mechanism] or give up on always having an efficient project selection [i.e., have a project selection $k(\theta)$ that does not satisfy (23.C.7) for some θ].

One special case in which a more positive result does obtain arises when there is at least one agent whose preferences are known. For notational purposes, let this agent be denoted "agent 0", and let there still be I agents, denoted $i = 1, \ldots, I$, whose preferences are private information (so that we are now letting there be $I + 1$ agents in total). The simplest case of this phenomenon, of course, occurs when agent 0 has no preferences over the project choice k, that is, when his preferences are $u_i(x) = \bar{m}_0 + t_0$. We saw one example of this kind in Example 23.B.4 when we considered auction settings (agent 0 is then the seller). Another example arises in the case of a public project when the project affects only a subset of the agents in the economy (so that agent 0 represents all of the other agents in the economy).

When there is such an agent, ex post efficiency of the social choice function still requires that (23.C.7) be satisfied; but now ex post efficiency is compatible with *any* transfer functions $t_1(\cdot), \ldots, t_I(\cdot)$ for the I agents with private information, as long as we set $t_0(\theta) = -\sum_{i \neq 0} t_i(\theta)$ for all θ. That is, in this $(I + 1)$-agent setting, the Groves mechanisms identified in Proposition 23.C.4 (in which only agents $i = 1, \ldots, I$ announce their types) are ex post efficient as long as we set the transfer of agent 0 to be $t_0(\theta) = -\sum_{i \neq 0} t_i(\theta)$ for all θ. In essence, the presence of an "outside" agent 0 who has no private information allows us to break the budget balance condition for those agents who do have privately observed types.

We should offer, however, one immediate caveat to this seemingly positive result: Up to this point, we have not worried about whether agents will find it in their interest to participate in the mechanism. As we will see in Section 23.E, when participation is voluntary, it may be that no ex post efficient social choice function is implementable in dominant strategies even when such an outside agent exists.

The differentiable case

It is common in applications to encounter cases in which $K = \mathbb{R}$, the $v_i(\cdot, \theta_i)$ functions are assumed to be twice continuously differentiable with $\partial^2 v_i(k, \theta_i)/\partial k^2 < 0$ and $\partial^2 v_i(k, \theta_i)/\partial k \, \partial \theta_i \neq 0$ at all (k, θ_i), and each θ_i is drawn from an interval $[\underline{\theta}_i, \bar{\theta}_i] \subset \mathbb{R}$ with $\underline{\theta}_i \neq \bar{\theta}_i$. In this case a great deal can be said about the set of social choice functions that can be truthfully implemented in dominant strategies. Exercise 23.C.9 develops this point fully. Here we will simply show how, in this environment, we can easily derive a number of our previous results.

Note first that for any continuously differentiable social choice function $f(\cdot) = (k(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$, if truth telling is a dominant strategy for agent i, then agent i's first-order condition implies that, for all θ_{-i},

$$\frac{\partial v_i(k(\theta_i, \theta_{-i}), \theta_i)}{\partial k} \frac{\partial k(\theta_i, \theta_{-i})}{\partial \theta_i} + \frac{\partial t_i(\theta_i, \theta_{-i})}{\partial \theta_i} = 0 \tag{23.C.13}$$

at all $\theta_i \in (\underline{\theta}_i, \bar{\theta}_i)$. Integrating (23.C.13) with respect to the variable θ_i, this implies that for all profiles of types (θ_i, θ_{-i}) we have

$$t_i(\theta_i, \theta_{-i}) = t_i(\underline{\theta}_i, \theta_{-i}) - \int_{\underline{\theta}_i}^{\theta_i} \frac{\partial v_i(k(s, \theta_{-i}), s)}{\partial k} \frac{\partial k(s, \theta_{-i})}{\partial s} \, ds. \tag{23.C.14}$$

Consider now any social choice function $f(\cdot) = (k^*(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ that satisfies

(23.C.7). Under our present assumptions $k^*(\cdot)$ must satisfy, for all θ,

$$\sum_{j=1}^{I} \frac{\partial v_j(k^*(\theta), \theta_j)}{\partial k} = 0. \tag{23.C.15}$$

Moreover, using the implicit function theorem and our assumptions on the $v_i(\cdot)$ functions, we see that $k^*(\cdot)$ is continuously differentiable and that it has nonzero partial derivatives, $\partial k^*(\theta)/\partial \theta_i \neq 0$ for all i.

We now substitute for $\partial v_i(k^*(s, \theta_{-i}), s)/\partial k$ in (23.C.14) using (23.C.15). Doing so, we derive that, for all profiles (θ_i, θ_{-i}),

$$\begin{aligned} t_i(\theta_i, \theta_{-i}) &= t_i(\underline{\theta}_i, \theta_{-i}) + \int_{\underline{\theta}_i}^{\theta_i} \left(\sum_{j \neq i} \frac{\partial v_j(k^*(s, \theta_{-i}), \theta_j)}{\partial k} \right) \frac{\partial k(s, \theta_{-i})}{\partial s} \, ds \\ &= t_i(\underline{\theta}_i, \theta_{-i}) + \int_{k^*(\underline{\theta}_i, \theta_{-i})}^{k^*(\theta_i, \theta_{-i})} \left(\sum_{j \neq i} \frac{\partial v_j(k, \theta_j)}{\partial k} \right) dk \\ &= t_i(\underline{\theta}_i, \theta_{-i}) + \sum_{j \neq i} v_j(k^*(\theta_i, \theta_{-i}), \theta_j) - \sum_{j \neq i} v_j(k^*(\underline{\theta}_i, \theta_{-i}), \theta_j). \end{aligned}$$

But this is true if and only if $t_i(\theta)$ satisfies (23.C.8). Thus, in this setting, Groves mechanisms are the only social choice functions satisfying (23.C.7) that are truthfully implementable in dominant strategies.[26]

Consider now the question of budget balance when there is no outside agent. We will show that satisfying (23.C.15) and budget balance is impossible in this differentiable setting when $I = 2$ [for $I > 2$ see Laffont and Maskin (1980) and Exercise 23.C.10]. By (23.C.13), for all $\theta = (\theta_1, \theta_2)$, we have

$$\frac{\partial t_1(\theta)}{\partial \theta_1} = -\frac{\partial v_1(k^*(\theta), \theta_1)}{\partial k} \frac{\partial k^*(\theta)}{\partial \theta_1}$$

and

$$\frac{\partial t_2(\theta)}{\partial \theta_2} = -\frac{\partial v_2(k^*(\theta), \theta_2)}{\partial k} \frac{\partial k^*(\theta)}{\partial \theta_2}.$$

Thus, for all $\theta = (\theta_1, \theta_2)$,

$$-\frac{\partial^2 t_1(\theta)}{\partial \theta_1 \, \partial \theta_2} = \frac{\partial^2 v_1(k^*(\theta), \theta_1)}{\partial k^2} \frac{\partial k^*(\theta)}{\partial \theta_2} \frac{\partial k^*(\theta)}{\partial \theta_1} + \frac{\partial v_1(k^*(\theta), \theta_1)}{\partial k} \frac{\partial^2 k^*(\theta)}{\partial \theta_1 \, \partial \theta_2} \tag{23.C.16}$$

and

$$-\frac{\partial^2 t_2(\theta)}{\partial \theta_1 \, \partial \theta_2} = \frac{\partial^2 v_2(k^*(\theta), \theta_2)}{\partial k^2} \frac{\partial k^*(\theta)}{\partial \theta_1} \frac{\partial k^*(\theta)}{\partial \theta_2} + \frac{\partial v_2(k^*(\theta), \theta_2)}{\partial k} \frac{\partial^2 k^*(\theta)}{\partial \theta_1 \, \partial \theta_2}, \tag{23.C.17}$$

If we have budget balance, then $t_1(\theta) = -t_2(\theta)$ for all θ, and so we must have $\partial^2 t_1(\theta)/\partial \theta_1 \partial \theta_2 = -\partial^2 t_2(\theta)/\partial \theta_1 \partial \theta_2$. But this would imply, by adding (23.C.16) and (23.C.17), and using (23.C.15), that

$$\left[\frac{\partial^2 v_1(k^*(\theta), \theta_1)}{\partial k^2} + \frac{\partial^2 v_2(k^*(\theta), \theta_2)}{\partial k^2} \right] \frac{\partial k^*(\theta)}{\partial \theta_1} \frac{\partial k^*(\theta)}{\partial \theta_2} = 0,$$

which is impossible under our assumptions.

---

26. This argument generalizes to any case in which $k^*(\cdot)$ is continuously differentiable.

23.D Bayesian Implementation

In this section, we study implementation in *Bayesian Nash equilibrium.*[27] Throughout we follow the notation introduced in Section 23.B: The vector of agents' types $\theta = (\theta_1, \ldots, \theta_I)$ is drawn from set $\Theta = \Theta_1 \times \cdots \times \Theta_I$ according to probability density $\phi(\cdot)$, and agent i's Bernoulli utility function over the alternatives in X given his type θ_i is $u_i(x, \theta_i)$. We also adopt the notational convention of writing $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_I)$ and $\theta = (\theta_i, \theta_{-i})$. A mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ is a collection of I sets $S_1, \ldots, S_I$, each S_i containing agent i's possible actions (or plans of action), and an outcome function $g: S \to X$, where $S = S_1 \times \cdots \times S_I$. As discussed in Section 23.B, the mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ combined with possible types $(\Theta_1, \ldots, \Theta_I)$, density $\phi(\cdot)$, and Bernoulli utility functions $(u_1(\cdot), \ldots, u_I(\cdot))$ defines a Bayesian game of incomplete information (see Section 8.E). We will also often write $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_I)$, $s = (s_i, s_{-i})$, and $s(\cdot) = (s_i(\cdot), s_{-i}(\cdot))$ where $s_{-i}(\cdot) = (s_1(\cdot), \ldots, s_{i-1}(\cdot), s_{i+1}(\cdot), \ldots, s_I(\cdot))$.

We begin by defining the concept of a Bayesian Nash equilibrium (see also Section 8.E) and specializing Definition 23.B.4 to the notion of implementation in Bayesian Nash equilibrium.[28]

Definition 23.D.1: The strategy profile $s^*(\cdot) = (s_1^*(\cdot), \ldots, s_I^*(\cdot))$ is a *Bayesian Nash equilibrium* of mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ if, for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i) | \theta_i] \geq E_{\theta_{-i}}[u_i(g(\hat{s}_i, s_{-i}^*(\theta_{-i})), \theta_i) | \theta_i]$$

for all $\hat{s}_i \in S_i$.

Definition 23.D.2: The mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ *implements the social choice function* $f(\cdot)$ *in Bayesian Nash equilibrium* if there is a Bayesian Nash equilibrium of Γ, $s^*(\cdot) = (s_1^*(\cdot), \ldots, s_I^*(\cdot))$, such that $g(s^*(\theta)) = f(\theta)$ for all $\theta \in \Theta$.

As with implementation in dominant strategies (see Section 23.C), we will see that a social choice function is Bayesian implementable if and only if it is truthfully implementable in the sense given in Definition 23.D.3.

Definition 23.D.3: The social choice function $f(\cdot)$ is *truthfully implementable in Bayesian Nash equilibrium* (or *Bayesian incentive compatible*) if $s_i^*(\theta_i) = \theta_i$ for all $\theta_i \in \Theta_i$ and $i = 1, \ldots, I$ is a Bayesian Nash equilibrium of the direct revelation mechanism $\Gamma = (\Theta_1, \ldots, \Theta_I, f(\cdot))$. That is, if for all $i = 1, \ldots, I$ and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i) | \theta_i] \geq E_{\theta_{-i}}[u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i) | \theta_i] \qquad (23.D.1)$$

for all $\hat{\theta}_i \in \Theta_i$.

The ability to restrict our inquiry, without loss of generality, to the question of whether $f(\cdot)$ is truthfully implementable is a consequence of the *revelation principle for Bayesian Nash equilibrium.*

27. Good sources for further reading on the subject of this section are Myerson (1991) and Fudenberg and Tirole (1991).

28. As in Section 8.E, we restrict our attention to pure strategy equilibria.

Proposition 23.D.1: (*The Revelation Principle for Bayesian Nash Equilibrium*) Suppose that there exists a mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ that implements the social choice function $f(\cdot)$ in Bayesian Nash equilibrium. Then $f(\cdot)$ is truthfully implementable in Bayesian Nash equilibrium.

Proof: If $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ implements $f(\cdot)$ in Bayesian Nash equilibrium, then there exists a profile of strategies $s^*(\cdot) = (s_1^*(\cdot), \ldots, s_I^*(\cdot))$ such that $g(s^*(\theta)) = f(\theta)$ for all θ, and for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(\hat{s}_i, s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \quad (23.D.2)$$

for all $\hat{s}_i \in S_i$. Condition (23.D.2) implies, in particular, that for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(g(s_i^*(\theta_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(g(s_i^*(\hat{\theta}_i), s_{-i}^*(\theta_{-i})), \theta_i)|\theta_i] \quad (23.D.3)$$

for all $\hat{\theta}_i \in \Theta_i$. Since $g(s^*(\theta)) = f(\theta)$ for all θ, (23.D.3) means that, for all i and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i] \geq E_{\theta_{-i}}[u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i)|\theta_i] \quad (23.D.4)$$

for all $\hat{\theta}_i \in \Theta_i$. But, this is precisely condition (23.D.1), the condition for $f(\cdot)$ to be truthfully implementable in Bayesian Nash equilibrium. ■

The basic idea behind the revelation principle for Bayesian Nash equilibrium parallels that given for the revelation principle for dominant strategy implementation (Proposition 23.C.1): If in mechanism $\Gamma = (S_1, \ldots, S_I, g(\cdot))$ each agent finds that, when his type is θ_i, choosing $s_i^*(\theta_i)$ is his best response to the other agents' strategies, then if we introduce a mediator who says "Tell me your type, θ_i, and I will play $s_i^*(\theta_i)$ for you," each agent will find truth telling to be an optimal strategy given that all other agents tell the truth. That is, truth telling will be a Bayesian Nash equilibrium of this direct revelation game.

The implication of the revelation principle is, once again, that to identify the set of implementable social choice functions (now in Bayesian Nash equilibrium) we need only identify those that are truthfully implementable.[29]

We can note immediately that the Bayesian implementation concept is a strictly weaker notion than the notion of dominant strategy implementation. Since every dominant strategy equilibrium is necessarily a Bayesian Nash equilibrium, any social choice function that is implementable in dominant strategies is a fortiori implementable in Bayesian Nash equilibrium. Intuitively put, when we compare the requirements for truthful implementation of a social choice function $f(\cdot)$ in dominant strategies and in Bayesian Nash equilibrium given in equations (23.C.3) and (23.D.1), respectively, we see that, with Bayesian implementation, truth telling need only give agent i his highest payoff *averaging over all possible types* θ_{-i} *that might arise for the other agents.* In contrast, the dominant strategy concept requires that truth telling be agent i's best strategy *for every possible* θ_{-i}. Thus, we can reasonably hope to be

29. Note that Proposition 23.D.1 is what we implicitly relied on in Section 14.C when, in studying the principal agent problem with hidden information, we restricted our focus to direct revelation mechanisms that induced truth telling by the agent. Formally, Proposition 23.D.1 tells us that the equilibrium outcome arising from any contract between the principal and the agent can be replicated using a direct revelation mechanism that induces the agent to truthfully reveal his type.

able to successfully implement a wider range of social choice functions in Bayesian Nash equilibrium than in dominant strategies. The drawback, of course, is that we can be less confident about this implementation relative to implementation in dominant strategies because it depends on the agents (and any outside mechanism designer) knowing the density $\phi(\cdot)$ of agents' types, as well as on the plausibility of the Nash assumption that the agents' have mutually correct expectations about each others' strategy choices (see Section 8.D).

In what follows in the remainder of this section, we first provide an example that illustrates that we can indeed implement a wider range of social choice functions with Bayesian implementation. Specifically, we show that in the quasilinear environment studied in Section 23.C, whenever the agents' types are statistically independent of one another, we can *always* Bayesian implement at least one ex post efficient social choice function (i.e., one that both has an efficient project choice and satisfies budget balance). In Section 23.C we saw that it may not be possible to accomplish this with dominant strategy implementation (recall Proposition 23.C.6 and the small-type discussion at the end of that section). After showing this, we then examine the properties of Bayesian implementable social choice functions in greater detail, concentrating on the special case in which agents have quasilinear utility functions that are linear in their type. As an application of this analysis, we prove the *revenue equivalence theorem* for auctions.

The Expected Externality Mechanism

Let us return to the quasilinear setting studied in Section 23.C. In particular, an alternative is now a vector $x = (k, t_1, \ldots, t_I)$, where k is an element of a finite set K, and $t_i \in \mathbb{R}$ is a transfer of a numeraire commodity ("money") to agent i. Agent i's utility function takes the quasilinear form

$$u_i(x, \theta_i) = v_i(k, \theta_i) + (\bar{m}_i + t_i), \tag{23.D.5}$$

where $\bar{m}_i$ is agent i's endowment of the numeraire.[30] For simplicity, we shall henceforth normalize $\bar{m}_i = 0$ for all i. We assume here that the I agents have no outside source of financing, and so $X = \{(k, t_1, \ldots, t_I): k \in K, t_i \in \mathbb{R} \text{ for all } i, \text{ and } \sum_i t_i \leq 0\}$. A social choice function in this environment takes the form $f(\cdot) = (k(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$. Note that if $f(\cdot)$ is ex post efficient then, for all $\theta \in \Theta$,

$$\sum_i v_i(k(\theta), \theta_i) \geq \sum_i v_i(k, \theta_i) \qquad \text{for all } k \in K \tag{23.D.6}$$

and

$$\sum_i t_i(\theta) = 0. \tag{23.D.7}$$

In Proposition 23.C.6 we saw that conditions exist in which no social choice

30. Unlike the analysis in Section 23.C (see Exercise 23.C.11), the developments that follow depend not only on preferences over certain outcomes having a quasilinear form but also on the fact that, with this Bernoulli utility function, each agent i is risk neutral with respect to lotteries over his monetary transfer.

function $f(\cdot) = (k(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ satisfying both (23.D.6) and (23.D.7) is truthfully implementable in dominant strategies. We will now show that it *is* possible to implement such a social choice function in Bayesian Nash equilibrium whenever the agents' types are statistically independent of one another [i.e., when the density $\phi(\cdot)$ has the form $\phi(\theta) = \phi_1(\theta_1) \times \cdots \times \phi_I(\theta_I)$].[31]

To verify this, let $k^*(\cdot)$ satisfy (23.D.6) and consider a social choice function $f(\cdot) = (k^*(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ in which, for all $i = 1, \ldots, I$,

$$t_i(\theta) = E_{\tilde{\theta}_{-i}}\left[\sum_{j \neq i} v_j(k^*(\theta_i, \tilde{\theta}_{-i}), \tilde{\theta}_j)\right] + h_i(\theta_{-i}), \tag{23.D.8}$$

where, for now, we take $h_i(\cdot)$ to be an arbitrary function of θ_{-i}. Note that the expectational term in (23.D.8) represents the *expected benefits* of agents $j \neq i$ when agent i announces his type to be θ_i and agents $j \neq i$ tell the truth. (As such, it is a function of only agent i's actual announcement θ_i—it is *not* a function of the actual announcements θ_{-i} of agents $j \neq i$.) Thus, the change in agent i's transfer when he changes his announced type is exactly equal to the *expected externality* of this change on agents $j \neq i$.

We check first that any social choice function $f(\cdot)$ with the form (23.D.8) is Bayesian incentive compatible. To see this, note that when agents $j \neq i$ announce their types truthfully, agent i finds that truth telling is his optimal strategy because (using statistical independence of θ_i and θ_{-i})

$$\begin{aligned} E_{\theta_{-i}}[v_i(k^*(\theta), \theta_i) + t_i(\theta)|\theta_i] &= E_{\theta_{-i}}\left[\sum_{j=1}^{I} v_j(k^*(\theta), \theta_j)\right] + E_{\theta_{-i}}[h_i(\theta_{-i})] \\ &\geq E_{\theta_{-i}}\left[\sum_{j=1}^{I} v_j(k^*(\hat{\theta}_i, \theta_{-i}), \theta_j)\right] + E_{\theta_{-i}}[h_i(\theta_{-i})] \\ &= E_{\theta_{-i}}[v_i(k^*(\hat{\theta}_i, \theta_{-i}), \theta_j) + t_i(\hat{\theta}_i, \theta_{-i})|\theta_i] \end{aligned}$$

for all $\hat{\theta}_i \in \Theta_i$, where the inequality follows because $k^*(\cdot)$ satisfies (23.D.6).

What remains is to show that we can choose the $h_i(\cdot)$ functions (for $i = 1, \ldots, I$) so that we also satisfy the budget balance condition (23.D.7). For notational ease, define $\xi_i(\theta_i) = E_{\tilde{\theta}_{-i}}[\sum_{j \neq i} v_j(k^*(\theta_i, \tilde{\theta}_{-i}), \tilde{\theta}_j)]$. We now let

$$h_i(\theta_{-i}) = -\left(\frac{1}{I-1}\right)\sum_{j \neq i} \xi_j(\theta_j), \tag{23.D.9}$$

for $i = 1, \ldots, I$. With this choice for the $h_i(\cdot)$ functions, we have

$$\begin{aligned} \sum_i t_i(\theta) &= \sum_i \xi_i(\theta_i) + \sum_i h_i(\theta_{-i}) \\ &= \sum_i \xi_i(\theta_i) - \left(\frac{1}{I-1}\right)\sum_i \sum_{j \neq i} \xi_j(\theta_j) \\ &= \sum_i \xi_i(\theta_i) - \left(\frac{1}{I-1}\right)\sum_i (I-1)\xi_i(\theta_i) \\ &= 0. \end{aligned}$$

31. See Fudenberg and Tirole (1991) for a discussion of the case of correlated types and for further references.

Intuitively, the form of the $h_i(\cdot)$ functions in (23.D.9) can be thought of as follows: We have seen that when the agents' types are $(\theta_1, \ldots, \theta_I)$, each agent $i = 1, \ldots, I$ receives a payment equal to $\xi_i(\theta_i)$ [the first term in (23.D.8)]. Now, if each agent contributes an equal $1/(I-1)$ share of all of the other agents' payments, the payments from a given agent i to each of the other $I-1$ agents will total $[1/(I-1)]\sum_{j\neq i}\xi_j(\theta_j)$, and agent i will receive from these agents in return payments that total to $\xi_i(\theta_i)$. Agent i's net transfer will therefore be $\xi_i(\theta_i) - (1/(I-1))\sum_{j\neq i}\xi_j(\theta_j)$.

This direct revelation mechanism is known as the *expected externality mechanism* [due to d'Aspremont and Gérard-Varet (1979) and Arrow (1979)]. In summary, we have shown that when agents' Bernoulli utility functions take the form in (23.D.5), and agents' types are statistically independent, there is an ex post efficient social choice function that is implementable in Bayesian Nash equilibrium.

Although this is an interesting result, it is not the end of the story, even when we restrict our attention to Bernoulli utility functions of the form (23.D.5) and a statistically independent distribution of types. The reason is that while the expected externality mechanism implements an ex post efficient social choice function, its transfer functions imply a particular distribution of utility across the various types of the agents, and we may wish to consider other mechanisms, possibly ones involving social choice functions that are not ex post efficient, that alter this distribution.

One reason why this may be important is that, in many applications of interest, agents are free to opt out of the mechanism, and so any mechanism that we wish to implement must not only be incentive compatible in the sense that we have studied so far, but must also satisfy *individual rationality* (or *participation*) *constraints* that assure that each agent i actually wishes to participate in the mechanism. If the expected externality mechanism does not satisfy these constraints, we will need to consider other mechanisms that do. We will have more to say about this issue in Sections 23.E and 23.F, but for now suffice it to say that for this reason, as well as others, we may be interested in identifying *all* of the social choice functions that are Bayesian implementable in this environment.

In the remainder of this section we do this for the special, but often-studied, class of cases in which agents' preferences take a form that is linear in their type, and their types are independently distributed.

Bayesian Incentive Compatibility with Linear Utility

Suppose now that each agent i's Bernoulli utility function takes the form

$$u_i(x, \theta_i) = \theta_i v_i(k) + (\bar{m}_i + t_i).$$

As before, we shall normalize $\bar{m}_i = 0$ for all i. We also suppose that each agent i's type lies in an interval $\Theta_i = [\underline{\theta}_i, \bar{\theta}_i] \subset \mathbb{R}$ with $\underline{\theta}_i \neq \bar{\theta}_i$, and that the agents' types are statistically independent. We let the distribution function of θ_i be denoted $\Phi_i(\cdot)$, and we assume that it has an associated density $\phi_i(\cdot)$ satisfying $\phi_i(\theta_i) > 0$ for all $\theta_i \in [\underline{\theta}_i, \bar{\theta}_i]$.

We begin by deriving a necessary and sufficient condition for a social choice function $f(\cdot) = (k(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ to be Bayesian incentive compatible. It is convenient to define $\bar{t}_i(\hat{\theta}_i) = E_{\theta_{-i}}[t_i(\hat{\theta}_i, \theta_{-i})]$; this is agent i's expected transfer given that he announces his type to be $\hat{\theta}_i$ and that all agents $j \neq i$ truthfully reveal their types. Likewise, we let $\bar{v}_i(\hat{\theta}_i) = E_{\theta_{-i}}[v_i(k(\hat{\theta}_i, \theta_{-i}))]$ denote agent i's expected "benefit"

conditional on announcing $\hat{\theta}_i$. Because of the form of agents' utility functions, we can write agent i's expected utility when he is type θ_i and announces his type to be $\hat{\theta}_i$ (assuming that all agents $j \neq i$ tell the truth) as[32]

$$E_{\theta_{-i}}[u_i(f(\hat{\theta}_i, \theta_{-i}), \theta_i)|\theta_i] = \theta_i \bar{v}_i(\hat{\theta}_i) + \bar{t}_i(\hat{\theta}_i). \tag{23.D.10}$$

It is also convenient to define for each i the function

$$U_i(\theta_i) = \theta_i \bar{v}_i(\theta_i) + \bar{t}_i(\theta_i),$$

giving agent i's expected utility from the mechanism conditional on his type being θ_i when he and all other agents report their true types.

Proposition 23.D.2: The social choice function $f(\cdot) = (k(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ is Bayesian incentive compatible if and only if, for all $i = 1, \ldots, I$,

(i) $\bar{v}_i(\cdot)$ is nondecreasing. (23.D.11)

(ii) $U_i(\theta_i) = U_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} \bar{v}_i(s)\, ds$ for all θ_i. (23.D.12)

Proof: (i) *Necessity.* Bayesian incentive compatibility implies that for each $\hat{\theta}_i > \theta_i$ we have

$$U_i(\theta_i) \geq \theta_i \bar{v}_i(\hat{\theta}_i) + \bar{t}_i(\hat{\theta}_i) = U_i(\hat{\theta}_i) + (\theta_i - \hat{\theta}_i)\bar{v}_i(\hat{\theta}_i)$$

and

$$U_i(\hat{\theta}_i) \geq \hat{\theta}_i \bar{v}_i(\theta_i) + \bar{t}_i(\theta_i) = U_i(\theta_i) + (\hat{\theta}_i - \theta_i)\bar{v}_i(\theta_i).$$

Thus,

$$\bar{v}_i(\hat{\theta}_i) \geq \frac{U_i(\hat{\theta}_i) - U_i(\theta_i)}{\hat{\theta}_i - \theta_i} \geq \bar{v}_i(\theta_i). \tag{23.D.13}$$

Expression (23.D.13) immediately implies that $\bar{v}_i(\cdot)$ must be nondecreasing (recall that we have taken $\hat{\theta}_i > \theta_i$). In addition, letting $\hat{\theta}_i \to \theta_i$ in (23.D.13) implies that for all θ_i we have

$$U_i'(\theta_i) = \bar{v}_i(\theta_i)$$

and so

$$U_i(\theta_i) = U_i(\underline{\theta}_i) + \int_{\underline{\theta}_i}^{\theta_i} \bar{v}_i(s)\, ds \qquad \text{for all } \theta_i.$$

(ii) *Sufficiency.* Consider any θ_i and $\hat{\theta}_i$ and suppose without loss of generality that $\theta_i > \hat{\theta}_i$. If (23.D.11) and (23.D.12) hold, then

$$\begin{aligned} U_i(\theta_i) - U_i(\hat{\theta}_i) &= \int_{\hat{\theta}_i}^{\theta_i} \bar{v}_i(s)\, ds \\ &\geq \int_{\hat{\theta}_i}^{\theta_i} \bar{v}_i(\hat{\theta}_i)\, ds \\ &= (\theta_i - \hat{\theta}_i)\bar{v}_i(\hat{\theta}_i). \end{aligned}$$

32. Observe that the agent's preferences here over his expected benefit $\bar{v}_i$ and expected transfer $\bar{t}_i$ satisfy the single-crossing property that played a prominent role in Sections 13.C and 14.C.

Hence,

$$U_i(\theta_i) \geq U_i(\hat{\theta}_i) + (\theta_i - \hat{\theta}_i)\bar{v}_i(\hat{\theta}_i) = \theta_i \bar{v}_i(\hat{\theta}_i) + \bar{t}_i(\hat{\theta}_i).$$

Similarly, we can derive that

$$U_i(\hat{\theta}_i) \geq U_i(\theta_i) + (\hat{\theta}_i - \theta_i)\bar{v}_i(\theta_i) = \hat{\theta}_i \bar{v}_i(\theta_i) + \bar{t}_i(\theta_i).$$

So $f(\cdot)$ is Bayesian incentive compatible. ■

Proposition 23.D.2 shows that to identify all Bayesian incentive compatible social choice functions in the linear setting, we can proceed as follows: First identify which functions $k(\cdot)$ lead every agent i's expected benefit function $\bar{v}_i(\cdot)$ to be nondecreasing. Then, for each such function, identify the expected transfer functions $\bar{t}_1(\cdot), \ldots, \bar{t}_I(\cdot)$ that satisfy condition (23.D.12) of the proposition. Substituting for $U_i(\cdot)$, these are precisely the expected transfer functions that satisfy, for $i = 1, \ldots, I$,

$$\bar{t}_i(\theta_i) = \bar{t}_i(\underline{\theta}_i) + \underline{\theta}_i v_i(\underline{\theta}_i) - \theta_i v_i(\theta_i) + \int_{\underline{\theta}_i}^{\theta_i} \bar{v}_i(s)\, ds$$

for some constant $\bar{t}_i(\underline{\theta}_i)$. Finally, choose any set of transfer functions $(t_1(\theta), \ldots, t_I(\theta))$ such that $E_{\theta_{-i}}[t_i(\theta_i, \theta_{-i})] = \bar{t}_i(\theta_i)$ for all θ_i. In general, there are many such functions $t_i(\cdot, \cdot)$; one, for example, is simply $t_i(\theta_i, \theta_{-i}) = \bar{t}_i(\theta_i)$.[33]

We now illustrate one implication of this characterization result for the auction setting introduced in Example 23.B.4. Some further implications of Proposition 23.D.2 are derived in Sections 23.E and 23.F.

Auctions: the revenue equivalence theorem

Let us consider again the auction setting introduced in Example 23.B.4: Agent 0 is the seller of an indivisible object from which he derives no value, and agents $1, \ldots, I$ are potential buyers.[34] It will be convenient, however, to generalize the set of possible alternatives relative to those considered in Example 23.B.4 by allowing for a *random* assignment of the object. Thus, we now take $y_i(\theta)$ to be buyer i's *probability* of getting the object when the vector of announced types is $\theta = (\theta_1, \ldots, \theta_I)$. Buyer i's expected utility when the profile of types for the I buyers is $\theta = (\theta_1, \ldots, \theta_I)$ is then $\theta_i y_i(\theta) + t_i(\theta)$. Note that buyer i is risk neutral with respect to lotteries both over transfers and over the allocation of the good.

This setting corresponds in the framework studied in Proposition 23.D.2 to the case where we take $k = (y_1, \ldots, y_I)$, $K = \{(y_1, \ldots, y_I): y_i \in [0, 1]$ for all $i = 1, \ldots, I$ and $\sum_i y_i \leq 1\}$, and $v_i(k) = y_i$. Thus, to apply Proposition 23.D.2 we can write $\bar{v}_i(\hat{\theta}_i) = \bar{y}_i(\hat{\theta}_i)$, where $\bar{y}_i(\hat{\theta}_i) = E_{\theta_{-i}}[y_i(\hat{\theta}_i, \theta_{-i})]$ is the probability that i gets the object conditional on announcing his type to be $\hat{\theta}_i$ when agents $j \neq i$ announce their types truthfully, and $U_i(\theta_i) = \theta_i \bar{y}_i(\theta_i) + \bar{t}_i(\theta_i)$.

---

33. However, if we wish the social choice function $f(\cdot) = (k(\cdot), t_1(\cdot), \ldots, t_I(\cdot))$ to satisfy some further properties, such as budget balance, only a subset (possibly an empty one) of the transfer functions generating the expected transfer functions $(\bar{t}_1(\theta_1), \ldots, \bar{t}_I(\theta_I))$ may have these properties.

34. We note that our assumption that the seller in an auction setting derives no value from the object is not necessary for the revenue equivalence theorem. (As we shall see, the result characterizes the expected revenues generated for the seller in different auctions, and so is valid for any utility function that the seller might have.) In the absence of this assumption, however, the seller in an auction will generally care about more than just the expected revenue he receives.

We can now establish a remarkable result, known as the *revenue equivalence theorem.*[35]

Proposition 23.D.3: (*The Revenue Equivalence Theorem*) Consider an auction setting with I risk-neutral buyers, in which buyer i's valuation is drawn from an interval $[\underline{\theta}_i, \bar{\theta}_i]$ with $\underline{\theta}_i \neq \bar{\theta}_i$ and a strictly positive density $\phi_i(\cdot) > 0$, and in which buyers' types are statistically independent. Suppose that a given pair of Bayesian Nash equilibria of two different auction procedures are such that for every buyer i: (i) For each possible realization of $(\theta_1, \ldots, \theta_I)$, buyer i has an identical probability of getting the good in the two auctions; and (ii) Buyer i has the same expected utility level in the two auctions when his valuation for the object is at its lowest possible level. Then these equilibria of the two auctions generate the same expected revenue for the seller.

Proof: By the revelation principle, we know that the social choice function that is (indirectly) implemented by the equilibrium of any auction procedure must be Bayesian incentive compatible. Thus, we can establish the result by showing that if two Bayesian incentive compatible social choice functions in this auction setting have the same functions $(y_1(\theta), \ldots, y_I(\theta))$ and the same values of $(U_1(\underline{\theta}_1), \ldots, U_I(\underline{\theta}_I))$ then they generate the same expected revenue for the seller.

To show this, we derive an expression for the seller's expected revenue from an arbitrary Bayesian incentive compatible mechanism. Note, first, that the seller's expected revenue is equal to $\sum_{i=1}^{I} E[-t_i(\theta)]$. Now,

$$\begin{aligned} E[-t_i(\theta)] &= E_{\theta_i}[-\bar{t}_i(\theta_i)] \\ &= \int_{\underline{\theta}_i}^{\bar{\theta}_i} [\bar{y}_i(\theta_i)\theta_i - U_i(\theta_i)]\phi_i(\theta_i)\, d\theta_i \\ &= \int_{\underline{\theta}_i}^{\bar{\theta}_i} \left(\bar{y}_i(\theta_i)\theta_i - U_i(\underline{\theta}_i) - \int_{\underline{\theta}_i}^{\theta_i} \bar{y}_i(s)\, ds \right) \phi_i(\theta_i)\, d\theta_i \\ &= \left[\int_{\underline{\theta}_i}^{\bar{\theta}_i} \left(\bar{y}_i(\theta_i)\theta_i - \int_{\underline{\theta}_i}^{\theta_i} \bar{y}_i(s)\, ds \right) \phi_i(\theta_i)\, d\theta_i \right] - U_i(\underline{\theta}_i). \end{aligned}$$

Moreover, integration by parts implies that

$$\begin{aligned} \int_{\underline{\theta}_i}^{\bar{\theta}_i} \left(\int_{\underline{\theta}_i}^{\theta_i} \bar{y}_i(s)\, ds \right) \phi_i(\theta_i)\, d\theta_i &= \left(\int_{\underline{\theta}_i}^{\bar{\theta}_i} \bar{y}_i(\theta_i)\, d\theta_i \right) - \left(\int_{\underline{\theta}_i}^{\bar{\theta}_i} \bar{y}_i(\theta_i)\Phi_i(\theta_i)\, d\theta_i \right) \\ &= \int_{\underline{\theta}_i}^{\bar{\theta}_i} \bar{y}_i(\theta_i)(1 - \Phi_i(\theta_i))\, d\theta_i. \end{aligned}$$

Substituting, we see that

$$E[-\bar{t}_i(\theta_i)] = \left[\int_{\underline{\theta}_i}^{\bar{\theta}_i} \bar{y}_i(\theta_i) \left(\theta_i - \frac{1 - \Phi_i(\theta_i)}{\phi_i(\theta_i)} \right) \phi_i(\theta_i)\, d\theta_i \right] - U_i(\underline{\theta}_i), \qquad (23.D.14)$$

35. Versions of the revenue equivalence theorem have been derived by many authors; see McAfee and McMillan (1987) and Milgrom (1987) for references as well as for a further discussion of the result.

or, equivalently,

$$E[-\bar{t}_i(\theta_i)] = \left[\int_{\theta_1}^{\bar{\theta}_1} \cdots \int_{\theta_I}^{\bar{\theta}_I} y_i(\theta_1, \ldots, \theta_I)\left(\theta_i - \frac{1-\Phi_i(\theta_i)}{\phi_i(\theta_i)}\right)\left(\prod_{j=1}^{I} \phi_j(\theta_j)\right) d\theta_I \cdots d\theta_1\right] - U_i(\underline{\theta}_i). \tag{23.D.15}$$

Thus, the seller's expected revenue is equal to

$$\left[\int_{\theta_1}^{\bar{\theta}_1} \cdots \int_{\theta_I}^{\bar{\theta}_I}\left[\sum_{i=1}^{I} y_i(\theta_1, \ldots, \theta_I)\left(\theta_i - \frac{1-\Phi_i(\theta_i)}{\phi_i(\theta_i)}\right)\right]\left(\prod_{j=1}^{I} \phi_j(\theta_j)\right) d\theta_I \cdots d\theta_1\right] - \sum_{i=1}^{I} U_i(\underline{\theta}_i). \tag{23.D.16}$$

By inspection of (23.D.16), we see that any two Bayesian incentive compatible social choice functions that generate the same functions $(y_1(\theta), \ldots, y_I(\theta))$ and the same values of $(U_1(\underline{\theta}_1), \ldots, U_I(\underline{\theta}_I))$ generate the same expected revenue for the seller. ■

As an example of the application of Proposition 23.D.3, consider the equilibria of the first-price and second-price sealed-bid auctions that we identified in Examples 23.B.5 and 23.B.6 (where the buyers' valuations were independently drawn from the uniform distribution on [0, 1]). For these equilibria, the conditions of the revenue equivalence theorem are satisfied: in both auctions the buyer with the highest valuation always gets the good and a buyer with a zero valuation has an expected utility of zero. Thus, the revenue equivalence theorem tells us that the seller receives exactly the same level of expected revenue in these equilibria of the two auctions (you can confirm this fact in Exercise 23.D.3). More generally, it can be shown that in any *symmetric auction setting* (i.e., one where the buyers' valuations are independently drawn from identical distributions), the conditions of the revenue equivalence theorem will be met for *any* Bayesian Nash equilibrium of the first-price sealed-bid auction and the (dominant strategy) equilibrium of the second-price sealed-bid auction (see Exercise 23.D.4 for a consideration of symmetric equilibria in these settings). We can conclude from Proposition 23.D.3, therefore, that in any such setting the first-price and second-price sealed-bid auctions generate exactly the same revenue for the seller.

23.E Participation Constraints

In Sections 23.B to 23.D, we have studied the constraints that the presence of private information puts on the set of implementable social choice functions. Our analysis up to this point, however, has assumed implicitly that each agent i has no choice but to participate in any mechanism chosen by the mechanism designer. That is, agent i's discretion was limited to choosing his optimal actions within those allowed by the mechanism.

In many applications, however, agents' participation in the mechanism is *voluntary*. As a result, the social choice function that is to be implemented by a mechanism must not only be incentive compatible but must also satisfy certain *participation* (or *individual rationality*) *constraints* if it is to be successfully implemented. In this section, we provide a brief discussion of these additional

constraints on the set of implementable social choice functions. By way of motivating our study, Example 23.E.1 provides a simple illustration of how the presence of participation constraints may limit the set of social choice functions that can be successfully implemented.

Example 23.E.1: *Participation Constraints in Public Project Choice.* Consider the following simple example of public project choice (recall our initial discussion of public project choice in Example 23.B.3). A decision must be made whether to do a given project or not, so that $K = \{0, 1\}$. There are two agents, 1 and 2. For each agent i, $\Theta_i = \{\underline{\theta}, \bar{\theta}\}$, so that each agent either has a valuation of $\underline{\theta}$, or a valuation of $\bar{\theta}$. We shall assume that $\bar{\theta} > 2\underline{\theta} > 0$. The cost of the project is $c \in (2\underline{\theta}, \bar{\theta})$. Suppose that we want to implement a social choice function having an ex post efficient project choice; that is, one that has $k^*(\theta_1, \theta_2) = 1$ if either θ_1 or θ_2 is equal to $\bar{\theta}$, and $k^*(\theta_1, \theta_2) = 0$ if $\theta_1 = \theta_2 = \underline{\theta}$. In the absence of the need to insure voluntary participation, we know from Section 23.C that we can implement some such social choice function in dominant strategies using a Groves scheme.

Suppose, however, that each agent has the option of withdrawing from the mechanism at any time (perhaps by withdrawing from the group), and that, if he does, he will not enjoy the benefits of the project if it is done, but will also avoid paying any monetary transfers. Can we implement a social choice function that achieves voluntary participation and that has an ex post efficient project choice?[36] The answer is "no." To see this, note that if agent 1 can withdraw at any time, then to insure his participation it must be that $t_1(\underline{\theta}, \bar{\theta}) \geq -\underline{\theta}$. That is, it must be that whenever his valuation for the project is $\underline{\theta}$, he pays no more than $\underline{\theta}$ toward the cost of the project. Now consider what agent 1's transfer must be when both agents announce that they have valuation $\bar{\theta}$: If truth telling is to be a dominant strategy, then $t_1(\bar{\theta}, \bar{\theta})$ must satisfy

$$\bar{\theta}k^*(\bar{\theta}, \bar{\theta}) + t_1(\bar{\theta}, \bar{\theta}) \geq \bar{\theta}k^*(\underline{\theta}, \bar{\theta}) + t_1(\underline{\theta}, \bar{\theta}),$$

or, substituting for $k^*(\bar{\theta}, \bar{\theta})$ and $k^*(\underline{\theta}, \bar{\theta})$,

$$\bar{\theta} + t_1(\bar{\theta}, \bar{\theta}) \geq \bar{\theta} + t_1(\underline{\theta}, \bar{\theta}).$$

Since $t_1(\underline{\theta}, \bar{\theta}) \geq -\underline{\theta}$, this implies that $t_1(\bar{\theta}, \bar{\theta}) \geq -\underline{\theta}$. Thus, we conclude that agent 1 must not make a contribution toward the cost of the project that exceeds $\underline{\theta}$ when $(\theta_1, \theta_2) = (\bar{\theta}, \bar{\theta})$. Moreover, by symmetry, we have exactly the same constraint for agent 2's transfer when $(\theta_1, \theta_2) = (\bar{\theta}, \bar{\theta})$, namely, $t_2(\bar{\theta}, \bar{\theta}) \geq -\underline{\theta}$. Hence, $t_1(\bar{\theta}, \bar{\theta}) + t_2(\bar{\theta}, \bar{\theta}) \geq -2\underline{\theta}$. But if this is so, then because $2\underline{\theta} < c$, the feasibility condition $t_1(\bar{\theta}, \bar{\theta}) + t_2(\bar{\theta}, \bar{\theta}) \leq -c$ cannot be satisfied. We conclude, therefore, that it is impossible to implement a social choice function with an ex post efficient project choice when the agents can withdraw from the mechanism at any time.

Note also that the presence of an "outside agent" (say "agent 0") who does not care about the project decision does not help at all here when that agent can also withdraw from the mechanism at any time. This is because, to insure this agent's participation, his transfer $t_0(\theta_1, \theta_2)$ must be nonnegative for every realization of

36. Note that any social choice function that *fails* to have both agents participate is necessarily *ex post* inefficient because one of the agents is excluded from the benefits of the project.

(θ_1, θ_2). In particular, we must have $t_0(\bar{\theta}, \bar{\theta}) \geq 0$, and so we must fail to satisfy the feasibility condition $t_0(\bar{\theta}, \bar{\theta}) + t_1(\bar{\theta}, \bar{\theta}) + t_2(\bar{\theta}, \bar{\theta}) \leq -c$. ■

As a general matter, we can distinguish among three stages at which participation constraints may be relevant in any particular application. First, as in Example 23.E.1, an agent i may be able to withdraw from the mechanism at the *ex post stage* that arises after the agents have announced their types and an outcome in X has been chosen. Formally, suppose that agent i can receive a utility of $\bar{u}_i(\theta_i)$ by withdrawing from the mechanism when his type is θ_i.[37] Then, to insure agent i's participation, we must satisfy the *ex post participation* (or *individual rationality*) *constraints*[38]

$$u_i(f(\theta_i, \theta_{-i}), \theta_i) \geq \bar{u}_i(\theta_i) \qquad \text{for all } (\theta_i, \theta_{-i}). \tag{23.E.1}$$

In other circumstances, agent i may only be able to withdraw from the mechanism at the *interim stage* that arises after the agents have each learned their type but before they have chosen their actions in the mechanism. Letting $U_i(\theta_i|f) = E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i]$ denote agent i's *interim expected utility* from social choice function $f(\cdot)$ when his type is θ_i, agent i will participate in a mechanism that implements social choice function $f(\cdot)$ when he is of type θ_i if and only if $U_i(\theta_i|f)$ is not less than $\bar{u}_i(\theta_i)$. Thus, *interim participation* (or *individual rationality*) *constraints* for agent i require that

$$U_i(\theta_i|f) = E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta_i)|\theta_i] \geq \bar{u}_i(\theta_i) \qquad \text{for all } \theta_i. \tag{23.E.2}$$

In still other cases, agent i might only be able to refuse to participate at the *ex ante stage* that arises before the agents learn their types. Letting $U_i(f) = E_{\theta_i}[U_i(\theta_i|f)] = E[u_i(f(\theta_i, \theta_{-i}), \theta_i)]$ denote agent i's *ex ante expected utility* from a mechanism that implements social choice function $f(\cdot)$, the *ex ante participation* (or *individual rationality*) *constraint* for agent i is

$$U_i(f) \geq E_{\theta_i}[\bar{u}_i(\theta_i)]. \tag{23.E.3}$$

Participation constraints are of the ex ante variety when the agents can agree to be bound by the mechanism prior to learning their types. When, instead, agents know their types prior to the time at which they can agree to be bound by the mechanism, we face interim participation constraints.[39] Finally, if there is no way to bind the

37. We assume that agent i's utility from withdrawal depends only on his own type.

38. We assume throughout that it is always optimal to insure that each agent is always willing to participate. In fact, however, there is no loss of generality from assuming this: When agents can "not participate," any outcome that can arise when some subset I' of the I agents does not participate, say x', should be included in the set X. Because we can always have the mechanism select x' in the circumstances when this subset of agents would have refused to participate, if the set X is defined appropriately we can always replicate the outcome of any mechanism that causes nonparticipation with a mechanism in which all agents are always willing to participate.

39. Recall that the assumption in a Bayesian game that types are drawn from a common prior density is often merely a modeling device for how agents form beliefs about each others' types (see Section 8.E). That is, for analytical purposes we may be representing a setting in which agents' types are already determined but are only privately observed by assuming that there has been a prior random draw of types from a commonly known distribution; but there may not actually be any such prior stage at which the agents could possibly interact.

agents to the assigned outcomes of the mechanism against their will, then we face ex post participation constraints.[40]

Note that if $f(\cdot)$ satisfies (23.E.1), then it satisfies (23.E.2); and, in turn, if it satisfies (23.E.2), then it satisfies (23.E.3). However, the reverse is not true. Thus, the constraints imposed by voluntary participation are most severe when agents can withdraw at the ex post stage, and least severe when they can withdraw only at the ex ante stage.

In summary, when agents' types are privately observed, the set of social choice functions that can be successfully implemented are those that satisfy not only the conditions identified in Sections 23.C and 23.D for incentive compatibility (in, respectively, either a dominant strategy or Bayesian sense, depending on the equilibrium concept we employ) but also any participation constraints that are relevant in the environment under study.

In the remainder of this section, we illustrate further the limitations on the set of implementable social choice functions that may be caused by participation constraints by studying the important *Myerson–Satterthwaite theorem* [due to Myerson and Satterthwaite (1983)].

The Myerson–Satterthwaite Theorem

Consider again the bilateral trade setting introduced in Example 23.B.4. Agent 1 is the seller of an indivisible object and has a valuation for the object that lies in the interval $\Theta_1 = [\underline{\theta}_1, \bar{\theta}_1] \subset \mathbb{R}$; agent 2 is the buyer and has a valuation that lies in $\Theta_2 = [\underline{\theta}_2, \bar{\theta}_2] \subset \mathbb{R}$. The two valuations are statistically independent, and θ_i has distribution function $\Phi_i(\cdot)$ with an associated density $\phi_i(\cdot)$ satisfying $\phi_i(\theta_i) > 0$ for all $\theta_i \in [\underline{\theta}_i, \bar{\theta}_i]$. We let $y_i(\theta)$ denote the probability that agent i receives the good given types $\theta = (\theta_1, \theta_2)$, and so agent i's expected utility given θ is $\theta_i y_i(\theta) + t_i(\theta)$ (we normalize $\bar{m}_i = 0$ for all i).

The expected externality mechanism studied in Section 23.D shows that in this setting we can Bayesian implement an ex post efficient social choice function (or what, in this environment, we might call a "trading rule"). A problem arises with the expected externality mechanism, however, when trade is voluntary. In this case, every type of buyer and seller must have nonnegative expected gains from trade if he is to participate. In particular, if a seller of type θ_1 is to participate in a mechanism that implements social choice function $f(\cdot)$, that is, if participation in the mechanism is to be *individually rational* for this type of seller, it must be that $U_1(\theta_1 | f) \geq \theta_1$, because this seller can achieve an expected utility of θ_1 by not participating in the mechanism and simply consuming the good. Likewise, a buyer of type θ_2 can always earn zero by refusing to participate, and so we must have $U_2(\theta_2 | f) \geq 0$. Unfortunately, these interim participation constraints are not satisfied in the expected externality mechanism (you are asked to verify this in Exercise 23.E.1).

The *Myerson–Satterthwaite Theorem* tells us the following disappointing piece of

40. For example, if the mechanism can lead an agent into bankruptcy, the provisions of bankruptcy law provide an effective lower bound on ex post utilities.

news: Whenever gains from trade are possible, but are not certain,[41] there is *no* ex post efficient social choice function that is both Bayesian incentive compatible and satisfies these interim participation constraints. Thus, under the conditions of the theorem, the presence of both private information and voluntary participation implies that it is impossible to achieve ex post efficiency. (For an illustration of the result for specific functional forms, see Exercise 23.E.7.)

Proposition 23.E.1: (*The Myerson Satterthwaite Theorem*) Consider a bilateral trade setting in which the buyer and seller are risk neutral, the valuations θ_1 and θ_2 are independently drawn from the intervals $[\underline{\theta}_1, \bar{\theta}_1] \subset \mathbb{R}$ and $[\underline{\theta}_2, \bar{\theta}_2] \subset \mathbb{R}$ with strictly positive densities, and $(\underline{\theta}_1, \bar{\theta}_1) \cap (\underline{\theta}_2, \bar{\theta}_2) \neq \varnothing$. Then there is no Bayesian incentive compatible social choice function that is ex post efficient and gives every buyer type and every seller type nonnegative expected gains from participation.

Proof: The argument consists of two steps:

Step 1: In any Bayesian incentive compatible and interim individually rational social choice function $f(\cdot) = [y_1(\cdot), y_2(\cdot), t_1(\cdot), t_2(\cdot)]$ *in which* $y_1(\theta_1, \theta_2) + y_2(\theta_1, \theta_2) = 1$ *and* $t_1(\theta_1, \theta_2) + t_2(\theta_1, \theta_2) = 0$, *we must have*

$$\int_{\underline{\theta}_1}^{\bar{\theta}_1} \int_{\underline{\theta}_2}^{\bar{\theta}_2} y_2(\theta_1, \theta_2)\left[\left(\theta_2 - \frac{1 - \Phi_2(\theta_2)}{\phi_2(\theta_2)}\right) - \left(\theta_1 + \frac{\Phi_1(\theta_1)}{\phi_1(\theta_1)}\right)\right]\phi_1(\theta_1)\phi_2(\theta_2)\, d\theta_2\, d\theta_1 \geq 0. \tag{23.E.4}$$

To see this, note first that the same argument that leads to (23.D.15) can be applied here to give [throughout the proof we suppress the argument f in $U_i(\theta_i|f)$ and simply write $U_i(\theta_i)$]:

$$E[-\bar{t}_2(\theta_2)] = \left[\int_{\underline{\theta}_1}^{\bar{\theta}_1} \int_{\underline{\theta}_2}^{\bar{\theta}_2} y_2(\theta_1, \theta_2)\left(\theta_2 - \frac{1 - \Phi_2(\theta_2)}{\phi_2(\theta_2)}\right)\phi_1(\theta_1)\phi_2(\theta_2)\, d\theta_2\, d\theta_1\right] - U_2(\underline{\theta}_2). \tag{23.E.5}$$

Also, because (23.D.12) implies that

$$U_1(\underline{\theta}_1) = U_1(\bar{\theta}_1) - \int_{\underline{\theta}_1}^{\bar{\theta}_1} \int_{\underline{\theta}_2}^{\bar{\theta}_2} y_1(\theta_1, \theta_2)\phi_2(\theta_2)\, d\theta_2\, d\theta_1,$$

condition (23.D.15) also implies that

$$E[-\bar{t}_1(\theta_1)] = \left[\int_{\underline{\theta}_1}^{\bar{\theta}_1} \int_{\underline{\theta}_2}^{\bar{\theta}_2} y_1(\theta_1, \theta_2)\left(\theta_1 + \frac{\Phi_1(\theta_1)}{\phi_1(\theta_1)}\right)\phi_1(\theta_1)\phi_2(\theta_2)\, d\theta_2\, d\theta_1\right] - U_1(\bar{\theta}_1). \tag{23.E.6}$$

Then, since $y_1(\theta_1, \theta_2) = 1 - y_2(\theta_1, \theta_2)$ we have

$$E[-\bar{t}_1(\theta_1)] = \left[\int_{\underline{\theta}_1}^{\bar{\theta}_1} \int_{\underline{\theta}_2}^{\bar{\theta}_2} \left(\theta_1 + \frac{\Phi_1(\theta_1)}{\phi_1(\theta_1)}\right)\phi_1(\theta_1)\phi_2(\theta_2)\, d\theta_2\, d\theta_1\right]$$
$$- \left[\int_{\underline{\theta}_1}^{\bar{\theta}_1} \int_{\underline{\theta}_2}^{\bar{\theta}_2} y_2(\theta_1, \theta_2)\left(\theta_1 + \frac{\Phi_1(\theta_1)}{\phi_1(\theta_1)}\right)\phi_1(\theta_1)\phi_2(\theta_2)\, d\theta_2\, d\theta_1\right] - U_1(\bar{\theta}_1).$$

41. That is, whenever $(\underline{\theta}_1, \bar{\theta}_1) \cap (\underline{\theta}_2, \bar{\theta}_2) \neq \varnothing$ (or equivalently, $\bar{\theta}_2 > \underline{\theta}_1$ and $\bar{\theta}_1 > \underline{\theta}_2$), so that for some realizations of $\theta = (\theta_1, \theta_2)$ there are gains from trade but for others there are not.

But

$$\left[\int_{\underline{\theta}_1}^{\bar{\theta}_1}\int_{\underline{\theta}_2}^{\bar{\theta}_2}\left(\theta_1+\frac{\Phi_1(\theta_1)}{\phi_1(\theta_1)}\right)\phi_1(\theta_1)\phi_2(\theta_2)\,d\theta_2\,d\theta_1\right]=\left[\int_{\underline{\theta}_1}^{\bar{\theta}_1}[\theta_1\phi_1(\theta_1)+\Phi_1(\theta_1)]\,d\theta_1\right]$$

$$=[\theta_1\Phi_1(\theta_1)]_{\underline{\theta}_1}^{\bar{\theta}_1}$$

$$=\bar{\theta}_1.$$

Thus,

$$E[-\bar{t}_1(\theta_1)]=\bar{\theta}_1-\left[\int_{\underline{\theta}_1}^{\bar{\theta}_1}\int_{\underline{\theta}_2}^{\bar{\theta}_2}y_2(\theta_1,\theta_2)\left(\theta_1+\frac{\Phi_1(\theta_1)}{\phi_1(\theta_1)}\right)\phi_1(\theta_1)\phi_2(\theta_2)\,d\theta_2\,d\theta_1\right]-U_1(\bar{\theta}_1). \tag{23.E.7}$$

Now, the fact that $t_1(\theta_1,\theta_2)+t_2(\theta_1,\theta_2)=0$ implies that $E[-t_1(\theta_1,\theta_2)]+E[-t_2(\theta_1,\theta_2)]=0$. So, adding (23.E.5) and (23.E.7) we see that

$$[U_1(\bar{\theta}_1)-\bar{\theta}_1]+U_2(\underline{\theta}_2)=$$

$$\int_{\underline{\theta}_1}^{\bar{\theta}_1}\int_{\underline{\theta}_2}^{\bar{\theta}_2}y_2(\theta_1,\theta_2)\left[\left(\theta_2-\frac{1-\Phi_2(\theta_2)}{\phi_2(\theta_2)}\right)-\left(\theta_1+\frac{\Phi_1(\theta_1)}{\phi_1(\theta_1)}\right)\right]\phi_1(\theta_1)\phi_2(\theta_2)\,d\theta_2\,d\theta_1.$$

But individual rationality implies that $U_1(\bar{\theta}_1)\geq\bar{\theta}_1$ and $U_2(\underline{\theta}_2)\geq 0$, which establishes (23.E.4).

Step 2: Condition (23.E.4) cannot be satisfied if $y_2(\theta_1,\theta_2)=1$ *whenever* $\theta_2>\theta_1$ *and* $y_2(\theta_1,\theta_2)=0$ *whenever* $\theta_2<\theta_1$.

Suppose it were. Then the left-hand side of (23.E.4) could be written as

$$\int_{\underline{\theta}_2}^{\bar{\theta}_2}\int_{\underline{\theta}_1}^{\text{Min}\{\theta_2,\theta_1\}}\left[\left(\theta_2-\frac{1-\Phi_2(\theta_2)}{\phi_2(\theta_2)}-\theta_1\right)\phi_1(\theta_1)-\Phi_1(\theta_1)\right]\phi_2(\theta_2)\,d\theta_1\,d\theta_2$$

$$=\int_{\underline{\theta}_2}^{\bar{\theta}_2}\left[\left(\theta_2-\frac{1-\Phi_2(\theta_2)}{\phi_2(\theta_2)}-\theta_1\right)\Phi_1(\theta_1)\right]_{\underline{\theta}_1}^{\text{Min}\{\theta_2,\bar{\theta}_1\}}\phi_2(\theta_2)\,d\theta_2$$

$$=\int_{\underline{\theta}_2}^{\bar{\theta}_2}\left[\left(\theta_2-\frac{1-\Phi_2(\theta_2)}{\phi_2(\theta_2)}-\text{Min}\{\theta_2,\bar{\theta}_1\}\right)\Phi_1(\text{Min}\{\theta_2,\bar{\theta}_1\})\right]\phi_2(\theta_2)\,d\theta_2$$

$$=-\int_{\underline{\theta}_2}^{\bar{\theta}_1}[1-\Phi_2(\theta_2)]\Phi_1(\theta_2)\,d\theta_2+\int_{\bar{\theta}_1}^{\bar{\theta}_2}[(\theta_2-\bar{\theta}_1)\phi_2(\theta_2)+(\Phi_2(\theta_2)-1)]\,d\theta_2$$

$$=-\int_{\underline{\theta}_2}^{\bar{\theta}_1}[1-\Phi_2(\theta_2)]\Phi_1(\theta_2)\,d\theta_2+[(\theta_2-\bar{\theta}_1)(\Phi_2(\theta_2)-1)]_{\bar{\theta}_1}^{\bar{\theta}_2}$$

$$=-\int_{\underline{\theta}_2}^{\bar{\theta}_1}[1-\Phi_2(\theta_2)]\Phi_1(\theta_2)\,d\theta_2$$

$$<0,$$

where the inequality follows because $\bar{\theta}_1>\underline{\theta}_2$ and $\underline{\theta}_1<\bar{\theta}_2$. This contradicts (23.E.4) and completes the argument. ■

Recalling the revelation principle for Bayesian Nash equilibrium (Proposition 23.D.1), the implication of the Myerson–Satterthwaite theorem can be put as follows: Consider *any* voluntary trading institution that regulates trade between the buyer and the seller. This includes, for example, any bargaining process in which the parties can make offers and counteroffers to each other, as well as any arbitration mechanism in which the parties tell a third party their types and this third party then decides

whether trade will occur and at what price.[42] By the revelation principle, we know that the social choice function that is indirectly implemented in a Bayesian Nash equilibrium[43] of such a mechanism must be Bayesian incentive compatible. Moreover, since participation is voluntary, this social choice function $f(\cdot)$ must satisfy the interim individual rationality constraints that $U_1(\theta_1|f) \geq \theta_1$ for all θ_1 and $U_2(\theta_2|f) \geq 0$ for all θ_2. Thus, the Myerson–Satterthwaite theorem tells us that, under its assumptions, no voluntary trading institution can have a Bayesian Nash equilibrium that leads to an ex post efficient outcome for all realizations of the buyer's and seller's valuations.

23.F Optimal Bayesian Mechanisms

In Sections 23.B to 23.E we have been concerned with the identification of implementable social choice functions in environments characterized by incomplete information about agents' preferences. In this section, we shift our focus to the welfare evaluaton of implementable social choice functions. We begin by developing several welfare criteria that extend the notion of Pareto efficiency that we have used throughout the book in the context of economies with complete information to these incomplete information settings. With these welfare notions in hand, we then discuss several examples that illustrate the characterization of optimal social choice functions (and, by implication, the optimal direct revelation mechanisms that implement them). We restrict our focus throughout this section to implementation in Bayesian Nash equilibria, discussed in detail in Section 23.D. Unless otherwise noted, we also adopt the assumptions and notation of Section 23.D. Good sources for further reading on the subject of this section are Holmstrom and Myerson (1983), Myerson (1991), and Fudenberg and Tirole (1991).

For economies in which agents' preferences are known with certainty, the concept of Pareto efficiency (or Pareto optimality) provides a minimal test that any welfare optimal outcome $x \in X$ should pass: There should be no other feasible outcome $\hat{x} \in X$ with the property that some agents are strictly better off with outcome $\hat{x}$ than with outcome x, and no agent is worse off.

The extension of this welfare test to social choice functions in settings of incomplete information should read something like the following:

> The social choice function $f(\cdot)$ is efficient if it is feasible and if there is no other feasible social choice function that makes some agents strictly better off, and no agents worse off.

To operationalize this idea, however, we need to be more specific about two things: First, what exactly do we mean by a social choice function being "feasible"? Second,

42. Strictly speaking, for a direct application of Proposition 23.E.1, the date of delivery and consumption of the good must be fixed (so the bargaining processes studied in Appendix A of Chapter 9 would not count). But through a suitable reinterpretation Proposition 23.E.1 can be applied to settings in which trade may take place over real time, where not only delivery of the good matters but also the *time of delivery* (see Exercise 23.E.4 for details).

43. And, hence, in any perfect Bayesian or sequential equilibrium (see Section 9.C).

precisely what do we mean when we say that no other feasible social choice function "makes some agents strictly better off, and no agent worse off"?

Let us consider the first of these issues. The identification of the set of feasible social choice functions when agents' preferences are private information has been discussed extensively in Sections 23.D and 23.E. Suppose that we define the set

$$F_{BIC} = \{f: \Theta \to X: f(\cdot) \text{ is Bayesian incentive compatible}\}. \quad (23.F.1)$$

The elements of set F_{BIC} in any particular application are the social choice functions that satisfy condition (23.D.1), the condition that assures that there is a Bayesian Nash equilibrium of the direct revelation mechanism $\Gamma = (\Theta_1, \ldots, \Theta_I, f(\cdot))$ in which truth telling is each agent's equilibrium strategy.

Likewise, following the discussion in Section 23.E, we can also define the set

$$F_{IR} = \{f: \Theta \to X: f(\cdot) \text{ is individually rational}\}. \quad (23.F.2)$$

The set F_{IR} contains those social choice functions that satisfy whichever of the three types of individual rationality (or participation) constraints (23.E.1)–(23.E.3) are relevant in the application being studied. If no individual rationality constraints are relevant (i.e., if agents' participation is not voluntary), then we simply have $F_{IR} = \{f: \Theta \to X\}$, the set of all possible social choice functions.

The content of our discussion in Sections 23.D and 23.E is therefore that the set of feasible social choice functions in environments in which agents' types are private information is precisely $F^* = F_{BIC} \cap F_{IR}$. Following Myerson (1991), we call this the *incentive feasible set* to emphasize that it is the set of feasible social choice functions when, because of incomplete information, incentive compatibility conditions must be satisfied.

Now consider the second issue: What do we mean when we say that no other feasible social choice function would "make some agents strictly better off, and no agents worse off"? The critical issue here has to do with the *timing* of our welfare analysis. In particular, is the welfare analysis occurring *before* the agents (privately) learn their types, or *after*? The former amounts to a welfare analysis conducted at what we called in Section 23.E the *ex ante stage* (the point in time at which agents have not yet learned their types); the latter corresponds to what we called in Section 23.E the *interim stage* (the point in time after each agent has learned his type, but before the agents' types are publicly revealed). To formally define the different welfare criteria that arise in these two cases, let us once again denote by $U_i(\theta_i|f)$ agent i's expected utility from social choice function $f(\cdot)$ conditional on being of type θ_i. Also let $U_i(f) = E_{\theta_i}[U_i(\theta_i|f)]$ denote agent i's ex ante expected utility from social choice function $f(\cdot)$. We can now state Definitions 23.F.1 and 23.F.2.

Definition 23.F.1: Given any set of feasible social choice functions F, the social choice function $f(\cdot) \in F$ is *ex ante efficient in F* if there is no $\hat{f}(\cdot) \in F$ having the property that $U_i(\hat{f}) \geq U_i(f)$ for all $i = 1, \ldots, I$, and $U_i(\hat{f}) > U_i(f)$ for some i.

Definition 23.F.2: Given any set of feasible social choice functions F, the social choice function $f(\cdot) \in F$ is *interim efficient in F* if there is no $\hat{f}(\cdot) \in F$ having the property that $U_i(\theta_i|\hat{f}) \geq U_i(\theta_i|f)$ for all $\theta_i \in \Theta_i$ and all $i = 1, \ldots, I$, and $U_i(\theta_i|\hat{f}) > U_i(\theta_i|f)$ for some i and $\theta_i \in \Theta_i$.

The motivation for the ex ante efficiency test is straightforward: If agents have not yet learned their types, then when comparing two feasible social choice functions

we should evaluate each agent's well-being using his expected utility over all of his possible types. However, when our welfare analysis occurs after agents have (privately) learned their types, things are a bit trickier. Although the agents each know their types, we—as outsiders—do not know them. Thus, the appropriate notion for us to adopt in saying that one social choice function $\hat{f}(\cdot)$ welfare dominates another social choice function $f(\cdot)$ is that $\hat{f}(\cdot)$ makes every possible type of every agent at least as well off as does $f(\cdot)$, and makes some type of some agent strictly better off. This leads to the concept of interim efficiency given in Definition 23.F.2.

Proposition 23.F.1 compares these two notions of efficiency.

Proposition 23.F.1: Given any set of feasible social choice functions F, if the social choice function $f(\cdot) \in F$ is ex ante efficient in F, then it is also interim efficient in F.

Proof: Suppose that $f(\cdot)$ is ex ante efficient in F but is not interim efficient in F. Then there exists an $\hat{f}(\cdot) \in F$ such that $U_i(\theta_i|\hat{f}) \geq U_i(\theta_i|f)$ for all $\theta_i \in \Theta_i$ and all $i = 1, \ldots, I$, and $U_i(\theta_i|\hat{f}) > U_i(\theta_i|f)$ for some i and $\theta_i \in \Theta_i$. But since, for all i, $U_i(f) = E_{\theta_i}[U_i(\theta_i|f)]$ and $U_i(\hat{f}) = E_{\theta_i}[U_i(\theta_i|\hat{f})]$, it follows that $U_i(\hat{f}) \geq U_i(f)$ for all $i = 1, \ldots, I$, and $U_i(\hat{f}) > U_i(f)$ for some i, contradicting the hypothesis that $f(\cdot)$ is ex ante efficient in F. ■

The ex ante efficiency concept is more demanding than is interim efficiency (and so fewer social choice functions $f(\cdot)$ pass the ex ante efficiency test) because a social choice function $\hat{f}(\cdot)$ can raise every agent's ex ante expected utility relative to the social choice function $f(\cdot)$ even though $\hat{f}(\cdot)$ may lead some type of some agent i to have a lower expected utility than he does with $f(\cdot)$.

Putting together the elements developed above, we conclude that when agents' types are already determined at the time we are conducting our welfare analysis, the proper notion of efficiency of a social choice function in an environment with incomplete information is interim efficiency in F^*, the set of Bayesian incentive compatible and individually rational social choice functions.[44] On the other hand, if our analysis is conducted prior to agents learning their types, then the proper notion of efficiency is ex ante efficiency in F^*.[45] These two notions are often called simply *ex ante incentive efficiency* and *interim incentive efficiency* [the terminology is due to Holmstrom and Myerson (1983)], where the modifier "incentive" is meant to convey the point that the set F^* is being used.[46]

These two welfare notions differ from the ex post efficiency criterion introduced in Definition 23.B.2. To see their relationship to it more clearly, Definition 23.F.3

44. These cases often correspond to situations in which our assumption that the agents' types are drawn from a known prior distribution is being used merely as a device to model agents' beliefs about each others' types, as described in Section 8.E, rather than as a description of any actual prior time at which the agents could interact or our welfare analysis might have been done.

45. This case often arises in contracting problems when, at the time of contracting, the agents anticipate that they will later come to acquire private information about their types. Then the natural welfare standard to use in comparing different contracts (i.e., different mechanisms) is the ex ante criterion. The principal agent model studied in Section 14.C and Example 23.F.1 below is an example along these lines.

46. However, since the relevant individual rationality constraints vary from one application to another, it is usually clearer to describe precisely the set F within which efficiency is being evaluated.

develops the ex post efficiency notion in a manner that parallels Definitions 23.F.1 and 23.F.2.

Definition 23.F.3: Given any set of feasible social choice functions F, the social choice function $f(\cdot) \in F$ is *ex post efficient in* F if there is no $\hat{f}(\cdot) \in F$ having the property that $u_i(\hat{f}(\theta), \theta_i) \geq u_i(f(\theta), \theta_i)$ for all $i = 1, \ldots, I$ and all $\theta \in \Theta$, and $u_i(\hat{f}(\theta), \theta_i) > u_i(f(\theta), \theta_i)$ for some i and $\theta \in \Theta$.

The ex post efficiency test in Definition 23.F.3 conducts its welfare evaluation at the ex post stage at which all agents' information has been publicly revealed. Using this definition, we see that a social choice function $f(\cdot)$ is ex post efficient in the sense of Definition 23.B.2 if and only if it is ex post efficient in the sense of Definition 23.F.3 when we take $F = \{f: \Theta \to X\}$.

Note that the criterion of ex post efficiency in $\{f: \Theta \to X\}$, or more generally, of ex post efficiency in F_{IR} when individual rationality constraints are present, ignores issues of incentive compatibility. As a result, it is appropriate as a welfare criterion only if agents' types are in fact publicly observable. Because $F^* \subset F_{IR}$, allocations that are ex ante or interim incentive efficient need not be ex post efficient in this sense. Indeed, the Myerson–Satterthwaite theorem (Proposition 23.E.1) provides an illustration of this phenomenon for the bilateral trade setting: under its assumptions, no element of F^* is ex post efficient. Examples 23.F.1 to 23.F.3 provide further illustrations. (For one way in which the notion of ex post efficiency is nevertheless still of interest in settings with privately observed types, see Exercise 23.F.1.)

Note also that even in settings in which agents' types are public information, the use of ex post efficiency in F_{IR} as our welfare criterion is appropriate only when agents' types are already determined. When our welfare analysis instead occurs prior to agents learning their types, the appropriate notion is instead the stronger criterion that $f(\cdot)$ be ex ante efficient in F_{IR}. These two notions are sometimes called *ex post classical efficiency* and *ex ante classical efficiency* [again, the terminology is due to Holmstrom and Myerson (1983)] to indicate that no incentive constraints are involved in defining the feasible set of social choice functions.

In the remainder of this section we study three examples in which we characterize welfare optimal social choice functions. In Examples 23.F.1 and 23.F.2, it is supposed that one agent who receives no private information chooses a mechanism to maximize his expected utility subject to both incentive compatibility constraints and interim individual rationality constraints for the other agents. These two examples therefore amount to a characterization of one particular interim incentive efficient mechanism. In Example 23.F.3, we provide a full characterization of the sets of interim and ex ante incentive efficient social choice functions for a simple setting of bilateral trade with adverse selection.

Example 23.F.1: *A Principal–Agent Problem with Hidden Information.* In Section 14.C we studied principal–agent problems with hidden information for the case in which the agent has two possible types. Here we consider the case where the agent may have a continuum of types. Recall from Section 14.C that in the principal–agent problem with hidden information, the principal faces the problem of designing an optimal (i.e., payoff maximizing) contract for an agent who will come to possess private information. In doing so, the principal faces both incentive constraints and

a reservation utility constraint for the agent. Recall also from Section 14.C that, in the limiting case in which the agent is infinitely risk averse, the agent must be guaranteed his reservation utility for each possible type he may come to have, and so this contracting problem is identical to the contracting problem that would arise if the agent already knew his type at the time of contracting. Here we shall set things up directly in these terms, assuming that the agent already possesses this information when contracting occurs. With this formulation, the principal's optimal contract can be viewed as implementing one particular interim incentive efficient social choice function. (When the agent actually does not know his type at the time of contracting and is infinitely risk averse, then this social choice function is also ex ante incentive efficient.)

To introduce our notation, we suppose that the agent (individual 1) may take some observable action $e \in \mathbb{R}_+$ (his "effort" or "task" level) and receives a monetary payment from the principal of t_1. The agent's type is drawn from the interval $[\underline{\theta}, \bar{\theta}]$, where $\underline{\theta} < \bar{\theta} < 0$, according to the distribution function $\Phi(\cdot)$ which has an associated density function $\phi(\cdot)$ that is strictly positive on $[\underline{\theta}, \bar{\theta}]$. We assume that this distribution satisfies the property that $[\theta - ((1 - \Phi(\theta))/\phi(\theta))]$ is nondecreasing in θ.[47]

The agent's Bernoulli utility function when his type is θ is $u_1(e, t_1, \theta) = t_1 + \theta g(e)$, where $g(\cdot)$ is a differentiable function with $g(0) = 0$, $g(e) > 0$ for $e > 0$, $g'(0) = 0$, $g'(e) > 0$ for $e > 0$, and $g''(\cdot) > 0$; that is, $\theta g(e)$ represents the agent's disutility of effort (recall that $\theta < 0$), with higher effort levels leading to an increasing level of disutility to the agent. Note that a larger (i.e., less negative) level of θ lowers, at any level of e, both the agent's total level of disutility and his marginal disutility from any increase in e. As noted above, we suppose that the agent must be guaranteed an expected utility level of at least $\bar{u}$ for each possible type he may have.

The principal (individual 0) has no private information. His Bernoulli utility function is $u_0(e, t_0) = v(e) + t_0$, where t_0 is his net transfer and $v(\cdot)$ is a differentiable function satisfying $v'(\cdot) > 0$ and $v''(\cdot) < 0$.

A contract between the principal and the agent can be viewed as specifying a mechanism in the sense we have used throughout this chapter. By the revelation principle for Bayesian Nash equilibrium (Proposition 23.D.1), the equilibrium outcome induced by such a contract, formally a social choice function that maps each possible agent type into effort and transfer levels, can always be duplicated using a direct revelation mechanism that induces truth telling. Thus, the principal can confine his search for an optimal contract to the set of Bayesian incentive compatible social choice functions $f(\cdot) = (e(\cdot), t_0(\cdot), t_1(\cdot))$ that give the agent an expected utility of at least $\bar{u}$ for every possible value of θ. In what follows, we shall (without loss of generality) restrict attention in our search for the principal's optimal contract to contracts that have $t_0(\theta) = -t_1(\theta)$ for all θ (i.e., that involve no waste of numeraire).

The principal's problem can therefore be stated as

$$\begin{aligned} \underset{f(\cdot)=(e(\cdot),t_1(\cdot))}{\text{Max}} \quad & E[v(e(\theta)) - t_1(\theta)] \\ \text{s.t. } & f(\cdot) \text{ is Bayesian incentive compatible and} \\ & \text{individually rational.} \end{aligned}$$

47. For a discussion of how the analysis changes when this assumption is not satisfied, see Fudenberg and Tirole (1991).

The present model falls into the class of models with linear utility studied in Section 23.D [specifically, in the notation of Proposition 23.D.2, $k = e$, $v_1(k) = g(e)$, and $\bar{v}_1(\theta) = g(e(\theta))$ here]. Letting $U_1(\theta) = t_1(\theta) + \theta g(e(\theta))$ denote the agent's utility if his type is θ and he tells the truth, Proposition 23.D.2 can be used to restate the principal's problem in terms of choosing the functions $e(\cdot)$ and $U_1(\cdot)$ to solve

$$\begin{aligned} \underset{e(\cdot),U_1(\cdot)}{\text{Max}} \quad & E[v(e(\theta)) + \theta g(e(\theta)) - U_1(\theta)] \\ \text{s.t.} \quad & \text{(i) } e(\cdot) \text{ is nondecreasing} \\ & \text{(ii) } U_1(\theta) = U_1(\underline{\theta}) + \int_{\underline{\theta}}^{\theta} g(e(s))\, ds \text{ for all } \theta \\ & \text{(iii) } U_1(\theta) \geq \bar{u} \text{ for all } \theta. \end{aligned} \tag{23.F.3}$$

Constraints (i) and (ii) are the necessary and sufficient conditions for the principal's contract to be Bayesian incentive compatible, adapted from Proposition 23.D.2 [constraint (i) follows because $g(\cdot)$ is increasing in e], while constraint (iii) is the agent's individual rationality constraint.

Note first that if constraint (ii) is satisfied, then constraint (iii) will be satisfied if and only if $U_1(\underline{\theta}) \geq \bar{u}$. As a result, we can replace constraint (iii) with

$$\text{(iii')} \quad U_1(\underline{\theta}) \geq \bar{u}.$$

Next, substituting for $U_1(\theta)$ in the objective function from constraint (ii), and then integrating by parts in a fashion similar to the steps leading to (23.D.14), problem (23.F.3) can be restated as

$$\begin{aligned} \underset{e(\cdot),U_1(\underline{\theta})}{\text{Max}} \quad & \left[\int_{\underline{\theta}}^{\bar{\theta}} \left\{v(e(\theta)) + g(e(\theta))\left(\theta - \frac{1-\Phi(\theta)}{\phi(\theta)}\right)\right\} \phi(\theta)\, d\theta\right] - U_1(\underline{\theta}) \\ \text{s.t.} \quad & \text{(i) } e(\cdot) \text{ is nondecreasing} \\ & \text{(iii')} \; U_1(\underline{\theta}) \geq \bar{u}. \end{aligned} \tag{23.F.4}$$

It is now immediate from (23.F.4) that in any solution we must in fact have $U_1(\underline{\theta}) = \bar{u}$. Thus, we can write the principal's problem as one of choosing $e(\cdot)$ to solve

$$\begin{aligned} \underset{e(\cdot)}{\text{Max}} \quad & \left[\int_{\underline{\theta}}^{\bar{\theta}} \left\{v(e(\theta)) + g(e(\theta))\left(\theta - \frac{1-\Phi(\theta)}{\phi(\theta)}\right)\right\} \phi(\theta)\, d\theta\right] - \bar{u} \\ \text{s.t.} \quad & \text{(i) } e(\cdot) \text{ is nondecreasing.} \end{aligned} \tag{23.F.5}$$

Suppose for the moment that we can ignore constraint (i). Then the optimal function $e(\cdot)$ must satisfy the first-order condition[48]

$$v'(e(\theta)) + g'(e(\theta))\left(\theta - \frac{1-\Phi(\theta)}{\phi(\theta)}\right) = 0 \qquad \text{for all } \theta. \tag{23.F.6}$$

But note that, under our assumption that $[\theta - ((1-\Phi(\theta))/\phi(\theta))]$ is nondecreasing in θ, the implicit function theorem applied to (23.F.6) tells us that any solution $e(\cdot)$ to this relaxed problem must in fact be nondecreasing. Thus, (23.F.6) characterizes the solution to the principal's actual problem (see Section M.K of the Mathematical Appendix). The optimal $U_1(\cdot)$ [and, hence, $t_1(\cdot)$] is then calculated from constraint (ii) of (23.F.3) using this optimal $e(\cdot)$ and the fact that $U_1(\underline{\theta}) = \bar{u}$.

48. It can be shown that under our assumptions, the optimal contract is interior, that is, has $e(\theta) > 0$ for (almost) all θ.

It is interesting to compare this solution with the optimal contract for the case in which the agent's type is observable. This contract solves

$$\begin{aligned} \underset{e(\cdot),t_1(\cdot)}{\text{Max}} \quad & E[v(e(\theta)) - t_1(\theta)] \\ & \text{s.t. } t_1(\theta) + \theta g(e(\theta)) \geq \bar{u} \text{ for all } \theta. \end{aligned}$$

Hence, the optimal task level in this complete information contract is the level $e^*(\theta)$ that satisfies, for all θ,

$$v'(e^*(\theta)) + g'(e^*(\theta))\theta = 0.$$

Note that $e^*(\theta)$ is the level that arises in any ex post (classically) efficient social choice function. In contrast, the principal's optimal $e(\cdot)$ when θ is private information is such that

$$v'(e(\theta)) + g'(e(\theta))\theta \begin{cases} > 0 & \text{at all } \theta < \bar{\theta}, \\ = 0 & \text{at } \theta = \bar{\theta}. \end{cases}$$

We see then that $e(\theta) < e^*(\theta)$ for all $\theta < \bar{\theta}$, and $e(\bar{\theta}) = e^*(\bar{\theta})$. This is a version of the same result that we saw for the two-type case in Section 14.C. In the optimal contract, the type of agent with the lowest disutility from effort (here type $\bar{\theta}$; in Section 14.C, type θ_H) takes an ex post efficient action, while all other types have their effort levels distorted downward. The reason is also the same: doing so helps reduce the amount the agent's utility exceeds his reservation utility for types $\theta > \underline{\theta}$ (his so-called *information rents*). To see this point heuristically, suppose that starting with some function $e(\cdot)$ we lower $e(\hat{\theta})$ by an amount $de < 0$ for some type $\hat{\theta} \in (\underline{\theta}, \bar{\theta})$ and lower this type's transfer to keep his utility unchanged.[49] The decrease in the transfer paid to type $\hat{\theta}$ is $\hat{\theta}g'(e(\hat{\theta}))\,de$, while the direct effect on the principal is $v'(e(\hat{\theta}))\,de$. At the same time, according to constraint (ii), this change in $e(\hat{\theta})$ lowers the utility level, and hence the transfer, that must be given to all types $\theta > \hat{\theta}$ by exactly $g'(e(\hat{\theta}))\,de$. The expected value of this reduction in the transfers paid to these types is $-(1 - \Phi(\hat{\theta}))g'(e(\hat{\theta}))\,de$. If the original $e(\cdot)$ is an optimum, the sum of the first two changes in the principal's profits (those for type $\hat{\theta}$) weighted by the density of type $\hat{\theta}$, $[v'(e(\hat{\theta})) + \hat{\theta}g'(e(\hat{\theta}))]\phi(\hat{\theta})\,de$, plus the reduction in payments to types $\theta > \hat{\theta}$, $(1 - \Phi(\hat{\theta}))g'(e(\hat{\theta}))\,de$, must equal zero. This gives exactly (23.F.6). ■

Example 23.F.2: *Optimal Auctions.* We consider again the auction setting introduced in Example 23.B.4. Here we determine the optimal auction for the seller of an indivisible object (agent 0) when there are I buyers, indexed by $i = 1, \ldots, I$. Each buyer has a Bernoulli utility function $\theta_i y_i(\theta) + t_i(\theta)$, where $y_i(\theta)$ is the probability that agent i gets the good when the agents' types are $\theta = (\theta_1, \ldots, \theta_I)$. In addition, each buyer i's type is independently drawn according to the distribution function $\Phi_i(\cdot)$ on $[\underline{\theta}_i, \bar{\theta}_i] \subset \mathbb{R}$ with $\underline{\theta}_i \neq \bar{\theta}_i$ and associated density $\phi_i(\cdot)$ that is strictly positive on $[\underline{\theta}_i, \bar{\theta}_i]$. We assume also that, for $i = 1, \ldots, I$,

$$\theta_i - \frac{1 - \Phi_i(\theta_i)}{\phi_i(\theta_i)}$$

is nondecreasing in θ_i.[50]

49. We say "heuristically" because to do this rigorously we need to perform this reduction in e over an interval of types and then take limits.

50. For a discussion of the case in which this assumption is not met, see Myerson (1981).

A social choice function in this environment is a function $f(\cdot) = (y_0(\cdot), \ldots, y_I(\cdot), t_0(\cdot), \ldots, t_I(\cdot))$ having the properties that, for all $\theta \in \Theta$, $y_i(\theta) \in [0, 1]$ for all i, $\sum_{i \neq 0} y_i(\theta) = 1 - y_0(\theta)$, and $t_0(\theta) = -\sum_{i \neq 0} t_i(\theta)$.[51] The seller wishes to choose the Bayesian incentive compatible social choice function that maximizes his expected revenue $E_\theta[t_0(\theta)] = -E_\theta[\sum_{i \neq 0} t_i(\theta)]$ but faces the interim individual rationality constraints that $U_i(\theta_i) = \theta_i \bar{y}_i(\theta_i) + \bar{t}_i(\theta_i) \geq 0$ for all θ_i and $i \neq 0$ [as in Section 23.D, $\bar{y}_i(\theta_i)$ and $\bar{t}_i(\theta_i)$ are agent i's probability of getting the good and expected transfer conditional on announcing his type to be θ_i] because buyers are always free not to participate. The seller's optimal choice is therefore a particular element of the set of interim incentive efficient social choice functions.

The seller's problem can be written as one of choosing functions $y_1(\cdot), \ldots, y_I(\cdot)$ and $U_1(\cdot), \ldots, U_I(\cdot)$ to solve

$$\max_{(y_i(\cdot), U_i(\cdot))_{i=1}^I} \sum_{i \neq 0} \int_{\underline{\theta}_i}^{\theta_i} [\bar{y}_i(\theta_i)\theta_i - U_i(\theta_i)]\phi_i(\theta_i)\, d\theta_i \tag{23.F.7}$$

$$\text{s.t.} \quad \text{(i)} \ \bar{y}_i(\cdot) \text{ is nondecreasing for all } i \neq 0.$$

$$\text{(ii)} \ \text{For all } \theta\text{: } y_i(\theta) \in [0, 1] \text{ for all } i \neq 0, \ \textstyle\sum_{i \neq 0} y_i(\theta) \leq 1.$$

$$\text{(iii)} \ U_i(\theta_i) = U_i(\underline{\theta}_i) + \textstyle\int_{\underline{\theta}_i}^{\theta_i} \bar{y}_i(s)\, ds \text{ for all } i \neq 0 \text{ and } \theta_i.$$

$$\text{(iv)} \ U_i(\theta_i) \geq 0 \text{ for all } i \neq 0 \text{ and } \theta_i.$$

We note first that if constraint (iii) is satisfied then constraint (iv) will be satisfied if and only if $U_i(\underline{\theta}_i) \geq 0$ for all $i \neq 0$. As a result, we can replace constraint (iv) with

$$\text{(iv}') \quad U_i(\underline{\theta}_i) \geq 0 \text{ for all } i \neq 0 \text{ and } \theta_i.$$

Next, substituting into the objective function for $U_i(\theta_i)$ using constraint (iii), and following the same steps that led to (23.D.16), the seller's problem can be written as one of choosing the $y_i(\cdot)$ functions and the values $U_1(\underline{\theta}_1), \ldots, U_I(\underline{\theta}_I)$ to maximize

$$\int_{\underline{\theta}_1}^{\theta_1} \cdots \int_{\underline{\theta}_I}^{\theta_I} \left[\sum_{i=1}^{I} y_i(\theta_1, \ldots, \theta_I) \left(\theta_i - \frac{1 - \Phi_i(\theta_i)}{\phi_i(\theta_i)} \right) \right] \left[\prod_{i=1}^{I} \phi_i(\theta_i) \right] d\theta_I \cdots d\theta_1 - \sum_{i=1}^{I} U_i(\underline{\theta}_i)$$

subject to constraints (i), (ii), and (iv'). It is evident that the solution must have $U_i(\underline{\theta}_i) = 0$ for all $i = 1, \ldots, I$. Hence, the seller's problem reduces to choosing functions $y_1(\cdot), \ldots, y_I(\cdot)$ to maximize

$$\int_{\underline{\theta}_1}^{\theta_1} \cdots \int_{\underline{\theta}_I}^{\theta_I} \left[\sum_{i=1}^{I} y_i(\theta_1, \ldots, \theta_I) \left(\theta_i - \frac{1 - \Phi_i(\theta_i)}{\phi_i(\theta_i)} \right) \right] \left[\prod_{i=1}^{I} \phi_i(\theta_i) \right] d\theta_I \cdots d\theta_1 \tag{23.F.8}$$

subject to constraints (i) and (ii).

Let us ignore constraint (i) for the moment. Define

$$J_i(\theta_i) = \theta_i - \frac{1 - \Phi_i(\theta_i)}{\phi_i(\theta_i)}.$$

Then inspection of (23.F.8) indicates that $y_1(\cdot), \ldots, y_I(\cdot)$ is a solution to this relaxed

51. Once again we restrict attention, without loss of generality, to social choice functions involving no waste of either the numeraire or the good (there is always an optimal social choice function for the seller with this form).

problem if and only if for all $i = 1, \ldots, I$ we have

$$y_i(\theta) = 1 \qquad \text{if } J_i(\theta_i) > \text{Max } \{0, \text{Max}_{h \neq i} J_h(\theta_h)\}$$

and (23.F.9)

$$y_i(\theta) = 0 \qquad \text{if } J_i(\theta_i) < \text{Max } \{0, \text{Max}_{h \neq i} J_h(\theta_h)\}.$$

[Note that $J_i(\theta_i) = \text{Max } \{0, \text{Max}_{h \neq i} J_h(\theta_h)\}$ is a zero probability event.] But, given our assumption that $J_i(\cdot)$ is nondecreasing in θ_i, (23.F.9) implies that $y_i(\cdot)$ is nondecreasing in θ_i, which in turn implies that $\bar{y}_i(\cdot)$ is nondecreasing. Thus the solution to this relaxed problem actually satisfies constraint (i), and so is a solution to the seller's overall problem (see Section M.K of the Mathematical Appendix). The optimal transfer functions can then be set as $t_i(\theta) = U_i(\theta_i) - \theta_i \bar{y}_i(\theta_i)$, where $U_i(\theta_i)$ is calculated from constraint (iii).

A few things should be noted about (23.F.9). First, observe that when the various agents have differing distribution functions $\Phi_i(\cdot)$, the agent i who has the largest value of $J_i(\theta_i)$ is *not* necessarily the same as the agent who has the highest valuation for the object. Thus, the seller's optimal auction need not be ex post (classically) efficient.

Second, in the case of *symmetric* bidders in which $\underline{\theta}_i = \underline{\theta}$ and $J_i(\cdot) = J(\cdot)$ for $i = 1, \ldots, I$, when $\underline{\theta} > 0$ is large enough so that $J(\underline{\theta}) > 0$, the optimal auction always gives the object to the bidder with the highest valuation and also leaves each bidder with an expected utility of zero when his valuation attains its lowest possible value. We can therefore conclude, using the revenue equivalence theorem (Proposition 23.D.3), that the first-price and second-price sealed-bid auctions are both optimal in this case.

Third, the optimal auction has a nice interpretation in terms of monopoly pricing. Consider, for example, the case in which $I = 1$ and $\underline{\theta}_1 = 0$. Then conditions (23.F.9) tell us that the optimal auction gives the object to the buyer (agent 1) if and only if $J_1(\theta_1) = [\theta_1 - ((1 - \Phi_1(\theta_1))/\phi_1(\theta_1))] > 0$. Suppose we think instead of the seller in this circumstance simply naming a price p and letting the buyer then decide whether to buy at this price. The seller's expected revenue from this scheme is $p(1 - \Phi_1(p))$, and so the first-order condition for his optimal posted price, say p^*, is $(1 - \Phi_1(p^*)) - p^*\phi_1(p^*) = 0$, or equivalently, $J_1(p^*) = 0$. Since $J_1(\cdot)$ is nondecreasing, we see that with this optimal posted price policy a buyer of type θ_1 gets the good with probability 1 if $J_1(\theta_1) > 0$, and with probability 0 if $J_1(\theta_1) < 0$, exactly as in the optimal auction derived above. Indeed, given the revenue equivalence theorem we can conclude that in this case this simple posted price scheme is an optimal mechanism for the seller. [For more on the monopoly interpretation of optimal auctions, see Exercise 23.F.5 and Bulow and Roberts (1989).] ■

Throughout the chapter, we have restricted attention to "private values" settings in which agents' utilities depend only on their own types. In a number of settings of economic interest, however, an agent i's utility depends not only his own type, θ_i, but also on the types of other agents, θ_{-i}. That is, agent i's Bernoulli utility function may take the form $u_i(x, \theta)$ rather than $u_i(x, \theta_i)$, where $\theta = (\theta_i, \theta_{-i})$. Fortunately, all of the concepts of implementation for Bayesian Nash equilibrium that we have studied in Sections 23.D to 23.F extend readily to this case. For example, we can say that the social choice function $f: \Theta \to X$ is Bayesian incentive compatible if for all i

and all $\theta_i \in \Theta_i$,

$$E_{\theta_{-i}}[u_i(f(\theta_i, \theta_{-i}), \theta) \mid \theta_i] \geq E_{\theta_{-i}}[u_i(f(\hat{\theta}_i, \theta_{-i}), \theta) \mid \theta_i] \qquad (23.F.10)$$

for all $\hat{\theta}_i \in \Theta_i$. Our third and last example, which studies a simple bilateral trade setting with adverse selection (see Section 13.B for more on adverse selection), falls within this class of models. Another difference from the analysis of Examples 23.F.1 and 23.F.2 is that here we shall characterize the *entire sets* of both ex ante and interim incentive efficient social choice functions.

Example 23.F.3: *Bilateral Trade with Adverse Selection* [from Myerson (1991)]. Consider a bilateral trade setting in which there is a seller (agent 1) and a potential buyer (agent 2) of one unit of an indivisible private good. The good may be of high quality or low quality, but only the seller observes which is the case. To model this, we let the seller have two possible types, so that $\Theta_1 = \{\theta_L, \theta_H\}$, and we assume that $\text{Prob}(\theta_H) = .2$. Both the buyer's and the seller's utilities from consumption of the good depend on the seller's type. In particular, letting y denote the probability that the buyer receives the good, and letting t denote the amount of any monetary transfer from the buyer to the seller, we suppose that[52]

$$\begin{aligned} u_1(y, t \mid \theta_L) &= t + 20(1 - y), \qquad & u_1(y, t \mid \theta_H) &= t + 40(1 - y), \\ u_2(y, t \mid \theta_L) &= 30y - t, & u_2(y, t \mid \theta_H) &= 50y - t. \end{aligned} \qquad (23.F.11)$$

A social choice function in this setting assigns a probability of trade and a transfer for each possible value of θ_1, and so can be represented by a vector (y_L, t_L, y_H, t_H). We suppose that trade is voluntary, and that as a result any feasible social choice function must satisfy interim individual rationality constraints for both the buyer and the seller. For the seller this means that, for each type he may have, his expected utility must be no less than his utility from refusing to participate and simply consuming the good. Hence, we must have

$$t_L + 20(1 - y_L) \geq 20 \qquad (\text{IR}_{1L}) \qquad (23.F.12)$$

$$t_H + 40(1 - y_H) \geq 40. \qquad (\text{IR}_{1H}) \qquad (23.F.13)$$

For the buyer, on the other hand, interim individual rationality simply requires that he receive nonnegative expected utility from participation (recall that he does not observe θ_1). Hence, we must have

$$.2(50y_H - t_H) + .8(30y_L - t_L) \geq 0. \qquad (\text{IR}_2) \qquad (23.F.14)$$

Note from (23.F.11) that if θ_1 were publicly observable then, for each value of θ_1, there would be gains from trade between the buyer and the seller. Because of this fact, any ex post (classically) efficient social choice function has $y_H = y_L = 1$ (see Exercise 23.F.8).

Because θ_1 is only privately observed, the set of feasible social choice functions is the incentive feasible set F^*, the set of Bayesian incentive compatible and (interim) individually rational social choice functions. In the present context, the social choice

52. We assume that there is no waste of either the good or the numeraire, so that the probability that either the buyer or the seller consumes the good is 1, and any transfer from the buyer goes to the seller.

function (y_L, t_L, y_H, t_H) is Bayesian incentive compatible if

$$t_H + 40(1 - y_H) \geq t_L + 40(1 - y_L) \qquad (IC_H) \qquad (23.F.15)$$

and

$$t_L + 20(1 - y_L) \geq t_H + 20(1 - y_H). \qquad (IC_L) \qquad (23.F.16)$$

Condition (23.F.15) requires that truth telling be an optimal strategy for the seller when his type is θ_H; (23.F.16) is the condition for truth telling to be optimal when his type is θ_L. Thus, (y_L, t_L, y_H, t_H) is a feasible social choice function if and only if it satisfies the incentive compatibility constraints (23.F.15)–(23.F.16) and the interim individual rationality constraints (23.F.12)–(23.F.14).

These constraints imply that *any* feasible social choice function possesses the following three properties (Exercise 23.F.9 asks you to establish these points):

(i) No feasible social choice function is ex post (classically) efficient.
(ii) In any feasible social choice function, $y_H \leq y_L$ and $t_H \leq t_L$.
(iii) In any feasible social choice function, the expected gains from trade for a low-quality seller are at least as large as the expected gains from trade for a high-quality seller, that is, $t_L - 20y_L \geq t_H - 40y_H$.

We now proceed to characterize the interim and ex ante incentive efficient social choice functions for this bilateral trade problem. To determine the interim incentive efficient social choice functions, we need to determine the (y_L, t_L, y_H, t_H) that solve, for each possible choice of $\bar{u}_{1H} \geq 0$ and $\bar{u}_2 \geq 0$, the following problem (we have simplified the incentive compatibility and individual rationality constraints by eliminating constants on both sides of the inequalities, and have removed a constant from the objective function as well[53]):

$$\begin{aligned} \underset{(y_l \in [0,1], t_l, y_H \in [0,1], t_H)}{\text{Max}} \quad & t_L - 20y_L \qquad (23.F.17) \\ \text{s.t.} \quad & \text{(i)} \; t_H - 40y_H \geq t_L - 40y_L \\ & \text{(ii)} \; t_L - 20y_L \geq t_H - 20y_H \\ & \text{(iii)} \; t_H - 40y_H \geq \bar{u}_{1H} \\ & \text{(iv)} \; .2(50y_H - t_H) + .8(30y_L - t_L) \geq \bar{u}_2. \end{aligned}$$

Problem (23.F.17) characterizes interim incentive efficient social choice functions by maximizing the interim expected utility of the type θ_L seller subject to giving the type θ_H seller an interim expected utility of at least $\bar{u}_{1H} \geq 0$, giving the buyer an interim expected utility of $\bar{u}_2 \geq 0$ (since the buyer acquires no private information, this is equivalent to giving him an ex ante expected utility of $\bar{u}_2$), and satisfying the seller's incentive compatibility constraints.

We now proceed to characterize the solution to problem (23.F.17) through a series of steps.

Step 1: Any solution to problem (23.F.17) has $y_L = 1$; that is, in any interim incentive efficient social choice function, trade is certain to occur when the good is of low quality.

To see this, suppose that $(y_L^*, t_L^*, y_H^*, t_H^*)$ solves (23.F.17) but that $y_L^* < 1$. Consider a change to social choice function $(\hat{y}_L, \hat{t}_L, \hat{y}_H, \hat{t}_H) = (y_L^* + \varepsilon, t_L^* + 30\varepsilon, y_H^*, t_H^*)$

53. In essence, we have expressed all of these in terms of the agents' *gains from trade.*

where $\varepsilon > 0$. For a sufficiently small $\varepsilon > 0$, this new social choice function satisfies all of the constraints of problem (23.F.17) (check this), and raises the value of the objective function but this contradicts the optimality of $(y_L^*, t_L^*, y_H^*, t_H^*)$.

Step 2: Any solution to problem (23.F.17) has $y_H < 1$; that is, in any interim incentive efficient social choice function, trade does not occur with certainty when the good is of high quality.

Given step 1, if a solution to (23.F.17), say $(y_L^*, t_L^*, y_H^*, t_H^*)$ has $y_H^* = 1$, then $(y_L^*, t_L^*, y_H^*, t_H^*)$ is ex post (classically) efficient (i.e., it has $y_L^* = y_H^* = 1$). But we have already noted above that no such social choice function is incentive feasible (i.e., is an element of F^*).

Step 3: In any solution to (23.F.17), constraint (ii) is binding (i.e., holds with equality).

Suppose that the social choice function $(y_L^*, t_L^*, y_H^*, t_H^*)$ is a solution to (23.F.17) in which constraint (ii) is not binding in the solution. Consider instead the social choice function $(\hat{y}_L, \hat{t}_L, \hat{y}_H, \hat{t}_H) = (y_L^*, t_L^* + \varepsilon, y_H^* + \varepsilon, t_H^* + 45\varepsilon)$ for $\varepsilon > 0$. For small enough $\varepsilon > 0$, this alternative social choice function satisfies all of the constraints of problem (23.F.17) (note that it satisfies $\hat{y}_H < 1$ because, by step 2, $y_H^* < 1$; check the other constraints too). Moreover, it yields a larger value of the objective function of (23.F.17) than $(y_L^*, t_L^*, y_H^*, t_H^*)$—a contradiction. This establishes step 3.

Step 4: If constraint (ii) binds and $y_L \geq y_H$, then constraint (i) is necessarily satisfied.

If constraint (ii) binds then $t_H - t_L = 20(y_H - y_L)$. If $y_L \geq y_H$ then this implies that $t_H - t_L \geq 40(y_H - y_L)$, or, equivalently, $t_H - 40y_H \geq t_L - 40y_L$. Hence, constraint (i) is satisfied.

Given steps 1 to 4, we can simplify problem (23.F.17). In particular, we see that (y_L, t_L, y_H, t_H) is interim incentive efficient if and only if $y_L = 1$ and (t_L, y_H, t_H) solve

$$\begin{aligned} \underset{(t_L \in [0,1],\, y_H, t_H \in [0,1])}{\text{Max}} \quad & t_L - 20 \\ \text{s.t. (ii')} \quad & t_L - 20 = t_H - 20y_H. \\ \text{(iii)} \quad & t_H - 40y_H \geq \bar{u}_{1H}. \\ \text{(iv)} \quad & .2(50y_H - t_H) + .8(30 - t_L) \geq \bar{u}_2. \end{aligned} \tag{23.F.18}$$

Substituting from constraint (ii') of problem (23.F.18) for t_L in its objective function and in constraint (iv) we see that we can determine the optimal values of (y_H, t_H) by solving

$$\begin{aligned} \underset{(y_H \in [0,1],\, t_H)}{\text{Max}} \quad & t_H - 20y_H \\ \text{s.t.} \quad \text{(ii)} \quad & t_H - 40y_H \geq \bar{u}_{1H}. \\ \text{(iv')} \quad & 26y_H - t_H + 8 \geq \bar{u}_2. \end{aligned} \tag{23.F.19}$$

The solution for a given pair of values of $\bar{u}_{1H} \geq 0$ and $\bar{u}_2 \geq 0$ is depicted in Figure 23.F.1. The pairs (y_H, t_H) that satisfy constraints (ii) and (iv') of (23.F.19) lie in the shaded set. Also drawn are two level sets of the objective function $t_H - 20y_H$. The optimal pair (y_H, t_H) for these values of $\bar{u}_{1H}$ and $\bar{u}_2$ is the point labeled (y_H^*, t_H^*). The corresponding values of t_L and y_L in this interim incentive efficient social choice function are then $y_L^* = 1$ and $t_L^* = 20 + t_H^* - 20y_H^*$.

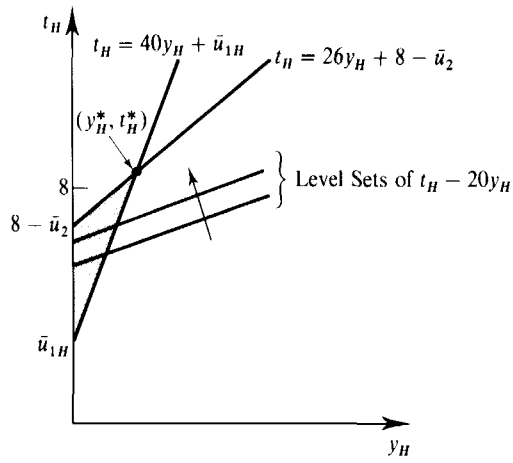


Figure 23.F.1 (left)
The optimal level of (y_H, t_H) in problem (23.F.19) for a given pair $(\bar{u}_H, \bar{u}_2) \geq 0$.

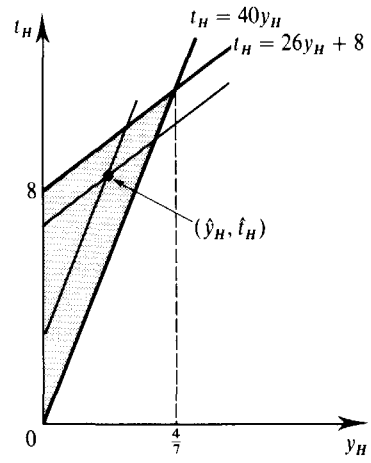


Figure 23.F.2 (right)
The shaded set contains those pairs (y_H, t_H) arising in interim incentive efficient social choice functions.

The shaded set in Figure 23.F.2, $\{(y_H, t_H)\colon y_H \in [0, 1],\ t_H \geq 40y_H,$ and $t_H \leq 26y_H + 8\}$, depicts *all* of the pairs (y_H, t_H) that arise in interim incentive efficient social choice functions. These are determined by performing the analysis in Figure 23.F.1 for each possible pair $(\bar{u}_{1H}, \bar{u}_2) \geq 0$ [a sample pair $(\hat{y}_H, \hat{t}_H)$ is also depicted in Figure 23.F.2]. Note that in any interim incentive efficient social choice function we have $y_H \leq 4/7$. As above, for each pair $(\hat{y}_H, \hat{t}_H)$ in this set, we can determine an interim incentive efficient social choice function $(\hat{y}_L, \hat{t}_L, \hat{y}_H, \hat{t}_H)$ by setting $\hat{y}_L = 1$ and $\hat{t}_L = 20 + \hat{t}_H - 20\hat{y}_H$.

Now, out of this set of interim incentive efficient social choice functions, which are ex ante incentive efficient? (Recall that the set of ex ante incentive efficient social choice functions is a subset of the interim incentive efficient set. Note also that although we now employ an ex ante welfare criterion, the set of participation constraints defining F^* continue to be the same interim participation constraints used above.) The buyer's and seller's ex ante expected utilities in an interim incentive efficient social choice function (y_L, t_L, y_H, t_H) are

$$U_1 = .8(t_L + 20(1 - y_L)) + .2(t_H + 40(1 - y_H))$$

and

$$U_2 = .8(30y_L - t_L) + .2(50y_H - t_H).$$

Since $y_L = 1$ and $t_L = 20 + t_H - 20y_H$ in any interim incentive efficient social choice function, these expected utilities can be written as functions of only (y_H, t_H) as follows:

$$U_1 = 24 + t_H - 24y_H, \qquad U_2 = 8 + 26y_H - t_H.$$

In Figure 23.F.3, for an arbitrary point $(\hat{y}_H, \hat{t}_H)$ in set $\{(y_H, t_H)\colon y_H \in [0, 1],\ t_H \geq 40y_H,$ and $t_H \leq 26y_H + 8\}$, the pairs (y_H, t_H) in the shaded set raise the ex ante expected utilities of both the buyer and the seller. As can be seen from this figure, no pair (y_H, t_H) that is not on the $t_H = 40y_H$ boundary of the set $\{(y_H, t_H)\colon y_H \in [0, 1],$ $t_H \geq 40y_H$, and $t_H \leq 26y_H + 8\}$ can be ex ante incentive efficient. Moreover, each pair (y_H, t_H) on this boundary *is* part of an ex ante incentive efficient social choice function. The set of (y_H, t_H) pairs in ex ante incentive efficient social choice functions