BA 513/STA 234: Ph.D. Seminar on Choice Theory
Professor Robert Nau
Spring Semester 2008

Readings for class #9: Social choice theory (updated March 10, 2008)

Primary readings:

1. "Social choices," chapter 6 of *Choices: An Introduction to Decision Theory* by Michael Resnik, 1987

2. "Social Choice Theory" by Amartya Sen, from *Handbook of Mathematical Economics, v. III*, 1987

3. Readings from *Rational Man and Irrational Society? An Introduction and Sourcebook*, edited by Brian Barry and Russell Hardin, 1982

   a. "Individual Preferences and Collective Decisions" by Brian Barry and Russell Hardin
   b. " Axiomatic Social Choice Theory: An Overview and Interpretation" by Charles Plott
   c. "Current Developments in the Theory of Social Choice" by Kenneth Arrow
   d. "Social Choice and Individual Values" by I.M.D. Little
   e. Welfare and Preference" by Kurt Baier
   f. "Utility, Strategy, and Social Decision Rules" by William Vickrey
   g. "Manipulation of Voting Schemes: A General Result" by Allan Gibbard
   h. "Epilog and Guide to Further Reading" by Brian Barry and Russell Hardin

4. "Incentives and Mechanism Design," chapter 23 from *Microeconomic Theory* by Mas-Colell, Whinston, and Green, 1995

5. See the web page of Vince Conitzer in Duke's Computer Science department for some interesting work on social choice mechanisms for artificial agents.

Our discussion of rational choice thus far has focused on choices made by individuals acting on their own behalf in games against nature or games against rational opponents. Social choice theory is concerned with a related but different problem, namely, how choices can or should be made on behalf of groups of individuals, either through institutional mechanisms such as voting or through the mediation of benevolent social planners. The literature of mathematical social choice dates back more than 200 years to the original work of Condorcet and Borda on voting systems in the 1780's, but modern social choice theory has its roots in the work of "Paretian" welfare economists such as Bergson and Samuelson in the 1930's, the reexamination of voting systems by Duncan Black in the 1940's, the introduction of axiomatic methods by Arrow, Nash,

and Harsanyi in the 1950's, and the more recent study of "game forms" by Gibbard and Satterthwaite, Maskin, and others. This week's readings present a survey of the classic work in social choice. The chapter by Resnik provides a textbook-level introduction, the long survey article by Sen provides more technical depth and a host of references, and the articles from the edited volume by Barry and Hardin include some classic (and very readable) papers with candid commentary by the editors.

Social choice theory generally uses the same basic tools and concepts as the rest of decision and game theory: individuals are described in terms of their *preferences* for the *consequences* they will receive under different *alternatives* that might be chosen. Those preferences are usually assumed to satisfy the axioms needed to ensure that they can be represented either by *ordinal utility functions* (as in consumer theory) or else by *cardinal utility functions* (as in expected utility theory). Thus, "consequentialist" social choice theory is concerned with the question of how the preferences of individuals can or should be used as data when making collective decisions.[1] The best-known results are all negative in character: Condorcet's parodox shows that majority voting leads to intransitive cycles in pairwise choices, Arrow's impossibility theorem shows that there is no entirely satisfactory rule for aggregating ordinal utilities, and Gibbard and Satterthwaite's theorem shows that all voting systems can be manipulated and that it is generally impossible to induce individuals to reveal their preferences truthfully in a collective choice situation. (Harsanyi's theorem on the aggregation of cardinal utilities stands out by comparison as a positive note, although it depends on very strong assumptions.) There are two conclusions that one might draw from these results. First, to some extent, the various paradoxes and impossibility theorems reflect genuine stresses and strains in a democratic society—real dilemmas that arise in public life. Second, and to a greater extent, they illustrate the limitations of social choice models that start from the assumption that the alternatives to be chosen and the preferences of the individuals are already determined and that the only "problem" to be solved is that of choosing among the given alternatives so as to best satisfy the given preferences.

The fundamental theorem of utilitarianism. Insofar as social choice theory seeks methods for making collective choices based on the preferences of individuals, and insofar as individual preferences are representable by ordinal or cardinal utility functions, the central problem in social choice theory can be framed as that of how to compare and aggregate the utilities of different individuals—essentially the same problem that was originally raised in a less formal manner by Bentham. As we have seen, utilitarianism fell into disrepute in the early 20th Century, but it was revived by von Neumann and Morgenstern's axiomatization of cardinal expected utility. Modern (i.e., post-vNM) utilitarianism comes in different flavors. Some authors—most notably Harsanyi—argue that cardinal utility is interpersonally comparable, or more precisely, that *differences* in cardinal utility between alternatives are interpersonally comparable. On this view, it is meaningful to ask whether Alice's gain in utility if society switches from policy $\boldsymbol{p}$ to policy $\boldsymbol{q}$ will be greater than Bob's loss in utility—i.e., whether the "total" cardinal utility of Alice and Bob is increased or decreased by a move from $\boldsymbol{p}$ to $\boldsymbol{q}$. Harsanyi proved the following

[1] There is also a non-consequentialist strand of social choice literature that focuses on the *processes and procedures* by which public choices are made. The latter strand of literature, which emphasizes individual rights and liberties more than the efficiency of outcomes, is exemplified by the work of James Buchanan in the 1950's and more recent work by Robert Nozick, Robert Sugden, Gaertner-Pattanaik-Suzumura and others. See "Individual Preference as the Basis of Social Choice" by Amartya Sen, in *Social Choice Re-examined*, edited by Arrow, Sen, and Suzumura (1997).

theorem: if (a) every individual has a cardinal utility function defined on lotteries over the same fixed set of social alternatives, and (b) a social planner also has a cardinal utility function defined over the same set of alternatives, and (c) the social planner's utility function satisfies a *Pareto condition* with respect to the individuals, then it follows that the social planner's utility function must be a (unique!) weighted sum of the individual utility functions. This result is the "fundamental theorem of utilitarianism" and, as you might expect, it can be proved by the same separating hyperplane argument that we have used to prove all the other fundamental theorems of rational choice. The Pareto condition states that if every individual weakly prefers $\boldsymbol{p}$ to $\boldsymbol{q}$, then society must weakly prefer $\boldsymbol{p}$ to $\boldsymbol{q}$, and if, in addition, at least one individual strictly prefers $\boldsymbol{p}$ to $\boldsymbol{q}$, then society must strictly prefer $\boldsymbol{p}$ to $\boldsymbol{q}$. To frame this condition in terms of utility functions, assume that there is a finite set of alternatives and that the objects of social choice are *lotteries* (i.e., probability distributions) defined over those alternatives. Assume that every individual satisfies the vNM axioms and the social planner does too. Then individual i's preferences are represented by a utility vector $\boldsymbol{u_i}$ such that lottery $\boldsymbol{p}$ is preferred to lottery $\boldsymbol{q}$ if and only if $\boldsymbol{p} \bullet \boldsymbol{u_i} \geq \boldsymbol{q} \bullet \boldsymbol{u_i}$. (When $\boldsymbol{p}$ is a probability vector, the vector product $\boldsymbol{p} \bullet \boldsymbol{u_i}$ is the expected utility of $\boldsymbol{p}$.) Let $\boldsymbol{u_0}$ denote the corresponding utility function for the social planner. In these terms, the Pareto condition requires that if $\boldsymbol{p} \bullet \boldsymbol{u_i} \geq \boldsymbol{q} \bullet \boldsymbol{u_i}$ for every individual i, then $\boldsymbol{p} \bullet \boldsymbol{u_0} \geq \boldsymbol{q} \bullet \boldsymbol{u_0}$, and if in addition $\boldsymbol{p} \bullet \boldsymbol{u_i} > \boldsymbol{q} \bullet \boldsymbol{u_i}$ for at least one individual i, then also $\boldsymbol{p} \bullet \boldsymbol{u_0} > \boldsymbol{q} \bullet \boldsymbol{u_0}$. Now consider the *open convex hull* of the vectors $\{\boldsymbol{u}_i, i > 0\}$, i.e., the set of all *positively weighted sums* of the individual utility vectors, and let this set be called U. By the separating hyperplane argument, exactly one of the following must be true: either (i) the vector $\boldsymbol{u_0}$ is contained in the convex set U, in which case the social planner's utility function is a positive weighted sum of the individual utility functions, or else (ii) there is a nontrivial hyperplane separating $\boldsymbol{u_0}$ from U. Let $\boldsymbol{v}$ denote the normal vector of the separating hyperplane, if one exists. This means that $\boldsymbol{v}$ must satisfy one of the following two conditions: either (a) $\boldsymbol{v} \bullet \boldsymbol{u_i} \geq 0$ for every $i > 0$ but meanwhile $\boldsymbol{v} \bullet \boldsymbol{u_0} < 0$, or else (b) $\boldsymbol{v} \bullet \boldsymbol{u_i} \geq 0$ for every $i > 0$, with $\boldsymbol{v} \bullet \boldsymbol{u_i} > 0$ for at least one i, but meanwhile $\boldsymbol{v} \bullet \boldsymbol{u_0} \leq 0$. (Intuitively, the individual utility vectors are all "above" the hyperplane whose normal vector is $\boldsymbol{v}$, while the social utility vector is "below" it.) Without loss of generality, such a vector $\boldsymbol{v}$ can be written as the difference of two probability vectors, i.e., $\boldsymbol{v} = \boldsymbol{p} - \boldsymbol{q}$.[2] Then condition (ii) of the separating hyperplane argument is precisely a violation of the Pareto condition: it says that there exist $\boldsymbol{p}$ and $\boldsymbol{q}$ such that every individual weakly prefers $\boldsymbol{p}$ to $\boldsymbol{q}$, while society does not, or at least one individual strictly prefers $\boldsymbol{p}$ to $\boldsymbol{q}$ (while everyone else at least weakly prefers $\boldsymbol{p}$ to $\boldsymbol{q}$) while the social planner does not. Hence, the social planner's utility function is a positive weighted sum of the individual utility functions if and only if the Pareto condition is not violated. *QED*

The following figures illustrate the geometry of Harsanyi's theorem in the case where there are three alternatives. The set of all differences between probability distributions over three alternatives is the set of all 3-vectors whose elements sum to zero and are less than unity in magnitude. Such vectors consist of all vectors $\boldsymbol{v} = (x, y, z)$ lying in the intersection of the cube

[2] To see that there is no loss of generality in writing $\boldsymbol{v} = \boldsymbol{p} - \boldsymbol{q}$, where $\boldsymbol{p}$ and $\boldsymbol{q}$ are probability distributions, note that without loss of generality it can be assumed that every utility vector is normalized so that its elements sum to zero, since they are unaffected by the addition of constants. Therefore, we can also add or subtract a constant from $\boldsymbol{v}$ without affecting any of the vector products $\boldsymbol{v} \bullet \boldsymbol{u_i}$, and in this way we can normalize $\boldsymbol{v}$ so that its elements, too, sum to zero. Furthermore we can scale $\boldsymbol{v}$ so that its positive and negative parts each sum to less than 1, and any such vector can be expressed as a difference of probability distributions.

defined by $-1 \le x \le 1$, $-1 \le y \le 1$, and $-1 \le z \le 1$ with the plane defined by $x+y+z = 0$. This intersection is a two-dimensional hexagon, as shown in Figure 1 below.

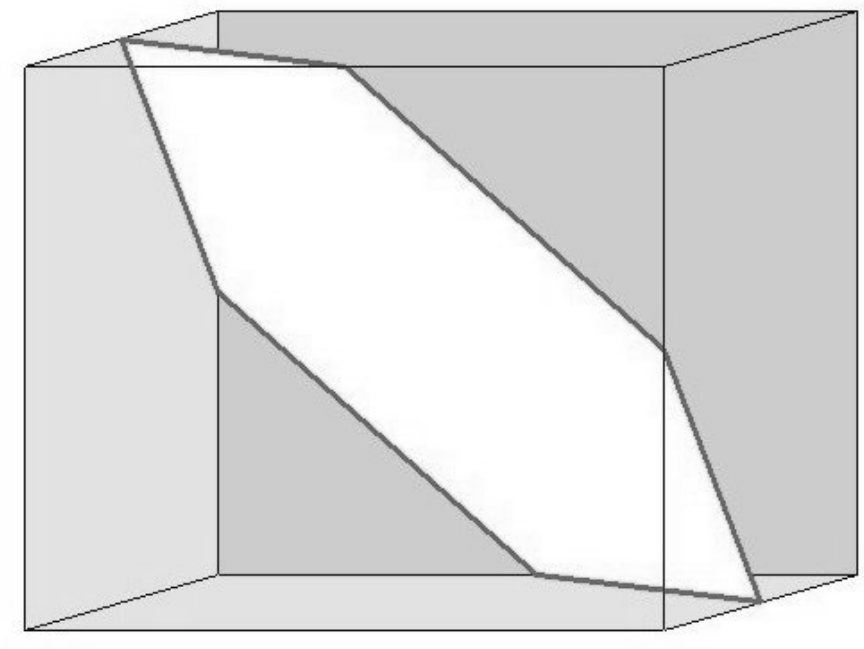

Figure 1. The cube represents the set of all 3-dimensional vectors whose *x*, *y*, and *z* coordinates are all between -1 and +1. The hexagon-shaped 2-dimensional set is the intersection of the cube with the plane defined by $x+y+z = 0$. All vectors that are differences between two probability distributions lie in the hexagon, since their elements sum to zero and are between -1 and +1. Without loss of generality, all utility functions over a set of 3 alternatives can also be represented by points in the hexagon.

The utility vectors of the individuals and the social planner can be plotted in the same hexagon by normalizing them so that their elements sum to zero and their maximum element has an absolute magnitude of unity. Thus, a typical utility vector extends from the center of the hexagon to a point on the boundary, as shown in Figure 2 below. Let $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ denote the utility vectors of individuals 1 and 2, respectively. Then individual 1 prefers $\boldsymbol{p}$ to $\boldsymbol{q}$ if $\boldsymbol{p} \bullet \boldsymbol{u}_1 \ge \boldsymbol{q} \bullet \boldsymbol{u}_1$, which is true if the vector $\boldsymbol{p} - \boldsymbol{q}$ lies above the line *AB* in the figure, whose normal vector is $\boldsymbol{u}_1$. Similarly, individual 2 prefers $\boldsymbol{p}$ to $\boldsymbol{q}$ if $\boldsymbol{p} - \boldsymbol{q}$ lies above the line *CD* in the figure, whose normal vector is $\boldsymbol{u}_2$. Thus, *both* individuals prefer $\boldsymbol{p}$ to $\boldsymbol{q}$ if $\boldsymbol{p} - \boldsymbol{q}$ lies in the area above both lines, which is the *light-shaded region* in the figure. The light-shaded region is called the "dual cone" generated by $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, while the "primal cone" of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, which is the set of convex combinations of them, is the *dark-shaded region*. . Meanwhile, letting $\boldsymbol{u}_0$ denote the utility vector of the social planner, the planner prefers $\boldsymbol{p}$ to $\boldsymbol{q}$ if $\boldsymbol{p} - \boldsymbol{q}$ lies above the line *EF* in the figure, whose normal vector is $\boldsymbol{u}_0$. The Pareto condition requires that the light-shaded area (i.e., the dual cone of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$) should lie strictly above the line *EF*, which (by the separating hyperplane argument) is true if and only if $\boldsymbol{u}_0$ lies in the interior of the primal cone generated by $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ (i.e., is a positive weighted average of them). Conversely, if $\boldsymbol{u}_0$ lay outside the primal cone of

$\boldsymbol{u_1}$ and $\boldsymbol{u_2}$, then some preferences shared by both individuals would not be shared by the social planner—i.e., the Pareto condition would be violated.

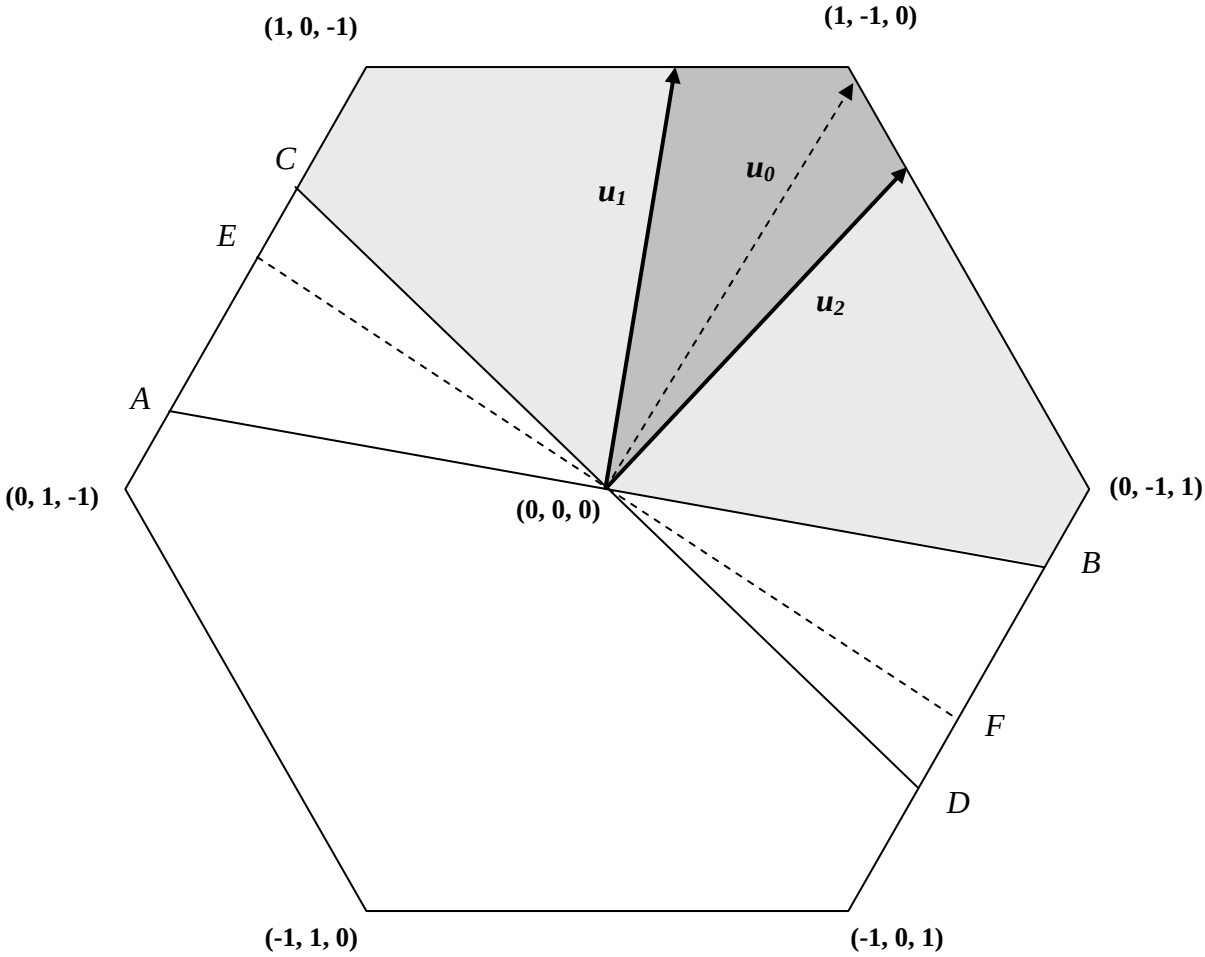

Figure 2. This is the hexagon-shaped region from the previous figure, redrawn in the plane. A utility function can be represented by a vector extending from the center to a point on the boundary. The set of all differences in probability distributions that are "preferred" under a given utility function is the set of all points lying on one side of a line drawn through the origin perpendicular to the utility vector. For example, the set of all differences in distributions that are preferred under u_1 is the set of points lying above the line *AB*.

Harsanyi's theorem appears to provide support for "additive utilitarianism," in which society prefers $\boldsymbol{p}$ over $\boldsymbol{q}$ if a weighted sum of all the individuals' utility differences between $\boldsymbol{p}$ and $\boldsymbol{q}$ is positive. Rather remarkably, under the conditions of Harsanyi's theorem, the relative weights assigned to different individuals are uniquely determined for any particular normalization of the individual and societal utility functions. Recall that a vNM utility function is unique only up to positive affine scaling. Harsanyi's theorem implies that if the utility function of an individual is rescaled, her weight in the social planner's utility function is changed in a reciprocal fashion, so she has the same influence regardless of the scaling of her own utility function.

Of course, this simple and powerful result is predicated on some very strong assumptions, most importantly on the *a priori* assumption that a societal cardinal utility function exists and is independently known. A much deeper problem in social choice theory, highlighted by Condorcet's paradox and Arrow's theorem, is whether and under what conditions a societal utility function—even an ordinal one—may be said to exist. As soon as a cardinal societal utility function is slapped on the table, the most difficult problem in social choice is assumed away. (The uniqueness of the utility weights in Harsanyi's theorem follows from the fact that the individual utility functions and the societal utility function are assumed into existence simultaneously instead of deriving the latter from the former.) The vNM axiom system that yields a cardinal utility function requires the objects of choice to be elements of a convex set of objective probabilistic lotteries, which is a highly abstract way of framing the choices available to an individual, and it becomes even more farfetched when applied to choices available to a society. (What would it mean for the social planner to choose an α chance of policy $\boldsymbol{p}$ and a $1-\alpha$ chance of policy $\boldsymbol{q}$?) Harsanyi argues, nevertheless, that interpersonal comparisons of cardinal utility differences are meaningful and practical. Such comparisons implicitly provide the rationale for redistributive social policies such as progressive taxation, in which a dollar is assumed to be worth more to a poor person than a rich one, or public health and safety initiatives in which government funds are allocated so as to yield the greatest social benefit in terms of quality-adjusted-life-years saved. Of course, the same argument can also be used to support positions on the other side of the aisle—e.g., that it is better to transfer wealth from poor fools to rich sophisticates who know better how to enjoy it, or to inflict suffering on some individuals in order to confer luxuries on others. It all depends on who is doing the social planning! This implication of additive utilitarianism is sometimes called the "repugnant conclusion." Harsanyi also uses a version of the *common prior assumption* to defend the additive utilitarian position. He proposes that, when participating in social decisions, individuals should imagine themselves to be behind a "veil of ignorance" in which they do not yet know their own "type"—i.e., who they are in society. From this hypothetical position of incomplete information, they should make the choices that would maximize the expected value of their utility, based on the distribution of types in the population, which is equivalent to maximizing the sum of everyone's utilities.

Standing in contrast to Harsanyi's school of additive utilitarianism there is a school of "leximin utilitarianism" championed by John Rawls (*A Theory of Justice*, 1971). Rawls argues, on moral grounds, that we should try to compare *levels* of welfare rather than differences, and that society's preferences should be lexicographically based on maximization of the welfare of the worst-off individual. In our earlier review of axiomatic utility theory, we saw that it is impossible to attach any meaning to absolute utility levels. Rawls therefore focuses on "primary goods" (basic necessities of life), rather than utilities, as the welfare measure whose level is compared between individuals. Rawls argues that people can roughly agree on the identity of the worst-off individual in any given social scenario and on whether he or she is better or worse off than the bottom-dwellers (who need not be the same individuals) in other scenarios that might be realized through changes in policy. Rawls rejects Harsanyi's notion of comparing welfare differences between individuals on the grounds that, unless and until the worst-off person is made better off, it makes no difference what happens to anyone else. Like Harsanyi, Rawls invokes a "veil of ignorance" argument to support his position, but he rejects the common prior assumption, arguing instead that when you are behind the veil you should imagine yourself in a

game against a malevolent opponent who is playing a minimax strategy against you. Therefore, if you don't yet know who you are, you should assume you will end up as the worst-off person instead of as the "average" person, and you should make social choices accordingly.

Both Harsanyi and Rawls assume that the problem of interpersonal welfare comparison can be solved uniquely, either because everyone is assumed to agree on the comparisons or because the comparisons are made by a representative social planner. Other authors have explored the idea that each individual might make his or her own interpersonal utility comparisons via the notion of "extended sympathy"—i.e., imagining onesself in someone else's place and judging whether the utility difference between *x* and *y* for you is greater or less than the utility between *z* and *w* for the other person. The fundamental problem of social choice then becomes the interpersonal aggregation of interpersonal utility comparisons. (Whew.) But when axioms are imposed on this sort of higher-level aggregation, Arrow's impossibility theorem rears its head: the only consistent aggregation schemes turn out to be dictatorial. So, in the end, the problem of interpersonal utility comparison is rather a muddle. Suzumura (1996) has described it as "the cloud over social choice theory" and (like Rawls and others) has suggested that it might be better to emphasize comparisons of more primitive and objectively quantifiable attributes such as primary goods, resources, or ability-to-function.

Arrow's impossibility theorem is the most famous result in social choice theory, but its significance for human affairs is still a matter for debate. Arrow considers the question of whether it is possible to construct a *social welfare function* (henceforth SWF), which is a *rule* by which the ordinal preferences of an arbitrary group of individuals can be aggregated to determine a social ordering of the same alternatives, in accordance to the following superficially reasonable axioms:

- O. Ordering: the social ordering should have the same properties as the individual orderings (e.g., completeness and transitivity) and should be determined only by the individual orderings
- U. Unrestricted domain: a social ordering should be determined for any logically possible specifications of individual preferences
- P. Pareto optimality: if everyone prefers *x* to *y*, then society should prefer *x* to *y*
- D. Non-Dictatorship: there is no individual whose preferences always prevail over those of all other individuals
- I. Independence of irrelevant alternatives: the social ordering of *x* and *y* should depend only on individual preferences between *x* and *y*, not preferences for any other alternatives

Arrow's theorem shows that this is impossible: no SWF can simultaneously satisfy all of these axioms. The proof (following Vickrey) proceeds in several ingenious steps. First, the notion of a *decisive set* is introduced. Define a set of individuals to be *decisive* for one alternative *x* over another alternative *y* if, whenever they all prefer *x* over *y*, society does too, when all other individuals have the opposite preferences. Then:

(i) Axioms O, U, I, and P imply that a set of individuals who are decisive for *x* over *y* are also decisive for all other pairs of alternatives, as follows:

Let set D be decisive for x over y. Suppose everyone in set D has $x > y > u$ while everyone else has $y > u > x$. Then $x > y$ must prevail, by the decisiveness of D. Meanwhile everyone agrees $y > u$, so $y > u$ must prevail by the Pareto rule, whence $x > u$ prevails by transitivity—even though only members of D have $x > u$ as individuals! Hence D is also decisive for x over u. Similarly:

$z > x > u$ in D, and $u > z > x$ elsewhere $\Rightarrow$ D is decisive for z over u

$z > u > w$ in D, and $u > w > z$ elsewhere $\Rightarrow$ D is decisive for z over w

(ii) There is always at least one decisive set, namely the set of all individuals.

(iii) Axioms O and U imply that any decisive set can be decomposed into two proper subsets, at least one of which is itself decisive, which eventually leads down to a decisive set of size 1—namely a dictator—in violation of axiom D, as follows:

Let D have proper subsets A and B, and let C be everyone else, with:

A: $x > y > u$,
B: $y > u > x$,
C: $u > x > y$

Since A∪B is decisive, $y > u$ must prevail. If also $x < y$ prevails, this must mean B is decisive for x over y. But if $x > y$ prevails, then $x > u$ must prevail by transitivity, in which case A is decisive. Gotcha!

In the 50+ years since Arrow first proved this result, various authors have pointed out that it isn't as surprising or as dismal as it might have appeared at first glance. Objections can be raised against most of the axioms, separately or in combination with each other, and against the whole enterprise of searching for a universal social welfare function. First, consider axiom O, which entails completeness and transitivity of both individual and social orderings. We have already seen that completeness is a dubious requirement (both normatively and empirically) when imposed on the preferences of an individual, and it is even more dubious when imposed on a group of individuals who are not of the same mind (unless they have had the opportunity to arbitrage-out their differences of opinion—but that is another story!). In any case, do we really need a complete social ordering of all the alternatives, or would it suffice to merely determine a "best" alternative or even a "good" alternative for the problem at hand? Would it be acceptable for society to occasionally be undecided, leaving some choices to be made by arbitrary or accidental tie-breaking rules? (Well, hopefully not by hanging chad…) The status of transitivity as a normative principle of rationality for individuals has been questioned by Peter Fishburn and Robert Sudgen, among others, and their arguments are even more compelling when applied to groups: if majority voting sometimes leads to intransitive cycles in pairwise comparisons, so what? Voters are not usually asked to make all possible pairwise comparisons when there are more than two candidates. Next, consider axiom U. Why should a SWF be required to operate on completely arbitrary individual preferences, no matter how perverse? Social norms, institutions, and evolutionary psychology may impose constraints or symmetries on individual preferences that could facilitate preference aggregation in some settings. Arrow observed that if preferences are "single peaked," a condition that Duncan Black had used earlier

to rationalize majority voting, it is possible to aggregate them in a way that satisfies all the other axioms. (Single peakedness is property that applies to situations in which there is some natural linear ordering of the alternatives, e.g., a left-to-right ordering of political candidates or a small-to-large ordering of amounts of money to be spent on a public project. Agents' preferences are single-peaked if they are ordered with respect to distance from the most-preferred alternative. For example, if the centrist candidate is most-preferred by a given voter, then that same voter should prefer the left-center candidate over the far-left candidate.) More generally, why should we let our social choices in *this* world be governed by considerations of what might have happened in some weirdly different hypothetical world? Axiom I, despite its seductive and value-laden title, has often been criticized for prohibiting the use of any data concerning *intensities* of preference between alternatives, which might otherwise provide a basis for making rational tradeoffs between the interests of different individuals. This axiom rules out otherwise-sensible preference aggregation methods based on scoring systems (e.g., point totals or weighted voting) or measures of cardinal utility (e.g., Harsanyi's theorem). It not only requires the social ordering to be determined from data on individual preferences: it requires the social ordering to be determined from *low quality data* on individual preferences. Maybe we should not be surprised that this turns out to be impossible. Even axiom D is not as uncontroversial as it might first appear: it is easy to imagine situations in which one individual perhaps ought to be given dictatorial discretion over some pairs of alternatives which affect her much more than they affect anyone else (e.g., whether to be ritually sacrificed). Finally, there is the question of how the axioms interact with each other. Each leverages the others, and the key steps in the proof of the theorem use a combination of two or more axioms to produce an extreme and surprising result—e.g., someone who has dictatorial discretion over any one pair of alternatives must have dictatorial discretion over all alternatives. In his critique of Arrow's theorem, I.M.D. Little states: "The conclusion, to my mind, is that it is foolish to accept or reject a set of ethical axioms one at a time. One must know the consequences before one can say whether one finds the set acceptable—which sets a limit to the usefulness of deductive techniques in ethics or in welfare economics."

Voting systems. Harsanyi's possibility theorem and Arrow's impossibility theorem assume that the preferences of the individuals in a society are somehow already known in great detail. In practice, the preferences of individuals in collective choice problems must be elicited or constructed through a mechanism such as voting. In the simplest situation, where there are only two alternatives, the most commonly used voting system is *majority voting*, in which every individual casts a single vote for his or her most-preferred alternative and the one with the most votes wins. Majority voting has much to recommend it in such situations: it is the only mechanism that has the desirable properties of *anonymity* (every voter is treated equally), *neutrality* (every alternative is treated equally), and *positive responsiveness* (if anyone's vote is changed, the outcome must change in the same direction, if at all). When there are three or more alternatives, more elaborate voting schemes must be considered. One possibility is to carry out majority voting between each *pair* of alternatives as a way of constructing society's binary preferences. If one alternative beats all its rivals in pairwise majority voting, it is called the *Condorcet winner* and is seemingly the socially preferred alternative. But social preferences elicited in this way need not be transitive, a phenomenon known as **Condorcet's paradox**. It is trivial to construct examples in which a majority prefers x over y, y over z, and z over x in pairwise comparisons. Recent work by Fuqua Ph.D. Ilia Tsetlin, in collaboration with Michael

Regenwetter and Bernard Grofman, shows that such cycles are unlikely to occur in practice if voter preferences are correlated to any degree. ("On the Probabilities of Correct or Incorrect Majority Preference Relations" by Tsetlin and Regenwetter, *Social Choice and Welfare* 20(2), 283-306, 2003, "Impartial Culture Maximizes the Probability of Majority Cycles" by Tsetlin, Regenwetter, and Grofman *Social Choice and Welfare* 21(3), 387-398.) The assumption of an "impartial culture"—i.e., uniformly random preferences—maximizes the probability of a cycle occurring, but this is obviously a highly unrealistic assumption, despite its frequent invocation in the social choice literature. Various methods other than pairwise majority voting can be used to determine an unambiguous winner (or more than one winner, if necessary) in situations involving three or more alternatives: *plurality voting*, *run-off voting*, *approval voting* (in which voters indicate all candidates that meet with their approval), the *Borda count* (in which voters assign points to candidates according to their preference rank), and the *single transferable vote* system (in which voters rank their top few candidates and excess votes received by a winning candidate beyond a quota needed for election are transferred to lower-ranked candidates in a proportional manner).

It has long been known that, in situations with three or more alternatives, most voting systems are subject to manipulation—that is, the voters may have incentives to misrepresent their true preferences in order to improve the chances that their most-preferred alternative will be selected. For example, in a 3-candidate race under plurality voting, supporters of a third-party candidate may have incentives to throw their support to the more preferred of the two major-party candidates (notwithstanding the Nader/Gore/Bush example). Under the Borda count, voters may have incentives to falsely label their second choice as their last choice to help ensure the election of their first choice. Under approval voting, voters may have incentives to approve of only their first choice, even if their second choice is privately acceptable. In the early 1970's, it was proved independently by Gibbard and Satterthwaite that this is true in general: all voting systems are manipulable. Gibbard actually proves a more general result, namely that in any "game form" (a general structure for a noncooperative game into which arbitrary preferences for outcomes can be plugged) at least one player fails to have a dominant strategy under some specifications of preferences. Any voting system that selects a winner by a deterministic function of voter responses is a special case of a game form, and if the voting system were non-manipulable, it would be a dominant strategy for every voter to reveal her true preferences. Gibbard's theorem shows that such a voting system is impossible. (Gibbard's original paper, minus the proof of the theorem, is one of the included readings from Barry and Hardin's book. The proof uses some of the same tricks as Arrow's impossibility theorem.)

Mechanism design. Since the publication of Gibbard and Satterthwaite's theorem, a considerable literature has grown up around the related topic of *implementation* and *mechanism design*—i.e., the construction of decentralized social choice mechanisms whose noncooperative equilibria yield efficient or otherwise desirable allocations of resources. A mechanism is a game form in which each individual sends a message from some set of possible messages, which ostensibly reveals private information, and the social outcome is then determined from the messages by a suitable rule. Ideally, for any specification of individual preferences and information, the induced game would have a unique (and transparent) Nash equilibrium that would yield the desired allocation. But here, too, impossibility theorems abound: it is generally impossible to design mechanisms which simultaneously (i) yield Pareto efficient allocations, (ii)

are incentive compatible (i.e., encourage individuals to reveal their private information truthfully), (iii) are individually rational (i.e., guarantee that individuals will be at least as well off by participating as by not participating), (iv) are valid for general preferences (e.g., nonlinear utility for money), and (v) have unique equilibria that are transparent and well-behaved (e.g., continuous).

Dominant-strategy vs. Bayesian implementation. Ideally, the mechanism would have the property that every player would have a *dominant strategy*, in which case the solution would be robust against imprecision or disagreement in the prior distribution over types. If the players do not have dominant strategies, then they must consider the distribution over other players' types when choosing their own strategies, in which case the *common prior assumption* is invoked and the solution concept of *Bayesian Nash equilibrium* is used. Nash implementation is less desirable for a number of reasons. First, it depends sensitively on common knowledge of preferences or distributions over preferences. Second, equilibria may not be unique (and usually aren't). Third, unless there is some kind of iteration of the choice process, it's hard to see how players would converge to a Nash equilibrium. Finally, if the solution is not continuous (as it sometimes is not), convergence to equilibrium may be implausible in any case.

Conceivably, a mechanism for implementing a social choice could require players to send signals from an arbitrary message space, which would then be decoded and processed in some complicated way. Fortunately this is not necessary. According to the **revelation principle**, if implementation is possible at all, then there is a mechanism in which every player's strategy is merely to truthfully report her preferences. There are two versions of the principle:

- If there is dominant strategy equilibrium, then there is a *truthful* dominant strategy equilibrium in which each player's strategy consists of truthful revelation of her type.

- If there is a Bayesian Nash equilibrium, then there is a *truthful* Bayesian Nash equilibrium in which each player's strategy consists of truthful revelation of her type.

Proof: for any equilibrium strategy profile specified as a mapping from players' types to a general message space, a mediator could choose the same outcome based on the players' reports of their types, in which case truth-telling would be optimal.

These results simplify the characterization of equilibria: it is only necessary to check to see whether truth-telling is optimal. (Interestingly, the coefficient vectors of incentive constraints often can be interpreted as payoff vectors of acceptable gambles that reveal the players' preferences, and the requirement of equilibrium is that, in light of those gambles, the players strategies should not lead to arbitrage. See my paper on "Joint Coherence in Games of Complete Information," *Management Science* 1992.)

An important special case of dominant strategy implementation is the **Groves-Clarke mechanism.** Assume that a single project must be selected from a list of projects, and players receive private benefits from different projects according to their types. Assume that the benefits are measured in money and players have linear utility for money. The mechanism is as follows:

- Let the players report their types (i.e., the private benefits they would receive from each project), and let an *efficient* (total benefit maximizing) project be selected on that basis.

- In addition, let each player receive a *transfer payment* equal to the sum of the other players' benefits, which represents the *externality* that her reported type imposes on the other players.

The rationale for the transfer payment is that if player *i* were to change her reported type from t_i to t_i^* , the change in her transfer payment would be *zero* if this did *not change* the project selection, holding the other players' reported types fixed. However, if the change in player *i*'s reported type *did* cause the project selection to change, then the change in her transfer payment would equal the change in total benefits to the other players. In particular, if player *i*'s change imposes a *negative externality* on the other players, her own transfer payment is reduced by exactly the same amount. Thus, agent *i* would be required to "internalize the externality."

The advantage of the Groves-Clarke mechanism is that it implements the selection of a total-benefit-maximizing project in dominant strategies. The drawbacks are that (i) it is not necessarily ex post efficient ("budget balancing")—i.e., the sum of transfer payments may be *negative*—and (ii) it only works with quasi-linear utility—i.e., it does not allow for "income effects". The first drawback is, alas, a special case of a more general impossibility result: according to the **Green-Laffont theorem,** under the unrestricted-domain assumption, there is no social choice function that is truthfully implementable in dominant strategies and is ex post efficient.

Another important impossibility result is the **Myerson-Satterthwaite theorem**, which states that, in a bilateral trade setting, there is no Bayesian incentive compatible social choice function ("trading rule") that is ex post efficient and gives every buyer type and every seller type nonnegative expected gains from participation. Hence, even the most elementary problem of microeconomics—namely how two consumers should exchange apples for bananas—cannot be finessed by clever mechanism design.

Concluding comments. So, the holy grail of social choice theory—a universal mechanism for making collective choices on the basis of individual preferences—has not been found, although some success has been achieved in particular applications by imposing restrictions on preferences and/or weakening some of the desiderata for ideal mechanisms. Should we be surprised or discouraged by these results? There are several grounds for concluding that we shouldn't.

First of all, as we have seen repeatedly, individual preferences generally do not provide enough information to uniquely determine the outcomes of even the simplest interactions between two or more individuals—e.g., haggling over an exchange of apples for bananas or playing a coordination game such as battle-of-the-sexes. Intuitively, other kinds of psychological variables also play a role—e.g., the individuals' relative degrees of power, patience, negotiating skill, or imagination. It may be asking too much to try to determine the outcomes of social choice problems on the basis of individual characteristics alone: social outcomes may be, in some cases, *fundamental measurements* of interpersonal variables.

Second, it is unduly restrictive to assume that social choices are made under conditions of common knowledge. Intuitively, the function of markets and other decentralized mechanisms is to help boundedly-rational individuals grope their way towards common understanding and control of a complex economic system. An important function of markets and other mechanisms is to make things common knowledge *which were not formerly common knowledge*, and agents who are especially "alert" or who otherwise have "uncommon knowledge" can earn rents above and beyond what they would be entitled to on the basis of their personal preferences for consumption and their private information gleaned (only) from sources with commonly-known statistical properties. Uncommon knowledge might mean knowing more about your opponent than he knows about you or being a better judge of "market psychology." If an ideal mechanism existed, it would deprive the "uncommon" agents of their natural competitive advantage. (Of course, in some cases this is precisely what is intended: the function of some real mechanisms is to protect individuals from the savagery of the "law of the jungle.")

Third, the question of how to choose among "given" alternatives on the basis of "given" preferences may not be the most important question to ask. The processes by which preferences are formed and alternatives are created or discovered are also of interest: in a complex environment, it is arguably more important to have good alternatives and to be guided by well-structured values than to have a choice function whose cardinal virtue is that it would also be suitable for choosing among bad alternatives under perverse values. A choice among evils is, after all, evil. On this view, a voting system or mechanism is merely the end stage of a *social decision process* in which issues are framed, rules are written, special interests and media forces are mobilized, beliefs and preferences are constructed, candidates are recruited and marketed, and policy options are imaginatively (or unimaginatively) created. Perhaps, then, the canvas should be enlarged to include what happens in the earlier stages.