13.4 Harsanyi's utilitarian theorems

John C. Harsanyi, a well-known economist and fellow Nobel Prize laureate, has proposed a radically different approach to social decision making. Briefly put, he defends a utilitarian solution to the problem of social choice, according to which the social preference ordering should be entirely determined by the sum total of individual utility levels in society. For example, if a single individual strongly prefers a high tax rate over a low tax rate, and all others disagree, then society should nevertheless prefer a high tax rate given that the preference of the single individual is sufficiently strong. Here is another equally surprising utilitarian conclusion: If a doctor can save five dying patients by killing a healthy person and transplant her organs to the five dying ones – without thereby causing any negative side-effects (such as decreased confidence in the healthcare system) – then the doctor should kill the healthy person.

In order to defend his utilitarian position, Harsanyi makes a number of assumptions. First of all, he rejects Arrow's view that individual preference orderings carry nothing but ordinal information. On Harsanyi's view, it is reasonable to assume that individual preference orderings satisfy the von Neumann and Morgenstern axioms for preferences over lotteries (or some equivalent set of axioms). This directly implies that rational individuals can represent their utility of a social state on an interval scale.

> **Individual rationality:** All individual preference orderings satisfy the von Neumann and Morgenstern axioms for preferences over lotteries. (See Section 5.2.)

To render this assumption more intelligible, we may imagine hypothetical lotteries over alternative social states. Suppose, for instance, that you are offered a lottery ticket that entitles you to a fifty-fifty chance of either living in a society with a high tax rate, or in one with a low tax rate. You are then asked to compare that lottery with a 'lottery' that entitles you to live in a society with a moderate tax rate with full certainty. Which lottery would you prefer? Given that your preferences over all social states, and all lotteries over social states, satisfy von Neumann and Morgenstern's axioms (or some equivalent set of axioms) it follows that your preferences can be represented by a utility function that measures your utility on an interval scale.

The next condition proposed by Harsanyi is a bit more abstract. Briefly put, Harsanyi asks us to imagine an individual (who may or may not be a fellow citizen) who evaluates all social states from a moral point of view. Let us refer to this individual as *the Chairperson*. If the Chairperson is a fellow citizen, then he has two separate preference orderings, viz. one personal preference ordering over all states that reflects his personal preference ordering, as well as a separate preference ordering over the same set of social states that reflects the social preference ordering. It is helpful to think of the Chairperson as an individual who is chosen at random from the entire population, and who is explicitly instructed to state two parallel preference orderings, viz. a personal preference ordering and a social one. As Harsanyi puts it, the social preference ordering is the preferences the Chairperson "exhibits in those – possibly quite rare – moments when he forces a special impartial and impersonal attitude, i.e. a *moral* attitude, upon himself" (Harsanyi 1979: 293). Of course, we do not yet *know* what the social preference ordering looks like, but Harsanyi shows that we can find out surprisingly much about it, given that it fulfils a number of structural conditions. Harsanyi's research question can thus be formulated as follows: What can be concluded about the Chairperson's social preference ordering, given that it fulfils certain structural conditions?

Before answering this question, we must of course clarify the structural conditions Harsanyi impose upon the Chairperson's social preference ordering. Consider the following condition:

> **Rationality of social preferences:** The Chairperson's social preference ordering satisfies the von Neumann and Morgenstern axioms for preferences over lotteries.

In order to assess the plausibility of this condition it does not suffice to ask, "What would *I* prefer in a choice between a lottery that gives us state *a* or *b*, and a lottery that gives us *c* or *d*?" In order to assess the new condition we must rather ask, "What would *the Chairperson* prefer in a choice between a lottery that gives us state *a* or *b*, and a lottery that gives us *c* or *d*?" Of course, it might be very difficult to answer such questions. However, note that Harsanyi's theorems will go through even if we are not able to tell *what* the Chairperson would prefer. All that matters is that we somehow know that the Chairperson's preferences, whatever they are, conform to the structural conditions proposed by von Neumann and Morgenstern.

The third condition proposed by Harsanyi is the Pareto condition.

Pareto: Suppose that a is preferred to b in at least one individual preference ordering, and that there is no individual preference ordering in which b is preferred to a. Then, a is preferred to b in the Chairperson's social preference ordering. Furthermore, if all individuals are indifferent, then so is the Chairperson in his social preference ordering.

The three conditions stated above imply that the Chairperson's social preference ordering must be a weighted sum of the individual preference orderings, in which the weight assigned to each individual preference ordering represents its moral importance relative to the others. In order to show this, it is helpful to introduce a slightly more technical vocabulary. From individual rationality it follows that individual preference orderings can be represented by utility functions that measure utility on an interval scale, and from rationality of social preferences it follows that the same holds true of the social preference ordering. Let $u_i(a)$ denote individual i's utility of state a, and let $u_s(a)$ denote the utility of a as reflected in the Chairperson's social preference ordering. Furthermore, let α be a real number between 0 and 1. Then,

Theorem 13.3 (Harsanyi's first theorem) Individual rationality, rationality of social preferences and Pareto together entail that:

$$u_s(a) = \sum_{i=1}^{n} \alpha_i \cdot u_i(a) \quad \text{with} \quad \alpha_i > 0 \text{ for } i = 1, \ldots, n \tag{1}$$

This theorem tells us that society's utility of state a is a weighted sum of all individuals' utility of that state. A proof will be given in Box 13.3. Meanwhile, note that the theorem does not guarantee that every individual preference ordering will be assigned the same weight. The theorem merely guarantees that each individual preference ordering is assigned *some* weight. However, utilitarians typically argue that all individual preference orderings should be assigned the *same* weight. Harsanyi thinks he can solve this problem by introducing a further assumption, which he formulates as follows.

Equal treatment of all individuals: If all individuals' utility functions $u_1, \ldots, u_n$ are expressed in equal utility units (as judged by the Chairperson, based on interpersonal utility comparisons), then the Chairperson's social utility function u_c must assign the same weight to all individual utility functions.

By adding this assumption to the previous ones, the following utilitarian conclusion can be proved.

Box 13.3 Proof of Harsanyi's theorem

We shall prove both theorems simultaneously, i.e. Theorems 13.3 and 13.4, by showing that $u_s(a) = \sum_{i=1}^{n} u_i(a)$ for every social state a. Without limiting the scope of the theorems, we stipulate that all individuals use a 0 to 1 utility scale, and that the Chairperson's social utility function starts at 0. (Its upper limit may of course exceed 1.) The utility numbers assigned by the individuals to a social state can be represented by a vector, i.e. by a finite and ordered sequence of real numbers. For instance, in a society with three individuals the vector [1/3, 0, 1] represents a state in which the first individual assigns utility 1/3 to the state in question, whereas the two others assign utility 0 and 1, respectively. Let us refer to such vectors as *state vectors*, and let the term *social utility number* refer to the numerical value of the Chairperson's social utility function for a state vector. Now consider the following lemma.

> **Lemma 1** Each state vector corresponds to one and only one social utility number.

By definition, a state vector corresponds to a social state. From *rationality of social preferences* it follows that the Chairperson's social utility function assigns some number to each social state. Hence, there is at least one social utility number that corresponds to each state vector. We also need to show that there cannot exist more than one such number. Of course, two or more social states could be represented by the same state vector (if all individuals are indifferent between them), so let us suppose for *reductio* that the Chairperson's social utility function assigns different social utility numbers u and v to two different social states with the same state vector. Now, since a single state vector can represent several social states just in case every individual is indifferent between the social states, it is helpful to apply *Pareto*: Since all individuals must be indifferent between the social states represented by u and v, it follows that the Chairperson must also be indifferent; hence, u and v are the same social utility numbers.

Lemma 1 directly entails that social utility is a function of state vectors. (This is trivial: Since each state vector corresponds to exactly one social utility number, it must be possible to capture this relationship by a function.) Hence, it holds that

$$u_s(a) = f[u_1(a), \ldots, u_n(a)] \quad (1)$$

Since (1) holds for all social states a, this equation can be abbreviated as $u_s = f[u_1, \ldots, u_n]$. It remains to show that $u_s = u_1 + \cdots + u_n$. To start with, consider the following claim:

$$kf[u_1, \ldots, u_n] = f[ku_1, \ldots, ku_n], \textit{where } 1 \geq k \geq 0 \tag{2}$$

Equation (2) says that it does not matter if we first multiply all individual utilities by a constant k and then apply the function f to the new state vector, or multiply k by the social utility number corresponding to the state vector. In the present exposition, Equation (2) will be accepted without proof. (For proof, see the proof of von Neumann and Morgernstern's theorem in Appendix B.) Now consider the state vector in which all individuals assign the number 0 to the state in question. Clearly, $u_s([0, \ldots, 0]) = 0$, because *Pareto* guarantees that every other state vector will be ranked above this vector, and hence assigned a number higher than 0. Next consider all unit vectors $[1, 0, \ldots, 0]$,$[0,1, \ldots, 0]$, and $[0, 0, \ldots, 1]$, in which exactly one ranks the state in question as the best one. Because of *equal treatment of all individuals*, u_s must assign the same social utility number to all unit vectors; let us stipulate that the number in question is 1.

In what follows, we only consider the case with a society that has two individuals. The case with societies of three or more individuals is analogous. Let u_1 and u_2 be fixed, and let L be a lottery that yields the social states $[u_1, 0]$ and $[0, u_2]$ with equal chances. From von Neumann and Morgenstern's expected utility theorem it follows that:

$$u_s(\mathrm{L}) = (1/2)u_s([u_1, 0]) + (1/2)u_s([0, u_2]) \tag{3}$$

By applying (2) to (3) we get:

$$u_s(\mathrm{L}) = u_s(1/2[u_1, 0] + 1/2[0, u_2]) \tag{4}$$

Note that each *individual's* expected utility of the lottery $1/2[u_1, 0] + 1/2[0, u_2]$ is $1/2u_1$ and $1/2u_2$, respectively. Hence, because of (1) it must also hold true that:

$$u_s(L) = f[(1/2)u_1, (1/2)u_2] \tag{5}$$

By applying (2) we get:

$$u_s(\mathrm{L}) = (1/2)f[u_1, u_2] \tag{6}$$

By applying (2) again we also find that:

$$u_s([u_1, 0]) = u_1 u_s([1, 0]) = u_1 \quad (7)$$

$$u_s([0, u_2]) = u_2 u_s([0, 1]) = u_2 \quad (8)$$

By combining (7) and (8) with (3) we obtain:

$$u_s(L) = (1/2)u_1 + (1/2)u_2 \quad (9)$$

Now we are almost done. We just have to put (9) and (6) together:

$$(1/2)f[u_1, u_2] = (1/2)u_1 + (1/2)u_2 \quad (10)$$

Hence,

$$f[u_1, u_2] = u_1 + u_2 \quad (11)$$

From (11) and (1) it follows directly that $u_s = u_1 + u_2$, which completes the proof for a society with two individuals. □

Theorem 13.4 (Harsanyi's second theorem) Given equal treatment of all individuals, the coefficients in Harsanyi's first theorem will be equal:

$$\alpha_1 = \cdots = \alpha_n \quad (2)$$

At this point it is natural to ask if Harsanyi's theorems are as powerful as they look. Has he really *proved* that society ought to distribute its resources according to utilitarian principles? Well, as shown above, the theorems do follow from the premises. Furthermore, Harsanyi's result does not violate Hume's law, according to which ethical 'ought-statements' cannot be derived from premises comprising no such ethical 'ought-statements'. The Pareto is an ethical premise, which Harsanyi uses for bridging the gap between rationality and ethics. The condition of *equal treatment of all individuals* also has some ethical content. So in one sense, Harsanyi really gives a proof of utilitarianism.

That said, no proof is better than the premises it is based upon. In Harsanyi's case, the most dubious premise is *equal treatment of all individuals*. This condition only makes sense if one believes that interpersonal comparisons of utility are possible. As pointed out above, this has been questioned by many

scholars. As such, the condition does not explain how interpersonal comparisons of utility could be made; it just *presupposes* that they are somehow possible. Furthermore, one may also question the normative content of *equal treatment of all individuals*. Why should everybody be treated equally? This is a substantial ethical question, that Harsanyi (and other utilitarians) ought to argue for, and not take for granted. For example, many ethicists would surely argue that some people ought to be treated better than others (i.e. $\alpha_i > \alpha_j$ for some individuals i and j), simply because they deserve it, or have certain rights that may not be violated, or are more virtuous, etcetera. My personal view is that the condition of *equal treatment of all individuals* is far too strong. However, even if correct, Harsanyi's first theorem (which does not rely on this condition) is still interesting, because it shows that the social utility function has to be additive. This indicates that *some* consequentialist ethical theory has to be accepted by anyone who is rational, at least as long as one thinks Harsanyi's premises make sense.

Exercises

13.1 A group of four people is about to select one out of six possible social states, a, b, c, d, e or f. (a) Which state will be selected if the decision is based on a traditional voting procedure? (Does the result depend on how the voting procedure is set up, given that all people get one vote each?) (b) Which state would be selected by the maximin rule? (It prescribes that society should prefer a state in which the worst-off person is as well off as possible.)

Anne: $a \succ b \succ c \succ d \succ e$
Bert: $b \succ a \succ c \succ e \succ d$
Carl: $a \succ c \succ d \succ b \succ e$
Diana: $a \succ b \succ c \succ d \succ e$

13.2 It follows from Arrow's impossibility theorem that social decisions cannot be based on majority voting procedures. (a) Exactly what is Arrow's criticism of majority voting? (b) Do you find his criticism convincing?

13.3 The Pareto principle entails that if we can improve the situation for the richest person in society, without thereby making things worse