Welfare

Bruno Salcedo*

Winter 2020

Before analyzing how institutions behave, let us ask how we would like them to behave. The answer can be straightforward in some cases. If a proposed public policy makes every member of society better off and it conforms to accepted notions of morality, then there is no reason for debate. Most reasonable people would agree that such policies are desirable from a social perspective. However, many real-life policies benefit some people while hurting others. We need a social criterion to evaluate policies that affect different individuals with (typically) different preferences.

These notes analyze different social criteria to rank alternative policies. The criteria we consider do not arise from exogenous moral values. Instead, they are derived from the individual values of the members of society. In that sense, these notes are about aggregating individual preferences into notions of social welfare.

The notes are structured as follows. Section 1 formalizes the problem of aggregating preferences and introduces the concept of *social welfare function*. Section 2 presents the classic result from Arrow (1950, 1951), which teaches us that there is no perfect way to aggregate preferences. Section 3 introduces the most popular criterion in Economics, which is due to Pareto (1909). It has the disadvantage of being incomplete. Sections 4–6 introduce ways of completing the Pareto criterion in settings with monetary transfers, or with cardinal measures of utility.

Reading these notes is required for both Intermediate Microeconomics II (2261) and Microeconomics II (9602). Those enrolled in 2261 may skip all the formal proofs from the appendix. Those enrolled in 9602 are responsible for learning all the material.

*Department of Economics, Western University · brunosalcedo.com · bsalcedo@uwo.ca

1. Individual Values and Social Choices

The essential components of a *social choice problem* are a set of individuals $I = \{1, \ldots, n\}$ and a set of alternatives $\mathscr{A}$. A typical individual is denoted by i, and typical alternatives are denoted by A, B, C, ... Let $\mathscr{P}$ be the set of all rational (i.e., complete and transitive) preference relations over alternatives.[1] A *preference profile* is a list $(\succcurlyeq_1, \ldots, \succcurlyeq_n) \in \mathscr{P}^n$, consisting of a preference relation for each individual. We would like to construct a social preference over alternatives. It is a principle to decide which outcomes are "good for society." It should be determined by the preferences of the individuals.

Definition 1 A *social welfare function* (SWF) is a mathematical function that takes as input a profile of individual preferences $(\succcurlyeq_1, \ldots, \succcurlyeq_n)$ and produces as output a social preference relation $\succcurlyeq^*$.

There are many different ways to aggregate individual preferences. To see this, let us consider a few examples of SWFs. One way to rank alternatives is to use a *dictatorial* SWF. Choose a fixed individual and name them the *dictator*. According to this SWF, social preferences should always match those of the dictator. That is, alternative A is socially preferred to alternative B if and only if the dictator prefers A to B, regardless of the preferences of other individuals.

If there are only two alternatives, society can rank them using the *simple majority* SWF. According to this SWF, the socially preferred alternative is the one preferred by at least half of the individuals. Simple majority is the criterion more often associated with democracy. There are many different ways to generalize it to settings with more than two alternatives.

The *plurality* SWF is one way to generalize simple majority. Count the number of votes that each alternative would receive in a general election (assuming that individuals vote sincerely for their top choice). The plurality criterion ranks alternatives according to these counts. For example, consider the preferences in Table 1. In a general election, alternatives A, B, and C would get 3, 2, and 1 votes, respectively. Hence, the plurality SWF ranks the alternatives as $A \succ^*_{\text{pl}} B \succ^*_{\text{pl}} C$.

[1] A preference relation $\succcurlyeq$ on $\mathscr{A}$ is complete if every pair of alternatives can be compared, i.e., either $A \succcurlyeq B$ or $B \succcurlyeq A$ or both. It is transitive if $A \succcurlyeq C$ whenever $A \succcurlyeq B$ and $B \succcurlyeq C$. Given a preference relation $\succcurlyeq$, I denote *strict preference* by $\succ$ and *indifference* by $\sim$. That is, $A \succ B$

1	2	3	4	5	6
A	A	A	B	B	C
B	B	B	A	C	B
C	C	C	C	A	A

Table 1 – Each column specifies the preferences of an individual, e.g., individual 1's preferences are $A \succ_1 B \succ_1 C$.

Condorcet (1785) championed a different method to generalize simple majority. According to the *Condorcet* SWF, alternative A is preferred to alternative B if and only if it would defeat B in a two-candidate election. That is, if and only if more than half the individuals rank A over B. In the example from Table 1, individuals 1, 2, and 3 prefer A to B, while individuals 4, 5, and 6 prefer B to A. Also, more than half of the people prefer A to C, and more than half prefer B to C. Hence, according to the Condorcet SWF, $A \sim^*_{\text{co}} B \succ^*_{\text{co}} C$.

Borda's SWF is another generalization of simple majority named after Borda (1781). According to this SWF, each alternative is awarded points by each individual. The number of points awarded to a given alternative by a given individual equals the number of alternatives that they ranks equal or below than than the alternative in question. Finally, alternatives are ranked by the total number of points received. Table 2 shows one way to compute the Borda points for the example from Table 1. For example, alternative A is awarded 3 points from individual 1 because $A \succ_1 B \succ_1 C$, and only 2 points from individual 4 because $B \succ_4 A \succ_4 C$. The total counts in the last column imply that $B \succ^*_{\text{bo}} A \succ^*_{\text{bo}} C$.

alternative	1	2	3	4	5	6	*total*
A	3	3	3	2	1	1	**13**
B	2	2	2	3	3	2	**14**
C	1	1	1	1	2	3	**9**

Table 2 – Borda counts for each alternative given the preferences from Table 1.

means that $A \succcurlyeq B$ and $B \not\succcurlyeq A$; while $A \sim B$ means that $A \succcurlyeq B$ and $B \succcurlyeq A$.

2. A Difficulty in the Concept of Social Welfare

In the previous section, we learned that there are different ways to rank social alternatives. Now, we will discuss what properties make a good SWF. Arrow (1950, 1951) proposed four minimal properties that most people find reasonable and appealing. Surprisingly, he found that it is impossible to construct a SWF satisfying all four properties simultaneously. This finding is known as *Arrow's Impossibility Theorem*. It teaches us that aggregating individual preferences is a difficult task and, in general, there is no perfect solution.

Condition U (Unanimity) If every individual strictly prefers A to B, then $A \succ^* B$.

Condition U requires that, if all individuals agree on the ranking of two alternatives, then so should Society. All the rules we have proposed satisfy this condition. Indeed, suppose that all individuals prefer A to B. Then, A would receive more votes than B in a general election. A would not lose a two-alternative election against B. And A would be awarded more Borda points than B by every single individual.

Condition UD (Universal Domain) $\succcurlyeq^* \in \mathscr{P}$ is well defined for *any* preference profile $(\succcurlyeq_1, \ldots, \succcurlyeq_n) \in \mathscr{P}^n$.

In some specific settings, it might be reasonable to impose restrictions on the preferences of individuals. For example, one might assume that most people prefer to reduce the time it takes to complete an infrastructure project *caeteris paribus*. However, at our level of abstraction, we should allow for any possible preference profile. And, if individual preferences exist, then so should Society's. UD actually requires more than that. It also requires social preferences to be rational, i.e., complete and transitive.[2]

Condorcet's SWF fails UD, because it can lead to intransitive social preferences with cycles of strict preference. Consider for instance the preferences from Table 3. Note that individuals 1 and 2 prefer A to B. Individuals 1 and 3 prefer B

[2]Failures of rationality of social preferences are problematic for the same reasons that failures of rationality can be problematic for individuals. I will not go into these reasons because I assume that you learned about them in either Econ 2260 or Econ 9601.

1	2	3
A	C	B
B	A	C
C	B	A

Table 3 – Instance of the Condorcet Paradox.

to C. And individuals 2 and 3 prefer C to A. Hence, Condorcet's SWF yields $A \succ^*_{\text{co}} B \succ^*_{\text{co}} C \succ^*_{\text{co}} A$. This phenomenon is known as the *Condorcet Paradox*.

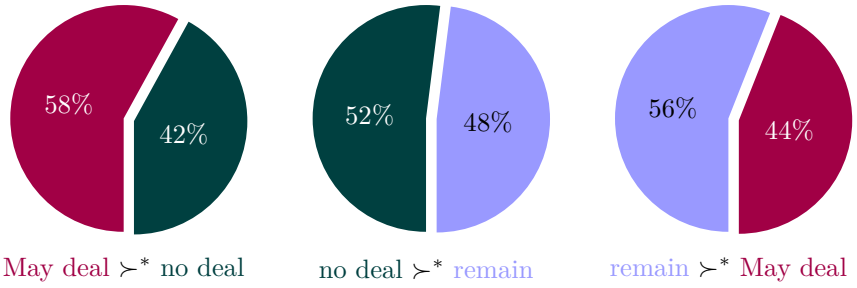

Figure 1 – "Thinking about your view of Brexit, for each of the following please say if it woudl be your first preference, second preference, or third preference." Deltapoll (2018)

The Condorcet Paradox can arise with significant probability, and has been observed in many real life instances (Van Deemen, 2014). Deltapoll (2018) reports the results from a poll of British citizens who were asked about their preferences over three alternatives: (*i*) leaving the EU with the deal that Prime Minister Theresa May was able to negotiate, (*ii*) leaving without a deal, and (*iii*) remaining a member of the EU. The results of the poll exhibit a Condorcet cycle. 58% ranked May's deal over leaving with no deal, 52% ranked no deal over remaining, and 56% ranked remaining over May's deal. See Figure 1. Perhaps this cycle is part of the reason why the British Parliament was unable to reach a decision and May's government ultimately collapsed.

Condition IIA (Independence of Irrelevant Alternatives) The social preferences over any two given alternatives depend *only* on the individual preferences over those two alternatives.

Condition IIA says that, given any three alternatives A, B, and C, the individual preferences over C should not play a role in determining the social preference between A and B. Consider for example the two preference profiles from Table 4. The only differences between the profile on the left and the profile on the right are in the preferences of individuals 6 and 7 over alternative C. The preferences of all individuals between A and B remain unchanged. Still, alternative A would win a general election given the preference profile on the left, while alternative B would win given the preferences on the right. Hence, plurality fails IIA.

1	2	3	4	5	6	7
A	A	A	B	B	C	C
B	B	B	C	C	B	B
C	C	C	A	A	A	A

1	2	3	4	5	6	7
A	A	A	B	B	B	B
B	B	B	C	C	A	A
C	C	C	A	A	C	C

Table 4 – Two preference profiles showing that plurality fails IIA.

Some people find IIA to be an appealing normative principle. But there are also practical reasons to require IIA. In order to apply a SWF in real life, it is necessary to elicit the preferences of the individuals. Under IIA, it is only necessary to elicit preferences over alternatives being considered. For SWFs that do not satisfy IIA, one might need to elicit a lot more information. And doing so might be costly or unfeasible.

Condition ND (Non-dictatorial) There does not exist a dictator.

Consider any dictatorial SWF. If the dictator's preferences are well defined and rational, then so is $\succ^*$. If everybody prefers A to B then so does the dictator. The social ranking between A and B only depends on the dictator relative ranking of A and B. Hence, dictatorial SWFs immediately satisfy IIA, UD, and U.

However, most people would agree that dictatorial SWFs are not good criteria to evaluate alternatives from a social perspective. Social preferences should take into account the preferences of more than one individual. Unfortunately, in rich enough environments, dictatorial SWFs are the only SWFs that satisfy the other three conditions proposed by Arrow.

Theorem 2.1 (Arrow (1950, 1951)) *If there are at least three alternatives, then there is no SWF that satisfies U, UD, IIA, and ND.*

3. An Incomplete Solution

One of the most popular approaches that economists use to rank alternatives is to use *Pareto dominance*, named after Pareto (1897, 1909).

Definition 2 Alternative A is said to *Pareto dominate* alternative B if and only if two conditions hold: (i) every individual weakly prefers A to B; and (ii) at least one individual strictly prefers A to B.

In other words, A Pareto dominates B if switching from B to A would make some individuals strictly better off, without making any individual worse off. In that case, we say that switching from A to B is a *Pareto improvement*. The Pareto criterion stipulates that if A Pareto dominates B, then society should strictly prefer A to B.[3]

We can construct a SWF around this criterion. Say that $A \succcurlyeq^*_{\text{pa}} B$ according to the *Paretian* SWF, if and only if $A \succcurlyeq_i B$ for every individual i. I claim that, according to this definition, we have $A \succ^*_{\text{pa}} B$ if and only if A Pareto dominates B. Trying to convince yourself of the validity of the claim can help you make sure that you understand the definitions.

The Paretian SWF satisfies conditions U, IIA, and ND. Therefore, we know from Arrow's Theorem that it must fail UD. It is always well defined and satisfies transitivity, but it fails completeness. If at least one member of society prefers A to B, and at least one member of society prefers B to A, then alternatives A and B are not comparable according to this criterion.

The lack of completeness can lead to situations in which there is no "best" alternative according to the Paretian SWF. That is, no alternative is socially (weakly or strictly) preferred to every other alternative. For instance, consider the preferences in Table 5. Alternative A is not preferred to C according to $\succcurlyeq^*_{\text{pa}}$, because of individual 1. Alternatives B and C are not preferred to A.

Even in such situations, the Paretian criterion can help to guide social choices. In the example from Table 5, it rules out alternative B, because switching from B to A is a Pareto improvement. In general settings, Society should not choose

[3] It is hard to find another idea in Economics which less controversial than this principle. But even a criterion as convincing as Pareto dominance is not without criticism. Sen (1970b) argues that the Pareto criterion can conflict with individual liberty. Gilboa et al. (2014) argue that the Pareto criterion can be inadequate in the presence of uncertainty and heterogeneous beliefs.

1	2	3	4	5	$\cdots$	n
C	A	A	A	A	$\cdots$	A
A	B	B	B	B	$\cdots$	B
B	C	C	C	C	$\cdots$	C

Table 5 – The pairs of alternatives A and C, and B and C are not comparable according to the Paretian SWF. However, $A \succ^*_{\text{pa}} B$.

an alternative if there is a different available alternative which Pareto dominates it. That is, Society should always choose *Pareto efficient* allocations.[4]

Definition 3 Alternative A is said to be *Pareto efficient* if it is not Pareto dominated by any other alternative. The set of all Pareto efficient allocations is called the *Pareto frontier*.

The concept of Pareto efficiency is very often misused. Pareto improvements are always desirable, but *not* every Pareto efficient alternative is preferable to all nonefficient alternatives. In the example from Table 5, alternative C is Pareto efficient, and alternative B is not. However, the Pareto criterion does not say that C is better than B. In order to avoid flawed reasoning, it is better to always think in terms of Pareto improvements.

Pareto dominance and Pareto efficiency can be visualized using utility functions. Suppose there are only two individuals, and let u_1 and u_2 be utility functions representing $\succcurlyeq_1$ and $\succcurlyeq_2$. We can associate each alternative A with a vector of utilities $\mathbf{u}(A) = (u_1(A), u_2(A))$, and plot these points in a Cartesian coordinate system like the one in Figure 2.

The shaded area represents the *utility possibilities set* U. It is the set of all utility vectors that can arise from some alternative. The Pareto frontier corresponds to the thick line demarcating the Northwest boundary of U. Alternative B Pareto dominates alternative A, because $u_1(B) > u_1(A)$ and $u_2(B) > u_2(A)$. Alternative C does not Pareto dominate A because $u_2(A) > u_2(C)$. The set of alternatives that Pareto dominate A corresponds to the hashed rectangle Northwest of $\mathbf{u}(A)$.

[4]The notions of "best" and "efficient" in this section correspond to the concepts of *maximum* and *maximal* from a branch of Mathematics called Order Theory. For complete preference relations, an alternative is a maximum if and only if it is maximal. For incomplete preferences, it is possible to have maximal elements which are not maximums. If you want to read more about this, a good place to start is the Wikipedia page for maximal and minimal elements.

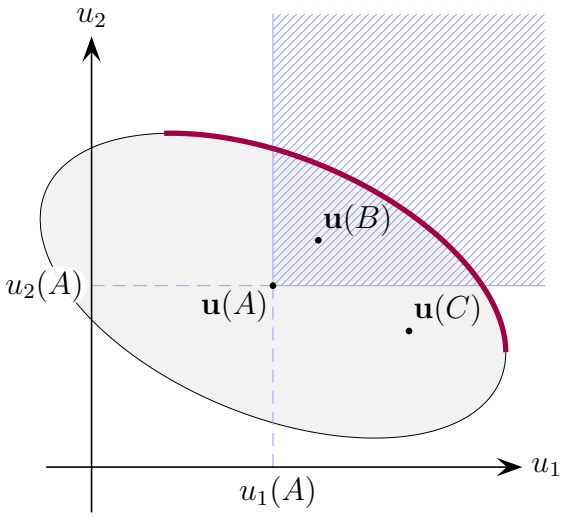

Figure 2 – Set of feasible utility vectors (gray oval), alternatives which Pareto dominate alternative A (hashed rectangle), and Pareto frontier (thick curve).

Let us conclude this section by noting that, under general conditions, there always exist Pareto efficient alternatives. Those enrolled in 2261 need not worry about the meaning of compactness and continuity. They are technical conditions. They are satisfied in all the models that we will consider in class, and most models taught in undergraduate courses.

Proposition 3.1 *If $\mathcal{A}$ is compact and nonempty, and every individual has continuous preferences, then the Pareto frontier is nonempty.*

The major drawback of the Pareto criterion is the fact that it is incomplete. The rest of these notes are devoting to finding ways of completing it. That is, we are looking for SWFs that always produce complete rankings of social alternatives, and never contradict the Pareto criterion. Unfortunately, none of the solutions we will consider are perfect.

4. If Utility Could Be Measured

The mainstream approach in contemporary Economics treats utility functions as nothing more than representations of preferences. That is why the input of SWFs consists of preference relations instead of utility functions. This approach makes it difficult to compare the utility of different individuals. For instance, suppose that society has a fortuitous surplus of $100 and must allocate it to either a billionaire or a pauper. One may argue that the pauper would extract more utility from this money than the billionaire. However, if both individuals prefer more money to less, this comparison cannot be based on preferences alone. In particular, the Pareto criterion cannot rank the two alternatives.

In contrast, before the XX$^{\text{th}}$ century, social philosophers often thought of utility as a measure of wellbeing. For example, Bentham (1789) proposed that maximizing utility should be the ultimate goal that both individuals and society should pursue. Suppose we were to accept this view. Further, suppose that we were able to measure utility functions using the same units of utility for all individuals. Later on, I will argue that these suppositions are very restrictive, but let us entertain the possibility for now. Under these suppositions, we would be able to rank social outcomes using cardinal social welfare functions defined as follows.

Definition 4 A *cardinal social welfare function* (CSWF) is a mathematical function that takes as inputs profiles of individual utility functions $(u_1, \ldots, u_n)$ expressed in terms of the same objective units and produces as output a social utility function w^*.

As with individual preferences, there are also many different ways to aggregate individual utilities. A common approach is to use the *utilitarian* CSWF, named after the work of Bentham (1789) and Mill (1863). The utilitarian criterion compares alternatives based on the total utility that individuals derive from it. The utilitarian CSWF is given by[5]

$$w^*_{\text{ut}}(A) = \sum_{i=1}^{n} u_i(A).$$

[5]The symbol "$\sum_{i=1}^{n} u_i(A)$" is read "*summation from* $i = 1$ *to* $i = n$ *of* $u_i(A)$." It is shorthand notation for "$u_1(A) + \ldots + u_n(A)$." We will use similar notation throughout the course.

Another possibility is the *max-min* CSWF, often associated with a notion of social justice proposed by Rawls (1971). The max-min criterion compares alternatives based on the utility of the least fortunate members of society. The social welfare assigned to an alternative A corresponds to the utility of the individual who is worse off given A. That is,

$$w^*_{\text{mm}}(A) = \min\Big\{u_1(A), u_2(A), \ldots, u_n(A)\Big\}.$$

The best alternative according to this criterion is the one that maximizes the minimum utility across individuals, hence the name max-min.

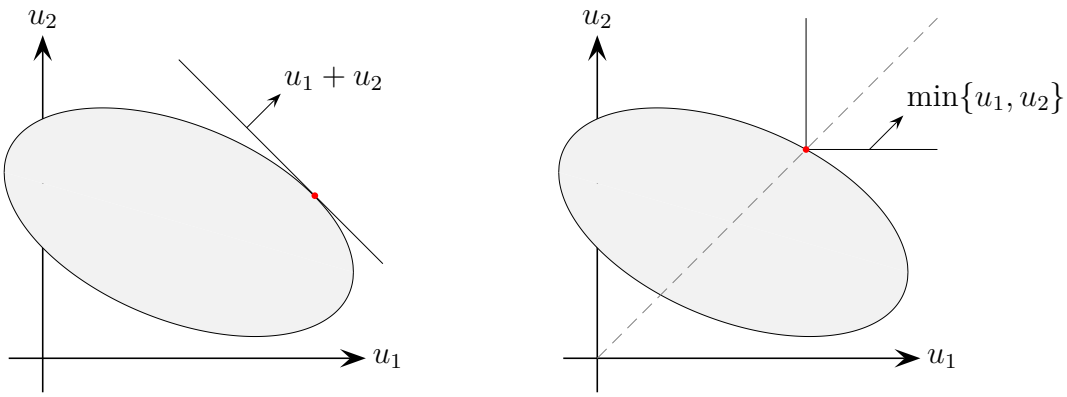

Figure 3 – Optimal alternatives according to the utilitarian criterion (left) and the max-min criterion (right).

Figure 3 illustrates the optimal alternative according to the utilitarian and max-min CSWFs for an example with two individuals. Both criteria select a point along the Pareto frontier, but each criterion chooses a different point. To understand this difference, it is helpful to make an analogy with a consumer choice problem. The utility possibilities set plays the role of the budget set. At the optimal points, the CSWF's indifference curve and the frontier of the utility possibilities set must be tangent to one another. Note that the indifference curves for the utilitarian CSWF are straight lines, while the indifference curves for the max-min CSWF are L-shaped.

Following the analogy, the utilitarian criterion treats the utility of different substitutes as perfect substitutes. In contrast, the max-min criterion treats them as perfect complements, and always chooses points along the 45-degree line. The two criteria take opposite stances on the priority between two social goals: to-

tal wellbeing and distribution of wellbeing. The max-min criterion prioritizes equality, while the utilitarian criterion maximizes total wellbeing regardless of its distribution .

The contrast between the utilitarian and max-min criteria illustrates that different CSWFs might involve different implicit value judgments. Unlike the Pareto criterion, some of these value judgments might be highly controversial. For example, different people might have different views about the relative importance of the distribution and the total level of wellbeing. When trying to measure cardinal welfare, it is crucial to be aware of the specific CSWF being used and the values it represents.

5. Comparing the Utility of Different Individuals

The utilitarian CSWF discussed in the previous section is ubiquitous in applied work and policymaking. Its popularity is due in part to the fact that it always provides answers to important questions. Unlike the Pareto criterion, the utilitarian criterion can always rank any pair of alternatives. Moreover, it can quantify changes in welfare. For instance, it allows practitioners to draw conclusions of the form: "the proposed policy would result in a 20% welfare increase." However, it is not always clear how to interpret these answers.

The reason is that CSWFs often involve comparing the level of utility of different agents. For example, the utilitarian CSWF adds them up. For interpersonal comparisons to make sense, the utility of different individuals must be expressed in the same units. It is meaningless to add three meters plus five feet without making a unit conversion first. How can we measure the utility functions of different people using the same units?

Some economists, such as Robbins (1945), have argued that we cannot because utility exists only inside people's minds, and we cannot measure it directly. Other economists have attempted to construct measures of happiness using survey data (Frey and Stutzer, 2002). However, this approach faces severe limitations that limit its applicability (Bond and Lang, 2019). Other economists hope that advances in Neuroeconomics might provide better ways to measure utility in the future. However, we still do not have the technology to do so (Bernheim, 2009,

Section II). For the time being, the prevailing way to measure utility is using preferences inferred from choice data.

A difficulty arises because the utility functions derived from preferences are unique up to monotone transformations. Hence, there can be different utility representations that are consistent with the same observed data. In turn, different utility representations can result in different measures of cardinal welfare.

Consider an example with three individuals and three alternatives. The left panel of Table 6 specifies the utilities of the individuals expressed in the same units. The utilitarian alternative is C. This is in part because individual 3 is much more sensitive than the other two. Nevertheless, the preferences of the individuals over these alternatives cannot directly reveal this gap in sensitivity. The utility functions in the right panel of Table 6 are also consistent with the preferences of the individuals. Based on these later utilities, alternative C would be ranked last by the utilitarian criterion.

	u_1	u_2	u_3
A	2	2	0
B	1	1	10
C	0	0	20

	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_3$
A	2	2	0
B	1	1	1
C	0	0	2

Table 6 – Two different utility representations for the same preferences.

One way to deal with this difficulty is to multiply the utility of different individuals by different positive numbers $\lambda_1, \ldots, \lambda_n$, often referred to as *Pareto weights*. With the right Pareto weights, doing so could potentially ensure that all the utilities are expressed in the same units. For example, suppose we multiply the utility functions $\hat{u}_i$ from the right panel of Table 6 by the Pareto weights $\lambda_1 = 1$, $\lambda_2 = 2$, and $\lambda_3 = 10$. Then, we would recover the utility functions from the left panel of the table, which are expressed in the same units.

Instead of the utilitarian CSWF, we could use the *Harnsanyi CSWF* named after Harsanyi (1955). The Harnsanyi CSWF given a list of Pareto weights $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ is given by

$$w_{\boldsymbol{\lambda}}(A) = \sum_{i=1}^{n} \lambda_i u_i(A).$$

With the right choice of Pareto weights, the Harsanyi CSWF could potentially measure the total wellbeing of the members of society. The question remains of

how to find the right weights. This is an important question, but the different attempts to solve it are beyond the scope of this course. Without a good way to choose weights, we must face the following result.

Proposition 5.1 *If $U = \{\mathbf{u}(A) \mid A \in \mathcal{A}\}$ is convex then, for every Pareto efficient alternative A, there exist Pareto weights $\boldsymbol{\lambda}$ such that A maximizes the Harsanyi CSWF given $\boldsymbol{\lambda}$.*

Proposition 5.1 teaches us two lessons. The first lesson is that we should be cautious using the utilitarian CSWF. The utilitarian criterion always chooses a point on the Pareto frontier. However, unless we are able to measure utilities with objective units, the chosen point is arbitrary. Any point that is Pareto efficient will maximize the sum of weighted utilities using some list of weights. Hence, without a good way of choosing Pareto weights, adding up utilities cannot rule out any alternatives that were not already ruled our by the Pareto SWF.

The second lesson is that CSWFs can be a powerful computational tool to characterize the Pareto frontier. The Pareto frontier may be difficult to characterize using Definition 3. Proposition 5.1 implies that, in convex setting, it can be characterized by maximizing Harsanyi CSWFs with different Pareto weights. In many settings, this is the easiest way to compute the Pareto frontier.

6. Money as a Measure of Utility

All the welfare criteria we have discussed so far have limitations in general environments. In particular, the Pareto criterion can be uninformative, and the utilitarian criterion can be misleading. In this section, we will focus on a special class of environments called *transferable-utility* environments. In such environments the utility of each individual can be measured in monetary units, and monetary transfers can be used to transfer utility between individuals with a 1:1 ratio. Because of that, the Pareto criterion is more informative, and the utilitarian criterion is easier to interpret.

Let us begin with an example. Anna and Bob would like to attend an event, but there is only one ticket remaining. The Pareto criterion alone cannot deter-

mine who should get the ticket. Giving the ticket to Anna and giving it to Bob are both Pareto efficient. Things can be very different if we consider the possibility of monetary transfers between Anna and Bob, and we knew how much each individual is willing to pay for the ticket.

Suppose that Bob is indifferent between \$100 and the ticket, while Anna is indifferent between \$200 and the ticket. I claim that, in that case, the Pareto criterion dictates that Anna should have the ticket. To see why, suppose that Bob had the ticket. Compared to that alternative, Bob could transfer the ticket to Ana, and Anna could pay \$150 to Bob. Doing so would make both Ana and Bob strictly better off, resulting in a Pareto improvement.

General transferable-utility environments are characterized by two features. The first feature is that the alternatives under consideration can be split into a monetary component and a non-monetary component. The non-monetary component is chosen from some fixed set $\mathcal{A}$, as before. The monetary component is a *transfer scheme*, described by a list $\mathbf{t} = (t_1, \ldots, t_n)$ of real numbers specifying a monetary transfers from each individual. That is, t_i is the amount of money that individual i has to pay. We restrict attention to transfer schemes that satisfy the *budget balance* condition $\sum_{i=1}^{n} t_i = 0$. An *extended alternative* is a pair $(A, \mathbf{t})$ consisting of a non-monetary alternative $A \in \mathcal{A}$ and a budget-balanced transfer scheme.

The second feature is that, for every individual i, $\succcurlyeq_i$ admits a utility representation of the form

$$u_i(A, \mathbf{t}) = v_i(A) - t_i.$$

Preferences with this property are called *quasilinear*. The key aspect of this utility representation is that there are no income effects, that is, the marginal benefit of money is constant. It does not depend on the chosen non-monetary alternative, nor on the size of the monetary transfer. Typically, quasilinearity is a good assumption when the size of the transfers is small, and choosing different alternatives would not have a significant impact on the wealth of any individual.

In transferable-utility environments, monetary transfers allow society to transfer utility between individuals with a 1:1 ratio. For example, if Ana pays \$150 to Bob, Ana's utility goes down by \$150 while Bob's utility goes up by the same amount.

In transferable-utility environments, the Pareto frontier is always a line with slope -1. To see why, consider the example in Figure 4. Without transfers, choos-

ing alternative A would results in the utilities $(v_1(A), v_2(A))$. With transferable utility, if individual x pays one dollar to individual 2, then 2's utility would go up by x units, and x's utility would go down by exactly the same amount. Hence, the pair utilities $(v_1(A) - x, v_2(A) + x)$ is feasible. Since x is arbitrary, the whole line of slope -1 containing the point $(v_1(A), v_2(A))$ can be attained with transfers.

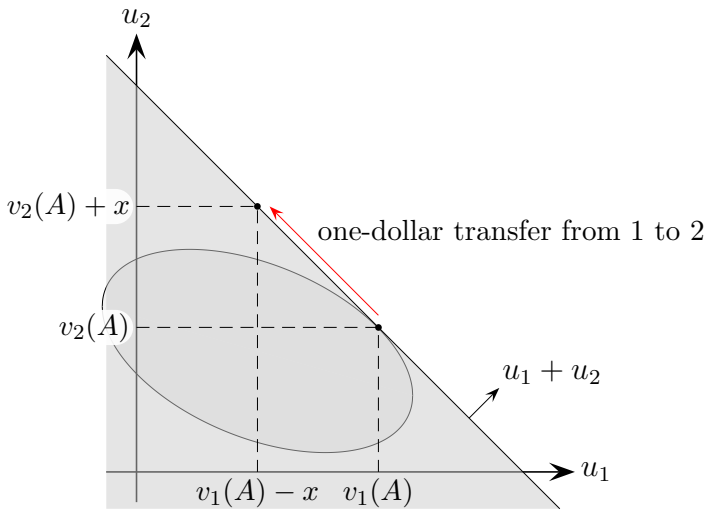

Figure 4 – Feasible utility with quasilinear preferences and monetary transfers.

In transferable utility environments, there is no trade off between the total level of wellbeing and its distribution. Since utility is transferable, any distribution is possible for any given level of welfare. Moreover, all Pareto efficient outcome must have the same total utility. Thus, the Pareto criterion requires that the chosen alternative must also be utilitarian. Hence, the Pareto criterion is no longer incomplete. This idea is formalized in the following proposition.

Proposition 6.1 *In an environment with monetary transfers and quasilinear preferences, an extended alternative* $(A, \mathbf{t})$ *is Pareto efficient if and only if the alternative* A *satisfies*

$$\sum_{i=1}^{n} v_i(A) \geq \sum_{i=1}^{n} v_i(B), \tag{1}$$

for every other alternative $B \in \mathscr{A}$.

Transferable-utility models are very attractive because they make welfare analysis easy. However, it is important to recall that the two key assumptions in such

models are restrictive. On one hand, unlimited monetary transfers are often not feasible from a political standpoint. Taxation and subsidies targeted to specific groups often face strong political opposition. On the other hand, the quasilinear assumption is also restrictive. Suppose for instance that we are interested in the welfare implications of a policy that affects unemployment. Offering welfare checks might be insufficient compensation to those who lose their jobs.

Some economists have advocated for the use of the utilitarian criterion (or other CSWFs) in quasilinear utility environments even without transfers. This is sometimes called the *Kaldor-Hicks criterion* in honor of Kaldor (1939) and Hicks (1939). The idea is that, quasilinearity allows to measure the utility of all agents in the same monetary units. Hence, one can compare the gains of those who benefit from a policy to the loses of those who are hurt by it.

Let us revisit the example with Anna and Bob, but now preclude monetary transfers. Without transfers, it is no longer the case that giving the ticket to Bob is Pareto dominated. However, the Kaldor-Hicks criterion still suggests that Anna should receive the ticket. One possibility is to make the argument that Anna's higher willingness to pay reveals that she will derive greater enjoyment from attending the event. Therefore, she should get the ticket in order to maximize the total enjoyment in society.

However, there is another possibility. It could be the case that Anna and Bob would derive the same enjoyment from the event, but Anna is much wealthier than Bob. In that case, Anna is willing to pay more simply because her marginal utility for money is lower. Just because it is possible to express the utility of all individuals in the same units, it does not mean that it is a good idea to do so. This example shows that, in societies with high wealth inequality, using the Kaldor-Hicks criterion carelessly could have the unintended consequence of prioritizing the wellbeing of the wealthy. There are different ways to prevent this, but they are beyond the scope of this class.

7. Where does that leave us?

The main takeaway from these notes should be that evaluating welfare is a difficult problem and there are no perfect solutions. That is why economists

employ a variety of tools. It is important to understand how the strengths and weaknesses of each tool in order to know how which tools to use and how to interpret them. We have focused on two of the most commonly used criteria to evaluate welfare, the Pareto criterion and the utilitarian criterion.

The Pareto criterion is the gold standard. Whenever it can provide an answer, it is the best criterion to use. However, it is incomplete and often fails to provide an answer. Moreover, in order to avoid flaw reasoning, it is better to think in terms of Pareto improvements and not in terms of Pareto efficiency. Recall that not every Pareto efficient alternative is preferable to every non-efficient one.

The utilitarian criterion is popular fallback option for situations in which the Pareto criterion fails to provide an answer. However, it must be used cautiously. The choice of a specific CSWF involves a moral stance on the relative importance of total wellbeing and its distribution. Moreover, the utilitarian criterion can be arbitrary unless there is a good way to make sure that the utilities of all individuals are expressed in the same units. Even when this is can be done (e.g., in environments with quasilinear preferences), our choice of units can influence our welfare computations.

Welfare analysis is much easier in models transferable utility environments with monetary transfers and quasilinear preferences. In such models, the Pareto and utilitarian criterion are closely related to one another, and they both help to rank social alternatives in a convincing manner. However, it is important to note that the assumptions of the model are restrictive. Despite their convenience, transferable utiliy models should not be applied to study arbitrary economic environments. There are other similar classes of models that also simplify welfare analysis, but they are beyond the scope of our class.

References

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346.

Arrow, K. J. (1951). *Social Choice and Individual Values*. Number 12 in Cowles commission for research in Economics monographs. John Wiley & Sons.

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*. Clare-

don Press.

Bernheim, B. D. (2009). On the potential of Neuroeconomics: A critical (but hopeful) appraisal. *American Economic Journal: Microeconomics*, 1(2):1–41.

Bond, T. N. and Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, 127(4):000–000.

Borda, J. (1781). Mémoire sur les élections au scrutin. In *Mémoires de l'Académie Royale des Sciences*, pages 657–665. Imprimerie Royale.

Burk, A. (1938). A reformulation of certain aspects of welfare economics. *The Quarterly Journal of Economics*, 52(2):310–334.

Chipman, J. S. (1976). The Paretian heritage. *Revue européenne des sciences sociales*, 14(37):165–171.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Imprimerie Royale.

Deltapoll (2018). Brexit deal survey. Retrieved July 27, 2919 from: http://www.deltapoll.co.uk/polls/brexit-deal-leave-remain.

Diamond, P. A. (1967). Cardinal welfare, individualistic ethics, and interpersonal comparison of utility: Comment. *The Journal of Political Economy*, 75(5):765.

Dinwiddy, J. and Twining, W. (1989). Bentham: Selected writings of John Dinwiddy. Oxford University Press.

Frey, B. S. and Stutzer, A. (2002). What can economists learn from happiness research? *Journal of Economic Literature*, 40(2):402–435.

Gilboa, I., Samuelson, L., and Schmeidler, D. (2014). No-betting-Pareto dominance. *Econometrica*, 82(4):1405–1442.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4):309–321.

Hicks, J. R. (1939). The foundations of welfare economics. *The Economic Journal*, 49(196):696–712.

Kaldor, N. (1939). Welfare propositions of economics and interpersonal comparisons of utility. *The Economic Journal*, 49(195):549–552.

Mill, J. S. (1863). *Utilitarianism*. Parker Son and Bourn.

Pareto, V. (1897). *Cours d'Économie Politique Professé a L'Université de Lausanne*, volume 2. F. Rouge & F. Pichon.

Pareto, V. (1909). *Manuel d'Économie Politique*. Giard & Brière.

Pattanaik, P. K. (2017). Social welfare function. In *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, 3rd edition.

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

Robbins, L. (1945). *Essay on the Nature & Significance of Economic Science*. Macmillan & Co., Limited, second (revised and extended) edition.

Samuelson, P. A. (1947). *Foundations of Economic Analysis*, volume 80 of *Harvard Economic Studies*. Harvard University Press.

Samuelson, P. A. (1981). Bergsonian welfare economics. In Rosefielde, S., editor, *Economic welfare and the economics of Soviet Socialism: essays in honor of Abram Bergson*, chapter 9, pages 223–66. Cambridge University Press.

Sen, A. (1970a). *Collective Choice and Social Welfare*. Harvard University Press.

Sen, A. (1970b). The impossibility of a Paretian liberal. *Journal of Political Economy*, 78(1):152–157.

Van Deemen, A. (2014). On the empirical relevance of Condorcet's paradox. *Public Choice*, 158(3–4):311–330.

A. Some Historic Remarks

A.1. The Summation of Utilities

The summation of utilities does *not* appear explicitly in the work of neither Bentham or Mill. However, it can be derived from two of the principles of utilitarianism. The first principle is that fostering leisure and avoiding pain and pleasure is the ultimate goal of individuals and should be the ultimate goal of society

> Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do, as well as to determine what we shall do (Bentham, 1789, pp. 1).

The second principle is that the pains and pleasures of all individuals should count exactly the same.

> [T]he very meaning of Utility... is a mere form of words without rational signification, unless one person's happiness, supposed equal in degree... is counted for exactly as much as another's. Those conditions being supplied, Bentham's dictum, "everybody to count for one, nobody for more than one," might be written under the principle of utility as an explanatory commentary (Mill, 1863, Chapter 5).

Interestingly John Dinwiddy attributes the following quote to Bentham:

> 'Tis vain to talk of adding quantities which after the addition will continue to be as distinct as they were before; one man's happiness will never be another man's happiness: a gain to one man is no gain to another: you might as well pretend to add 20 apples to 20 pears (Dinwiddy and Twining, 1989, pp. 49).

A.2. The Pareto Critique

Pareto was interested in the question of whether the outcomes of competitive markets are "optimal." He was pressed to come up with a criterion of optimality that made no utility comparisons across different people, because the utility of different people could be measured in different units. In Pareto's words:

> Nous ne pouvons ni comparer ni sommer celles-ci [dU^1, dU^2, etc.], car nous ignorons le rapport des unites en lesquelles elles sont exprimées (Pareto, 1897, pp. 93).

Which roughly translates to "we can neither compare nor sum these [individual utilities], because we do not know in which units they are expressed." His solution was to use a criterion that depends only on ordinal preferences. For more on the history of the Pareto criterion and other contributions from Vilfredo Pareto see Chipman (1976).

Lionel Robbins proposed a very extreme and influential version of the Pareto Critique. In the preface of the second edition of his 1932 essay, he argues that

> ...the aggregation or comparison of the different satisfactions of different individuals involves judgments of value rather than judgments of fact, and that such judgments are beyond the scope of positive economics (Robbins, 1945, pp. vii).

A.3. The term "Social Welfare Function"

The notion of SWF from Definition 1 is due to Arrow (1950, 1951). The notion of CSWF from Definition 4 is due to Sen (1970a), and it is often called a *Social*

Welfare Functional. However, the term "Social Welfare Function" can be traced back to Burk (1938)—who at some point changed his surname to Bergson—and Samuelson (1947).

The notion of *Bergson-Samuelson Welfare Functions* (BS-SWF) is different from the notions of SWF and CSFW discussed in these notes. One difference is that Bergson and Samuelson do allow for the use of cardinal information, although Samuelson's notion of welfare does *not* depend crucially on it. A more substantial difference is that BS-SWFs assign a welfare value to each alternative, given a *fixed* utility profile for individuals. In contrast, Arrow's SWFs are defined for *many different* preference profiles within a given domain. Some people would argue that this is an important difference. For example, Arrow's IIA cannot be defined for BS-SWFs. See Pattanaik (2017) for further clarification. Having one name for two different objects has been a source of confusion.

> [Arrow] used the same name for his unicorn that Bergson and other writers had used for their existent animals. So it is not particularly surprising that Arrow's readers, learning that he had proved the impossibility of a "social welfare function" should have formed the mistaken inference that there cannot exist a reasonable and well-behaved Bergsonian social welfare function (Samuelson, 1981, pp. 228).

A.4. The Veil of Ignorance

Rawls (1971) proposed that the principles of justice should be chosen behind a *veil of ignorance* so that

> . . . no one knows his place in society, his class position or social status; nor does he know his fortune in the distribution of natural assets and abilities, his intelligence and strength, and the like (Rawls, 1971, pp. 12).

Rawls used this original position of ignorance as a thought experiment to derive his principles of justice. One of these principles can be interpreted as the max-min rule:

> [E]ach person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others (Rawls, 1971, pp. 60).

The veil of ignorance was made popular throughout different disciplines by Rawls. However, it appears already in the work of Harsanyi (1955). Harsaniyi argues that an individual's welfare assessments should indicate

> ... what social situation he would choose if he did not know what his personal position would be in the new situation choosen (and in any of its alternatives) but rather had an equal *chance* of obtaining any of the social positions existing in this situation, from the highest down to the lowest (Harsanyi, 1955, pp. 316).

Interestingly, Harsanyi used the veil of ignorance as a basis to justify the weighted utilitarian CSWF and not the max-min CSWF. Harsanyi uses arguments from the theory of choice under uncertainty, including the so called sure-thing principle. His approach was later criticized by Diamond (1967), who argues that

> I am willing to accept the sure-thing principle for individual choice but not for social choice, since it seems reasonable for the individual to be concerned solely with the final states while society is also interested in the process of choice (Diamond, 1967, pp. 766).