# 7 Separability

## 7.1 The construction of value

When a possible outcome looks attractive, then this is usually because it has attractive aspects. It may also have unattractive aspects, but the attractive aspects (the "pros") outweigh the unattractive aspects (the "cons"). In this chapter, we will explore how this weighing of different aspects might work.

At the end of chapter 5 we saw that the utility of a proposition for an agent is determined by two factors: the agent's credences, and the agent's basic desires, reflected in the utility the agent assigns to her "concerns", where a concern is a proposition that settles everything the agent ultimately or intrinsically cares about.

Suppose, for example, that you have only two basic desires: to be rich and to be famous. The proposition *that you are both rich and famous* then settles everything you ultimately care about. All worlds in which you are both rich and famous are equally desirable for you, no matter what else happens in them. The same is true for the worlds in which you are *neither rich nor famous*, and for the worlds in which you are *rich and not famous*, and for the worlds in which you are *not rich and famous*. These four sets of worlds are your concerns.

In general, if there are $n$ propositions $A_1, A_2, \ldots, A_n$ that an agent ultimately cares about, then any consistent conjunction that can be formed from these propositions and their negations (such as $A_1 \wedge \neg A_2 \wedge A_3 \wedge \ldots \wedge \neg A_n$) has uniform utility, and is one of the agent's concerns.

It will be useful to have a label for an agent's utility function restricted to concerns. I will call it the agent's **value function**. An agent's value function represents the agent's intrinsic, belief-independent goals or motives or values.

The aim of the present chapter is to explore under what conditions an agent's value function is a matter of adding up "pros" and "cons". Suppose, in the above example, that your intrinsic desire for wealth is a little stronger than your desire for fame. A state in which you are *rich and not famous* (for short, $R \wedge \neg F$) is then better than

a state in which you are *not rich and famous* ($\neg R \wedge F$). Both kinds of state have a good aspect (a "pro") and a bad aspect (a "con"), but the $R$ aspect has more weight than the $F$ aspect. We might capture this by saying that $R$ contributes 2 amounts of utility to a state while $F$ contributes only 1, so that the total utility of $R \wedge \neg F$ is 2 and that of $\neg R \wedge F$ is 1. The utility of $R \wedge F$ would be 3, and the utility of $\neg R \wedge \neg F$ would be 0.

## 7.2 Additivity

Let's look at a slightly more complex example. You are looking for a flat to rent. You care about certain aspects of a flat such as size, location, and price. We'll call these aspects **attributes**. If a set of attributes comprises all the features (of a flat) that matter to you, then your preferences between possible flats are determined by your preferences between combinations of these attributes: if you prefer one flat to another, that's because you prefer the combined attributes of the first to those of the second.

So the desirability of any possible flat is determined by the desirability of every possible combination of attributes. We'll write these combinations as lists enclosed in angular brackets. For example, '$\langle 40\text{m}^2, \text{central}, £500 \rangle$' stands for any flat with a size of 40 $\text{m}^2$, central location, and monthly costs of £500. Let's assume that size, location, and price are all the attributes you care about. Your utility function then assigns the same value to all flats represented by the list $\langle 40\text{m}^2, \text{central}, £500 \rangle$.

In fact, of course, utility functions don't assign numbers to flats. Your preferences are defined over propositions, not over flats. When I say that you prefer one kind of flat over another, what I really mean is that you prefer living in one kind of flat over living in the other. The attributes we are interested in are really attributes of worlds, not of flats. To keep things simple, we currently assume that the only thing you ultimately care about is what kind of flat you are living in (or going to live in). Your concerns can therefore be represented by lists like $\langle 40\text{m}^2, \text{central}, £500 \rangle$. Any such list settles everything you ultimately care about.

In this toy example, your value function assigns a desirability score to all possible combinations of size, location, and price. If you're like most people, we can we say more about how these scores are determined. For example, you probably prefer cheaper flats to more expensive flats, and larger flats to smaller flats.

The "weighing up pros and cons" idea suggests that the overall score for a given flat is determined by adding up individual scores for the flat's properties. A cheap but small flat in a good location, for example, would get a high score for price, a low score for size, and a high score for location.

More formally, the idea is to identify the value of any given attribute list with the sum of **subvalues** assigned to the elements in the list. We might say that (living in) a 40 $m^2$ flat has, for you, a certain intrinsic (sub)value $V_S(40\text{m}^2) = 1$. The (sub)value you assign to central location might be $V_L(\text{central}) = 2$, while the (sub)value you assign to monthly costs of £500 is $V_P(£500) = -1$. The overall value of $\langle 40\text{m}^2, \text{central}, £500 \rangle$ would then be the sum of these subvalues:

$$V(\langle 40\text{m}^2, \text{central}, £500 \rangle) = V_S(40\text{m}^2) + V_L(\text{central}) + V_P(£500) = 2.$$

If a value function $V$ is determined by adding up subvalues in this manner, then $V$ is called **additive** relative to the attributes in question.

Additivity may seem to imply that you assign the same weight to all the attributes: that size, location, and price are equally important to you. To allow for different weights, we could introduce scaling factors $w_S, w_L, w_P$, into the determination of value, like so:

$$V(\langle 40\text{m}^2, \text{central}, £500 \rangle) = w_s \cdot V_S(40\text{m}^2) + w_L \cdot V_L(\text{central}) + w_P \cdot V_P(£500).$$

We can, however, omit the weights by folding them into the subvalues. We will let $V_S(200\text{m}^2)$ measure not just how awesome it would be to have a 200 $m^2$ flat, but also how important this feature is compared to price and location.

**Exercise 7.1** ††

Like utility functions, subvalue functions assign numbers to propositions that needn't be of uniform utility. Unlike utility functions, however, subvalue functions are insensitive to belief. For example, if you can afford to pay £600 in monthly rent, then $V_P(£300)$ is plausibly high, even though the utility you assign to renting a flat for £300 is plausibly low. Can you explain why?

**Exercise 7.2** †††

Additivity greatly simplifies an agent's psychology. Suppose an agent's basic desires pertain to 10 logically independent propositions $A_1, A_2, \ldots, A_{10}$. There are $2^{10} = 1024$ conjunctions of these propositions and their negations (such as $A_1 \wedge A_2 \wedge \neg A_3 \wedge \neg A_4 \wedge A_5 \wedge A_6 \wedge \neg A_7 \wedge A_8 \wedge A_9 \wedge \neg A_{10}$). To store the agent's value function in a database, we would therefore need to store up to 1024 numbers. How many numbers do we need to store in the database if the agent's value function is additive?

## 7.3 Separability

Under what conditions is value determined by adding subvalues? How are different subvalue functions related to one another? What do subvalue functions represent anyway? We can get some insight into these questions by following an idea from the previous chapter and study how an agent's value functions might be derived from their preferences – specifically, from their preferences over complete attribute lists, which we take to represent the agent's concerns.

The main motivation for starting with preferences is, as always, the problem of measurement. We need to explain what it means that your subvalue for a given attribute is 5 rather than 29. Since the numbers are supposed to reflect, among other things, the importance (or weight) of the relevant attribute in comparison to other attributes, it makes sense to determine the subvalues from their effect on the overall ranking of attribute lists.

So assume we have preference relations $\succ$, $\succsim$, $\sim$ between lists of attributes. To continue the illustration in terms of flats, if you prefer a central 40 $m^2$ flat for £500 to a central 60 $m^2$ for £800, then

$$\langle 40\text{m}^2, \text{central}, £500 \rangle \succ \langle 60\text{m}^2, \text{central}, £800 \rangle.$$

If, like most people, you prefer to pay less rather than more, then your subvalue function $V_P$ is a decreasing function of monthly costs: the higher the costs $c$, the lower $V_p(c)$. This doesn't mean that you prefer *any* cheaper flat to *any* more expensive flat. You probably don't prefer a 5 $m^2$ flat for £499 to a 60 $m^2$ flat for £500. The other attributes also matter.

In what sense, then, do you prefer cheaper flats to more expensive flats? We can cash this out as follows: whenever two flats agree in terms of size and location, and one is cheaper than the other, then you prefer the cheaper one.

Let's generalize this idea.

Consider an attribute list $\langle A_1, A_2, \ldots A_n \rangle$, and let $A_1'$ be an alternative to $A_1$. If, for example, the first position in an attribute list represents monthly costs, then $A_1$ might be £400 and $A_1'$ £500. We can now compare $\langle A_1, A_2, \ldots A_n \rangle$ to $\langle A_1', A_2, \ldots A_n \rangle$ – a hypothetical flat that's like the first in terms of size and location, but costs £100 more. If

$$\langle A_1, A_2, \ldots, A_n \rangle \succ \langle A_1', A_2, \ldots, A_n \rangle,$$

we say that you prefer $A_1$ to $A_1'$ *conditional on* $A_2, \ldots, A_n$.

Suppose you prefer $A_1$ to $A_1'$ conditional on any way of filling in the remainder $A_2, \ldots, A_n$ of the attribute list. In that case, we can say that your preference of $A_1$ over $A_1'$ is *independent* of the other attributes.

In the flat example, your preference of £400 over £500 is plausibly independent of the other attributes: whenever two possible flats agree in size and location, but one costs £400 and the other £500, you plausibly prefer the one for £400. (We are still assuming that size, location, and costs are all you care about.)

We can similarly consider alternatives properties $A_i$ and $A_i'$ that may figure at another position in an attribute list. If we find that your preferences between $A_i$ and $A_i'$ are always independent of the other attributes, we say that your preferences between attribute lists are **weakly separable**.

Weak separability means that your preference between two attribute lists that differ only in one position does not depend on the attributes in the other positions.

Consider the following preferences between four possible flats.

$$\langle 50\text{m}^2, \text{central}, £500 \rangle \succ \langle 40\text{m}^2, \text{beach}, £500 \rangle$$
$$\langle 40\text{m}^2, \text{beach}, £400 \rangle \succ \langle 50\text{m}^2, \text{central}, £400 \rangle$$

Among flats that cost £500, you prefer central 50 $\text{m}^2$ flats to 40 $\text{m}^2$ flats at the beach. But among flats that cost £400, your preferences are reversed: you prefer 40 $\text{m}^2$ beach flats to 50 $\text{m}^2$ central flats. In a sense, your preferences for size and location depend on price. But we don't have a violation of weak separability.

That's why weak separability is called 'weak'. To rule out the present kind of

dependence, we need to strengthen the concept of separability. Your preferences are **strongly separable** if your ranking of lists that differ in *one or more positions* does not depend on the attributes in the remaining positions, in which they do not differ. In the example, your ranking of $\langle 50\text{m}^2, \text{central}, - \rangle$ and $\langle 40\text{m}^2, \text{beach}, - \rangle$ depends on how the blank ('–') is filled in. Your preferences aren't strongly separable.

**Exercise 7.3** ††

Suppose all you care about is the degree of pleasure of you and your three friends, which we can represent by a list like $\langle 10, 1, 2, 3 \rangle$. Suppose further that you prefer states in which you all experience equal pleasure to states in which your degrees of pleasure are different. For example, you prefer $\langle 2, 2, 2, 2 \rangle$ to $\langle 2, 2, 2, 8 \rangle$, and you prefer $\langle 8, 8, 8, 8 \rangle$ to $\langle 8, 8, 8, 2 \rangle$. Are your preferences weakly separable? Are they strongly separable?

**Exercise 7.4** ††

Which of the following preferences violate weak separability, based on the information provided? Which violate strong separability?

| (a) | (b) | (c) |
|---|---|---|
| $\langle A_1, B_1, C_3 \rangle \succ \langle A_3, B_1, C_1 \rangle$ | $\langle A_1, B_3, C_1 \rangle \succ \langle A_1, B_3, C_2 \rangle$ | $\langle A_1, B_3, C_2 \rangle \succ \langle A_1, B_1, C_2 \rangle$ |
| $\langle A_3, B_2, C_1 \rangle \succ \langle A_1, B_2, C_3 \rangle$ | $\langle A_1, B_2, C_2 \rangle \succ \langle A_1, B_2, C_3 \rangle$ | $\langle A_2, B_3, C_2 \rangle \succ \langle A_2, B_1, C_2 \rangle$ |
| $\langle A_3, B_2, C_3 \rangle \succ \langle A_3, B_2, C_1 \rangle$ | $\langle A_3, B_2, C_3 \rangle \succ \langle A_3, B_1, C_3 \rangle$ | $\langle A_1, B_1, C_1 \rangle \succ \langle A_1, B_3, C_1 \rangle$ |

In 1960, Gérard Debreu proved that strong separability is exactly what is needed to ensure additivity.

To state Debreu's result, let's say that an agent's preferences over attribute lists have an **additive representation** if there is a value function $V$, assigning numbers to the lists, and there are subvalue functions $V_1, V_2, \ldots, V_n$, assigning numbers to the items on the lists, such that the following two conditions are satisfied. First, the preferences are represented by $V$. That is, for any two lists $A$ and $B$,

$$A \succ B \text{ iff } V(A) > V(B), \text{ and } A \sim B \text{ iff } V(A) = V(B).$$

Second, the value assigned to any list $\langle A_1, A_2, \ldots, A_n \rangle$ equals the sum of the subval-

ues assigned to the items on the list:

$$V(\langle A_1, A_2, \ldots, A_n \rangle) = V_1(A_1) + V_2(A_2) + \ldots + V_n(A_n).$$

Now, in essence, Debreu's theorem states that if preferences over attribute lists are complete and transitive, then they have an additive representation if and only if they are strongly separable.

A technical further condition is needed if the number of attribute combinations is uncountably infinite; we'll ignore that. Curiously, the result also requires that there are at least three attributes that matter to the agent. For two attributes, a different condition called 'double-cancellation' is required instead of strong separability. Double-cancellation says that if $\langle A_1, B_1 \rangle \succsim \langle A_2, B_2 \rangle$ and $\langle A_2, B_3 \rangle \succsim \langle A_3, B_1 \rangle$ then $\langle A_2, B_3 \rangle \succsim \langle A_3, B_2 \rangle$. But let's just focus on cases with at least three relevant attributes.

The proof of Debreu's theorem requires some serious maths. We will not get near it. I will mention, though, another interesting fact that falls out of the proof: if the agent's preferences are defined over a sufficiently rich set of possibilities, then the value function $V$ that additively represents the agent's intrinsic preferences is unique except for the choice of unit and zero.

This suggests a new response to the ordinalist challenge. The ordinalists claimed that utility assignments are arbitrary as long as they respect the agent's preference order. In response, one might argue that rational (intrinsic) preferences should be strongly separable and that an adequate representation of such preferences should involve an additive utility (or value) function. The only arbitrary aspect of a utility representation would then be the choice of unit and zero.

**Exercise 7.5** ††

Show that whenever $V$ additively represents an agent's preferences, then so does any function $V'$ that differs from $V$ only by the choice of zero and unit. That is, assume that $V$ additively represents an agent's preferences, so that for some subvalue functions $V_1, V_2, \ldots, V_n$,

$$V(\langle A_1, A_2, \ldots, A_n \rangle) = V_1(A_1) + V_2(A_2) + \ldots + V_n(A_n).$$

Assume $V'$ differs from $V$ only by a different choice of unit and zero, which means that there are numbers $x > 0$ and $y$ such that $V'(\langle A_1, A_2, \dots, A_n \rangle) = x \cdot V(\langle A_1, A_2, \dots, A_n \rangle) + y$. From these assumptions, show that there are subvalue functions $V'_1, V'_2, \dots, V'_n$ such that

$$V'(\langle A_1, A_2, \dots, A_n \rangle) = V'_1(A_1) + V'_2(A_2) + \dots + V'_n(A_n).$$

Why might one think that rational preferences should be separable? Remember that we are talking about preferences over "attribute lists" that settle everything the agent ultimately cares about, with each position in a list settling one question that intrinsically matters to the agent. In our toy example, these were the size, location, and costs of their flat. More realistically, items in the attribute list might be the agent's level of happiness, their social standing, the well-being of their relatives, etc. Now, if an agent has a basic desire for, say, happiness, then we would expect that increasing the level of happiness, while holding fixed everything else the agent cares about, always is a change for the better. That is, if two worlds $w_1$ and $w_2$ agree in all respects that matter to the agent except that the agent is happier in $w_1$ than in $w_2$, then we would expect the agent to prefer $w_1$ over $w_2$. From this perspective, separability might be understood as a condition on how to identify basic desires: if an agent's preferences over some attribute lists are not separable, then the attributes don't represent (all) the agent's basic (intrinsic) desires.

**Exercise 7.6** ††

Imagine you can freely choose four courses for next semester. You assess each course by a range of criteria (such as whether the course will teach you anything useful). On this basis, you determine an overall ranking of the courses and sign up for the top four. Why might this not be a good idea?

## 7.4 Separability across time

According to psychological hedonism, the only thing people ultimately care about is their personal pleasure. But pleasure isn't constant. The hedonist conjecture leaves open how people rank different ways pleasure can be distributed over a lifetime. Un-

less an agent just cares about their pleasure at a single point in time, a basic desire for pleasure is really a concern for a lot of things: pleasure now, pleasure tomorrow, pleasure the day after, and so on. We can think of these as the "attributes" in the agent's value function. The hedonist's value function somehow aggregates the value of pleasure experienced at different times.

To keep things simple, let's pretend that pleasure does not vary within any given day. We might then model a hedonist value function as a function that assigns numbers to lists like $\langle 1, 10, -1, 2, \ldots \rangle$, where the elements in the list specify the agent's degree of pleasure today (1), tomorrow (10), the day after (-1), and so on. Such attribute lists, in which successive positions correspond to successive points in time, are called **time streams**.

A hedonist agent would plausibly prefer more pleasure to less at any point in time, no matter how much pleasure there is before or afterwards. If so, their preferences between time streams are weakly separable. Strong separability is also plausible: whether the agent prefers a certain amount of pleasure on some days to a different amount of pleasure on these days should not depend on how much pleasure the agent has on other days. It follows by Debreu's theorem that the value the agent assigns to a time stream can be determined as the sum of the subvalues she assign to the individual parts of the stream. That is, if $p_1$, $p_2$, …, $p_n$ are the agent's degrees of pleasure on days $1, 2, \ldots, n$ respectively, then there are subvalue functions $V_1, V_2, \ldots, V_n$ such that

$$V(\langle p_1, p_2, \ldots, p_n \rangle) = V_1(p_1) + V_2(p_2) + \ldots + V_n(p_n).$$

We can say more if we make one further assumption. Suppose an agent prefers stream $\langle p_1, p_2, \ldots, p_n \rangle$ to an alternative $\langle p'_1, p'_2, \ldots, p'_n \rangle$. Now consider the same streams with all entries pushed one day into the future, and prefixed with the same degree of pleasure $p_0$. So the first stream turns into $\langle p_0, p_1, p_2, \ldots, p_n \rangle$ and the second into $\langle p_0, p'_1, p'_2, \ldots, p'_n \rangle$. Will the agent prefer the modified first stream to the modified second stream, given that she preferred the original first stream? If the answer is yes, then her preferences are called **stationary**. From a hedonist perspective, stationarity seems plausible: if there's more aggregated pleasure in $\langle p_1, p_2, \ldots, p_n \rangle$ than in $\langle p'_1, p'_2, \ldots, p'_n \rangle$, then there is also more pleasure in $\langle p_0, p_1, p_2, \ldots, p_n \rangle$ than in $\langle p_0, p'_1, p'_2, \ldots, p'_n \rangle$.

It is not hard to show that if preferences over time streams are separable and stationary (as well as transitive and complete), then they can be represented by a value

function of the form

$$V(\langle A_1, \ldots, A_n \rangle) = V_1(A_1) + \delta \cdot V_1(A_2) + \delta^2 \cdot V_1(A_3) \ldots + \delta^{n-1} \cdot V_1(A_n),$$

where $\delta$ is a fixed number. The interesting thing here is that the subvalue function for any time equals the subvalue function $V_1$ for the first time, scaled by an exponential **discounting factor** $\delta^i$.

If a hedonist has strongly separable and stationary preferences, then her preferences over time streams are fixed by two things: how much she values present pleasure, and how much she discounts the future. If $\delta = 1$, the agent values pleasure equally, no matter when it occurs. If $\delta = 1/2$, then one unit of pleasure tomorrow is worth half as much as to the agent as one unit today; the day after tomorrow it is worth a quarter; and so on.

**Exercise 7.7 †**

Consider the following streams of pleasure:

S1: $\langle 1, 2, 3, 4, 5, 6, 7, 8, 9 \rangle$
S2: $\langle 9, 8, 7, 6, 5, 4, 3, 2, 1 \rangle$
S3: $\langle 1, 9, 2, 8, 3, 7, 4, 6, 5 \rangle$
S4: $\langle 9, 1, 8, 2, 7, 3, 6, 4, 5 \rangle$
S5: $\langle 5, 5, 5, 5, 5, 5, 5, 5, 5 \rangle$

Assuming present pleasure is valued in proportion to its degree, so that $V_1(p) = p$ for all degrees of pleasure $p$, how would a hedonist agent with separable and stationary preferences rank these streams, provided that (a) $\delta = 1$, (b) $\delta < 1$, (c) $\delta > 1$? (You need to give three answers.)

Even if you're not a hedonist, you probably care about some things that can occur (and re-occur) at different times: talking to friends, going to concerts, having a glass of wine, etc. The formal results still apply. If your preferences over the relevant time streams are separable and stationary, then they are fixed by your subvalue function for the relevant events (talking to friends, etc.) right now and by a discounting parameter $\delta$.

Some have argued that stationarity and separability across times are requirements

of rationality. Some have even suggested that the only rationally defensible discounting factor is 1, on the ground that we should be impartial with respect to different parts of our life.

One argument in favour of stationarity is that it thought to be required to protect the agent from a kind of disagreement with her future self. To illustrate, suppose you prefer getting £100 now to getting £105 tomorrow, but you also prefer £105 in 11 days to £100 in 10 days. These preferences violate stationarity. For if you prefer $\langle £100, £0, \dots \rangle$ to $\langle £0, £105, \dots \rangle$, where the entries in the positions specifying how much money you get on successive days, then by stationarity you also prefer $\langle £0, £100, £0, \dots \rangle$ to $\langle £0, £0, £105, \dots \rangle$, and $\langle £0, £0, £100, £0, \dots \rangle$ to $\langle £0, £0, £0, £105, \dots \rangle$, and so on. £100 in 10 days should be preferred to £105 in 11 days. Now suppose your (non-stationary) preferences remain the same for the next 10 days. At the end of this time, you still prefer £100 now over £105 tomorrow. But your "now" is what used to be "in 10 days". Your new preferences disagree with those of your earlier self, in the sense that what you now regard as better is what your earlier self regarded as worse. This kind of disagreement is called **time inconsistency**.

Empirical studies suggest that time inconsistency is pervasive. People often prefer their future selves to study, eat well, and exercise, but choose burgers and TV for today.

These preferences do look problematic. Other violations of stationarity, and even separability across time, however, look fine. For example, suppose you value having a glass of wine every now and then. But only now and then; you don't want to have wine every day. It seems to follow that your preferences violate both separability and stationarity. You violate stationarity because even though you might prefer a stream $\langle \text{wine}, \text{no wine}, \text{no wine}, \dots \rangle$ to $\langle \text{no wine}, \text{no wine}, \text{no wine}, \dots \rangle$, your preference reverses if both streams are prefixed with wine (or many instances of wine). You violate separability because whether you regard having wine in $n$ days as desirable depends on whether you will have wine right before or after these days. Even if an agent only cares about pleasure, it is not obvious why a rational agent might not (say) prefer relatively constant levels of pleasure over wildly fluctuating levels, or the other way round.

So stationarity and separability over time don't look plausible as general requirements of rationality. But one might say something similar to what I said at the end of section 7.3. If your preferences over time streams are not separable, then arguably the items in the time streams do not represent all your basic desires. If, for example,

you have a preference for constant levels of pleasure, then your basic desires don't just pertain to how much pleasure you have today, how much pleasure you have tomorrow, and so on. You have a further basic desire: that your pleasure be constant from day to day.

**Exercise 7.8 ††**

Are your preferences in the wine example time-inconsistent, in the sense that what you prefer for your future self is not what your future self prefers for itself?

Let's briefly return to the fact that people often choose vice for today and virtue for tomorrow. What might show up here is that our preferences have different sources (as I emphasized in chapter 5). When we reflect on having fries or salad now, we are more influenced by spontaneous cravings than when we consider the same options for tomorrow.

We could represent different sources of value by different subvalue functions. We might, for example, have a subvalue function $V_c$ that measures the extent to which your present cravings are satisfied, and another subvalue function $V_m$ that measures to what extent you live in accordance with your moral convictions. Your overall value function is some kind of aggregate of these components. Here, too, separability is plausible. If, for example, one world is by your lights morally better than another, and the two worlds are equally good with respect to all your other motives (your cravings are equally satisfied in either, etc.), then you plausibly prefer the first world to the second. This suggests that different sources of value combine in an additive manner.

## 7.5 Separability across states

An agent faces a choice between some acts. According to the MEU Principle, the agent should evaluate each option $A$ by its expected utility

$$\mathrm{EU}(A) = \mathrm{U}(O_1) \cdot \mathrm{Cr}(S_1) + \mathrm{U}(O_2) \cdot \mathrm{Cr}(S_2) + \ldots + \mathrm{U}(O_n) \cdot \mathrm{Cr}(S_n),$$

where $S_1, S_2, \ldots, S_n$ are the relevant states and $O_1, O_2, \ldots, O_n$ are the outcomes of act $A$ in those states. Holding fixed the states, we can represent each available act by the list of its outcomes: $\langle O_1, O_2, \ldots, O_n \rangle$. In the mushroom problem from chapter 1, for example, eating the mushroom can be represented by the list $\langle$satisfied, dead$\rangle$, and not eating by $\langle$hungry, hungry$\rangle$, with the understanding that the first item in the list comes about if the mushroom is a paddy straw and the second if it is a death cap.

Suppose the agent ranks the available acts by their expected utility. Her preference over the relevant outcome lists then have an additive representation: they are represented by a function $V$ that assigns numbers to lists in such a way that the number assigned to any list is determined by adding up subvalues assigned to individual items on the list. This function $V$ is the EU function; the subvalues are the credence-weighted utilities of the outcomes. The subvalue of outcome $O_1$, for example, is $\mathrm{U}(O_1) \cdot \mathrm{Cr}(S_1)$.

By Debreu's theorem, rational preferences have an additive representation if and only if they are strongly separable. The MEU Principle therefore implies that an agent's preferences between the acts in a decision problem are (strongly) separable *across states*, meaning that the desirability of an outcome in one state does not depend on the outcomes in other states.

Admittedly, this is a very roundabout path to a fairly obvious result. I mention it for two reasons. First, it shows that the response to the ordinalist challenge from the previous section is closely related to the response that we met in chapter 6. Von Neumann and Ramsey, in effect, assume that rational preferences are separable across states, and that the right way to measure separable preferences construes the net utility of an act as the sum of certain values assigned to the individual outcomes.

Second, a general consequence of separability is that the relevant preferences are insensitive to "shapes" in the distribution of subvalues. For example, separable preferences cannot prefer even distributions to uneven distributions. This may seem to point at a problem with the MEU Principle. Consider the following schematic decision problem:

| | State 1 (½) | State 2 (½) |
|---|---|---|
| *A* | Outcome 1 (+10) | Outcome 1 (+10) |
| *B* | Outcome 2 (-10) | Outcome 3 (+30) |

Option *A* leads to a guaranteed outcome with utility 10, while option *B* leads either to a much better outcome or to a much worse one. The expected utilities are the same,

but one might think an agent might rationally prefer $A$ because the utility distribution $\langle 10, 10 \rangle$ is more even than $\langle$-10, 30$\rangle$. Intuitively, $A$ is safe, while $B$ is risky. We will return to this issue in the next chapter.

**Exercise 7.9** ††

Where in their axioms do Savage and von Neumann and Morgenstern assume a kind of separability across states?

## 7.6 Harsanyi's "proof of utilitarianism"

The ordinalist movement, which rejected the quantitative concept of utility, posed a challenge not only to the MEU Principle, but also to utilitarianism in ethics. According to utilitarianism, an act is right iff it brings about the best available state of the world, where the "goodness" of a state is measured by the sum of the utility of all people. Without a numerical (and not just ordinal) measure of utility, this measure of goodness breaks down. We need a new criterion for ranking states of the world.

One such criterion was proposed by Pareto. Recall that Pareto did not deny that people have preferences. If we want to rank two states of the world, we can still ask which of them people prefer. This allows us to define at least a partial order on the possible states:

**The Pareto Condition**

If everyone is indifferent between $A$ and $B$, then $A$ and $B$ are equally good; if at least one person prefers $A$ to $B$ and no one prefers $B$ to $A$, then $A$ is better than $B$.

Unlike classical utilitarianism, however, the Pareto Condition offers little moral guidance. For instance, while classical utilitarianism suggests that one should harvest the organs of an innocent person in order to save ten others, the Pareto Condition does not settle whether it would be better or worse to harvest the organs, given that the person to be sacrificed ranks the options differently than those who would be saved.

**Exercise 7.10 (The Condorcet Paradox) †**

A "democratic" strengthening of the Pareto condition might say that whenever *a majority* of people prefer $A$ to $B$, then $A$ is better than $B$. But consider the following scenario. There are three relevant states: $A, B, C$, and three people. Person 1 prefers $A$ to $B$ to $C$. Person 2 prefers $B$ to $C$ to $A$. Person 3 prefers $C$ to $A$ to $B$. If betterness is decided by majority vote, which of $A$ and $B$ is better? How about $A$ and $C$, and $B$ and $C$?

In 1955, John Harsanyi proved a remarkable theorem that seemed to rescue, and indeed vindicate, classical utilitarianism.

To begin, Harsanyi assumes that there is a betterness ordering between states of the world that extends to lotteries between such states. This is not yet a substantive premise, as we have not yet made any assumptions about the ordering.

Harsanyi's first premise is that the betterness order satisfies the axioms of von Neumann and Morgenstern. By von Neumann and Morgenstern's representation theorem, it follows that the order is represented by a ("social") utility function that ranks lotteries by their expected utility. The function is unique except for the choice of unit and zero.

Second, Harsanyi assumes that the betterness order satisfies the Pareto condition (both for states and for lotteries).

Finally, Harsanyi assumes that each person – of which he assumes for simplicity that there is a fixed number $n$ – has personal preferences between the relevant states and lotteries, and that these preferences also satisfy the von Neumann and Morgenstern axioms. The personal preferences are therefore represented by $n$ personal utility functions.

Note that the Pareto condition expresses a kind of separability of betterness across people. The assumption that social and personal utility rank lotteries by their expected utility, which follows from the von Neumann and Morgenstern construction, amounts to separability in another dimension, across states. As it turns out, Debreu's results can be strengthened for cases in which the relevant attributes are separable across two independent dimensions (here, people and states). Drawing on this result, Harsanyi showed that it follows from the above three assumptions that the individual and social preferences are represented by utility functions $\mathrm{U}_s$ and $\mathrm{U}_1, \mathrm{U}_2, \ldots, \mathrm{U}_n$ such that the social utility function $\mathrm{U}_s$ is simply the sum of the individual utility

functions $U_1, U_2, \ldots, U_n$: for any state or lottery $A$,

$$U_s(A) = U_1(A) + U_2(A) + \ldots + U_n(A).$$

And this looks just like classical utilitarianism.

On closer inspection, things are less clear-cut. For a start, recall that the utility functions established by von Neumann and Morgenstern's representation theorem have arbitrary units and zeroes. If according to one adequate representation of our preferences, my utility for a given state is 10 and yours is -1, then according to another, equally adequate representation, my utility for the state is 10000 and yours 0.07. Harsanyi's theorem only tells us is that there is *some* utility representation of our individual preferences relative to which our utilities add up to social utility. This is compatible with the assumption that social utility is almost entirely determined by the preferences of a single person, because their utilities are scaled so as to dwarf all the others. This does not look like classical utilitarianism. (We can't, for example, infer that innocent people should be slaughtered for their organs.)

Also, anyone who is not already a utilitarian should probably reject the Pareto Condition. The condition implies that the only thing that matters, from a moral perspective, is the satisfaction of people's preferences. If anything else had any moral weight – whether people's rights are respected, whether animals suffer, whether God's commands are obeyed, or whatever – then it could happen that everyone is indifferent between $A$ and $B$, and yet $A$ is actually better.

In general, if someone seems to offer a mathematical proof of a substantive normative principle, you can be sure that either the principle isn't really established or it has been smuggled in through the premises.

**Essay Question 7.1**

Do you think time inconsistency is a requirement of rationality? Can you explain why, or why not?

### Sources and Further Reading

The topic of this chapter is rarely discussed in mainstream philosophy, although its importance is occasionally recognized. See, for example, Philip Pettit, "Decision Theory and Folk Psychology" (1991). In economics, our topic is commonly known as "multi-attribute utility theory". Ralph L. Keeney and Howard Raiffa, *Decisions with Multiple Objectives* (1976/1993) is a classical, and very detailed, exposition. Paul Weirich, *Decision Space* (2001) explores the area from a more philosophical angle. The theorem by Debreu that I've referred to is from his 1960 article "Topological methods in cardinal utility". More results along the same line are surveyed in David Krantz et al., *Foundations of Measurement, Vol. I: Additive and Polynomial Representations* (1971).

For an in-depth discussion of preferences over time streams, including relevant empirical results, see Shane Frederick, George Loewenstein, and Ted O'Donoughue, *Time Discounting and Time Preference: A Critical Review* (2002).

A good introduction to Harsanyi's argument for utilitarianism is John Broome, "General and Personal Good: Harsanyi's Contribution to the Theory of Value" (2015).