5 Utility

5.1 Two conceptions of utility

Daniel Bernoulli realized that rational agents don't always maximize expected monetary payoff: £1000 has more utility for a pauper than for a rich man. But what is utility?

Until the early 20th century, utility was widely understood to be some kind of psychological quantity, often identified with degree of pleasure and absence of pain. On that account, an outcome has high utility for an agent to the extent that it increases the agent's pleasure and/or decreases her pain.

Let's assume for the sake of the argument that one can represent an agent's total amount of pleasure and pain by a single number – the agent's "degree of pleasure". Can we understand utility as degree of pleasure? The answer depends on what role we want the concept of utility to play.

One such role lies in ethics. According to **utilitarianism**, an act is morally right just in case it would bring about the greatest total utility for all people. In this context, identifying utility with degree of pleasure implies that only pleasure and pain have intrinsic moral value; everything else – autonomy, integrity, respect of human rights, and so on – would be morally relevant only insofar as it causes pleasure or pain. This assumption is known as **ethical hedonism**. We will not pursue it any further.

Exercise 5.1 †

Suppose that money has declining marginal utility, and that the utility of money is the same for all people, so that a net wealth of £1000 is as good for me as it is for you. Without any further assumptions about utility, it follows that if one person has more money than another, then their combined utility would increase if the wealthier person gave some of her money to the

poorer person, decreasing the gap in wealth. Explain why.

Another role for a concept of utility lies in the theory of practical rationality. According to the MEU Principle, practically rational agents choose acts that maximize the credence-weighted average of the utility of the possible outcomes. If we identify utility with degree of pleasure, the MEU principle turns into what we might call the 'MEP Principle':

The MEP Principle

Rational agents maximize their expected degree of pleasure.

An act's *expected degree of pleasure* is the probability-weighted average of the degree of pleasure that might result from the act.

The MEP Principle is a form of **psychological hedonism**. Psychological hedonism is the view that the only thing that ultimately motivates people is their own pleasure and pain.

The founding fathers of modern utilitarianism, Jeremy Bentham and John Stuart Mill, had sympathies for both ethical hedonism and psychological hedonism. As a consequence, the two conceptions of utility – the two roles associated with the word 'utility' – were not properly distinguished. Today, both kinds of hedonism have long fallen out of fashion, but the two conceptions are still often conflated.

For the most part, contemporary utilitarians hold that the standard of moral rightness is the total *welfare* or *well-being* produced by an act, which is not assumed to coincide with total degree of pleasure. Thus 'utility' is nowadays often used as a synonym for 'welfare' or 'well-being'. But the word is also widely used in the other sense, to denote whatever motivates (rational) agents.

Some have argued that the two uses actually coincide: that the only thing that motivates rational agents is their own welfare or well-being. This may or may not be true. But it needs to be backed up by data and argument; it does not become true through sloppy use of language.

In these notes, 'utility' is only used in the second sense. The utility of a outcome measures the extent to which the agent in question wants the outcome to obtain. We do not assume that the only thing agents ultimately want is to increase their degree

of pleasure, their welfare, their well-being, or anything like that.

Note that psychological hedonism, or the slightly more liberal claim that people only care about their welfare, is at most a contingent fact about humans. One can easily imagine agents who are motivated by other things. We can imagine a mother who knowingly takes on hardships for the benefit of her children, or a soldier who intentionally chooses a painful death in order to save her comrades. Psychological hedonists hold that humans would never consciously do such things: whenever an agent sacrifices her own good to benefit others, she mistakenly believes that her choice will actually make herself better off than the alternatives. Again, we don't need to argue over whether this is true. The important point is that utility, as we use the term, does not *mean* the same as degree of pleasure or welfare or well-being.

A hedonist might object that while it is conceivable that an agent is motivated by things other than her personal pleasure, such agents would be irrational. After all, the MEP Principle only states that *rational* agents maximize their expected degree of pleasure; it doesn't cover irrational agents.

This brings us to a tricky issue. What do we mean by 'rational'? The label 'rational' is sometimes associated with cold-hearted selfishness. On this usage, a rational agent always looks out for her own advantage, with no concern for others. This idea of "economic rationality" has its use, but it is not our topic. The kind of rationality we're interested in is a more minimal notion. Intuitively, it is the idea of "making sense". If you want to reduce animal suffering, and you know you can achieve this by eating less meat, then it makes sense that you eat less meat. If you are sure that a picnic will be cancelled if it is raining, and you see that it is raining, then it doesn't make sense to believe that the picnic will go ahead. The model we are studying is a model of agents who "make sense" in this kind of way.

Even if we were interested in the cold-hearted and selfish sense of rationality, we should not define utility as degree of pleasure or welfare. Consider a hypothetical agent who cares not just about herself, who sacrifices some of her own good to reduce the pain of others. The agent is "irrational" in the cold-hearted and selfish sense. But what is irrational about her? Does the fault lie in her beliefs, in her goals, or in the way she brings these together to make choices? Plausibly, the "fault" lies in her goals. Her concern for others is what goes against the standards of cold-hearted and selfish rationality. But if we were to define utility as degree of pleasure or welfare, we would have to say that the agent violates the basic norm of practical rationality, the MEU Principle.

The point generalizes. Consider a person in an abusive relationship who is manipulated into doing things that hurt or degrade her. We might reasonably think that the person shouldn't do these things; it is against her interest to do them. But what is at fault? Arguably, the fault lies in her (manipulated) desires. What the person does may well be in line with what she wants to achieve – in particular, with her strong desire to please her partner. But a healthy, self-respecting person, we think, should have other desires.

By understanding utility as a measure of whatever the agent in question desires, we do not automatically sanction these desires as rational or praiseworthy. Our usage of 'utility' allows us to say that the person in the abusive relationship shouldn't do what she is doing, because she should have different desires that would not support her actions.

5.2 Sources of utility

An outcome's utility measures the extent to which the agent is motivated to bring about the outcome. I will often say that this is the degree to which the agent *desires* the outcome, but we need to keep in mind that the word 'desire' can be misleading. For one thing, we need to cover "negative desire". Being hungry might have greater utility for you than being dead, even though you do not desire either. More importantly, 'desire' is often associated with a particular type of motivational state. I might say that I got up early in the morning despite my strong desire to stay in bed; I got up not because I desired to get up, but because I had to. On this usage, my desires contrast with my sense of duty.

Utility comprises everything that motivates the agent, all the reasons she has for and against a particular action. As such, 'utility' is an umbrella term for a diverse set of psychological states or events. We can be motivated by bodily cravings, by moral commitments, by our image of the kind of person we want to be, by an overwhelming feeling of terror or love, and so on. These factors need not be conscious. There is good evidence that our true motives are often not what we believe or say they are. An agent's utility function represents her true motives, and all of them.

Why should we believe that all the factors that motivate an agent can be amalgamated into a single numerical quantity? Would it not be better to allow for a whole range of utility functions: moral utility, emotional utility, and so on? We could cer-

tainly do that. But there are reasons to think that there must also be an amalgamated, all-things-considered utility (although the determinacy and numerical precision of utility functions is obviously an idealisation). When you face a decision, you have to make a single choice. You can't choose one act on moral grounds and a different act on emotional grounds. Somehow, all your motives and reasons have to be weighed against each other to arrive at an overall ranking of your options.

We will have a brief look at the weighing of different considerations in chapter 7, but to a large extent this is really a topic for empirical psychology and neuroscience. If it turns out that there are 23 distinct factors that influence our motivation in an intricate network of inhibition and reinforcement, then so be it. We will model the whole network by a single utility function, staying neutral on "lower-level" details that can vary from agent to agent. But it's important to keep in mind that a lot of interesting and complicated psychology is hiding in our seemingly simple concept of utility.

Our use of 'utility' matches the official usage in economics textbooks. In practice, however, economists, along with psychologists and other social scientists, often ignore most sources of human motivation and fall back onto a naive interpretation of utility in terms of material wealth.

Consider the following example.

Example 5.1 (The endowment effect)

Emily's favourite band is playing in town. The tickets cost £70, and Emily is not willing to pay that much. Emily's neighbour Fred has a ticket but can't go to the concert, so he sells his ticket to Emily for £50. The day before the concert, all the tickets are sold out, and another of Emily's neighbours, George, asks Emily if she would sell her ticket to him for £70. Emily declines.

The kind of behaviour displayed by Emily is quite common. It is also widely claimed to contradict expected utility theory. The idea is that if Emily isn't willing to sell the ticket for £70, then having the ticket is worth more than £70 to her, so she should have bought the ticket for £70 at the outset.

But it is not hard to understand what happened. By the time George approaches Emily, she has made plans for going to the concert; she is looking forward to the event. Giving up the ticket now would be a serious disappointment. It would also

mean that she has to make new plans for tomorrow evening. In short, for Emily, *giving up a ticket she previously owned* is worse than *never owning the ticket in the first place*. Given these preferences, her behaviour is perfectly in line with the MEU Principle.

Exercise 5.2 ††

Draw an adequate decision matrix for Emily's decision problem when she first considered buying the ticket, and another matrix for her decision problem when she was approached by George. (There is no relevant uncertainty, so the matrices have only one state.)

Emily's behaviour does not violate the MEU Principle. Indeed, no pattern of behaviour whatsoever can, all by itself, violate the MEU Principle. After all, for any pattern of behaviour, we can imagine that the agent has a basic desire to display just that pattern of behaviour. Displaying the behaviour then evidently maximizes expected utility.

If we are interested in the MEU Principle as a descriptive hypothesis about real people's choices, and we interpret 'utility' to measure whatever people care about, then the Principle is, in a sense, unfalsifiable. Whenever an agent seems to violate the MEU Principle, we can posit beliefs and desires that would make her choices conform to the principle. Psychologists and social scientists sometimes point at this fact in support of their decision to re-interpret 'utility' as a function of material goods. A scientific hypothesis, they assume, is only worth taking seriously if it can be falsified. But a lot of respectable scientific theories are unfalsifiable *in isolation*. Philosophers of science have long realized that one can generally only test scientific hypotheses in conjunction with a whole range of background assumptions.

The same is true for the MEU Principle, understood as a descriptive hypothesis about human behaviour. *Given* some assumptions about an agent's beliefs and desires, we can easily find that her choices do not conform to the MEU Principle. And often we have good evidence about the relevant beliefs and desires. For example, it is safe to assume that participants in the world chess tournament want to win their games, and that they are aware of the current position of the pieces in the game.

We are going to understand utility as a measure of whatever the agent cares about. It might be worth pointing out, however, that our model can be usefully applied

with other conceptions of utility. For example, one might ask what an agent should do, from a moral perspective, in a situation like the Miner's Problem (from chapter 1) where she lacks crucial information about the world. A tempting idea is that the agent should maximize expected *moral utility*, where the moral utility of an outcome is defined by our ethical theory (utilitarianism, perhaps). Similarly, the board of directors of a company may want to know what corporate decisions would ideally promote shareholder value in the light of such-and-such shared information. Here the relevant utility function could be derived from the stipulated goal of promoting shareholder value, and the "credence" function could be derived from the shared information. Neither of these needs to match the beliefs and desires of any individual member of the board.

Exercise 5.3 †††

Some choices can predictably change our desires. One might argue that in such a case, a rational agent should be guided not by her present desires, but by the desires she will have as a result of her choice.
For example, suppose you can decide right now how many drinks you will have tonight: zero, one, or two. (You have to order the drinks in advance and can't change the order at the time.) If you're sober, you prefer to have one drink rather than zero or two. But if you have a drink, you often prefer to have another. Draw a matrix for your decision problem, assuming that your goal is to maximize your expected future utility. (Can you see why we don't need to change the MEU Principle or our definition of utility?)

5.3 The structure of utility

Now that we know what utility is, let's have a closer look at its formal structure.

First of all, what are the bearers of utility? In ordinary language, we often say that people desire *things*: tea, cake, a concert ticket, a larger flat. This fits the economist tradition of identifying the bearers of utility with material goods, or "commodity bundles". But if we want to allow for the entire range of possible desires, we need a broader conception. Perhaps you desire that your friends are happy, that it won't rain tomorrow, that so-and-so will win the next elections. Here the object of desire isn't a particular thing, but a possible state of the world. Even when we say that people

desire things, plausibly the desire is really directed at a possible state of the world. When you desire tea, you desire to *drink the tea*. Your desire wouldn't be satisfied if I gave you a certificate of ownership for a cup of tea that is locked away in a safe.

So we'll assume that the objects of desire are the same kinds of things as the objects of belief: propositions, or possible states of the world. As in the case of belief, we don't distinguish between logically equivalent states of the world. If you assign high utility to drinking tea then you also assign high utility to *drinking tea or coffee but not coffee*.

Let's study how an agent's desires towards logically related propositions are related to one another. Suppose you assign high utility to the proposition that it won't rain tomorrow (perhaps because you want to go on a picnic). Then you should plausibly assign *low* utility to the proposition that it *will* rain. You can't hope that it will rain and also that it won't rain. In this respect, desire resembles belief: if you are confident that it will rain, you can't also be confident that it won't rain. The Negation Rule of probability captures the exact relationship between $\mathrm{Cr}(A)$ and $\mathrm{Cr}(\neg A)$, stating that $\mathrm{Cr}(\neg A) = 1 - \mathrm{Cr}(A)$. Does the rule also hold for utility? More generally, do utilities satisfy the Kolmogorov axioms? It will be instructive to go through the three axioms.

Kolmogorov's axiom (i) states that probabilities range from 0 to 1. If there are upper and lower bounds on utility, we could adopt axiom (i) for utilities as a convention of measurement: we simply use 1 for the upper bound and 0 for the lower bound. However, it is not obvious that there are such bounds. Couldn't there be an infinite series $A_1, A_2, A_3, \ldots$ of states of increasing utility in which the difference in utility between successive states is always the same? If there is such a series, then utility can't be measured by numbers between 0 and 1. Philosophers are divided over the question. Some think utility must be **bounded**, others think it can be unbounded. There are arguments for both sides. We will not pause to look at them.

Kolmogorov's axiom (ii) states that logically necessary propositions have probability 1. If utilities satisfy the probability axioms, this would mean that logically necessary propositions have maximal utility. However much you desire that it won't rain tomorrow, your desire that *it either will or won't rain* should be at least as great.

This does not look plausible. Intuitively, if something is certain to be the case, it makes no sense to desire it. But this could mean two things. It could mean that degrees of desire are not even defined for logically necessary propositions. Or it could mean that an agent should always be indifferent towards logically necessary propo-

sitions – neither wanting them to be the case nor wanting them to not be the case. Our common-sense conception of desire arguably sides with the first option: if you are certain of something, even asking how strongly you desire it seems odd. But the issue isn't clear. For our purposes, it proves more convenient to go with the second option. We will say that even logically necessary propositions have well-defined utility, and that their utility measures the point between "positive" and "negative" desire. If you positively want something to be the case, the utility you assign to it is greater than the utility of a tautology. If you want something not to be the case, its utility is lower than that of a tautology. Some authors make this more concrete by adopting a convention that logically necessary propositions always have utility 0.

Axiom (iii) states that if A and B are logically incompatible, then the probability of $A \vee B$ equals the sum of the probabilities of A and B. To illustrate, suppose there are three possible locations for a picnic: Alder Park, Buckeye Park, and Cedar Park. Alder Park and Buckeye Park would be convenient for you; Cedar Park would not. Now how much do you desire that the picnic takes place in *either Alder Park or Buckeye Park*? If axiom (iii) holds for utilities, then if you desire Alder Park and Buckeye Park to equal degree x, then your utility for the disjunction should be $2x$: you should be more pleased to learn that the picnic takes place in either Alder Park or Buckeye Park than to learn that it takes place in Alder Park. That's clearly wrong. Axiom (iii) also fails.

What is the true connection between the utility of $A \vee B$ and the utilities of A and B? Intuitively, if A and B have equal utility x, then the utility of $A \vee B$ should also be x. What if the utilities of A and B are not equal? What if, say, $\mathrm{U}(A) > \mathrm{U}(B)$? Then the utility of $A \vee B$ should plausibly lie in between the utilities of A and B:

$$\mathrm{U}(A) \geq \mathrm{U}(A \vee B) \geq \mathrm{U}(B).$$

That is, if Alder Park is your first preference and Buckeye your second, then the disjunction *either Alder Park or Buckeye Park* can't be worse than Buckeye Park or better than Alder Park. But where does $\mathrm{U}(A \vee B)$ lie in between $\mathrm{U}(A)$ and $\mathrm{U}(B)$? At the mid-point?

Suppose you prefer Alder Park to Buckeye Park, and Buckeye Park to Cedar Park. You think it is highly unlikely that the picnic will take place in Buckeye Park. Now how pleased would you be to learn the picnic won't take place in Cedar Park – equivalently, that it will take place either in Alder Park or in Buckeye Park? You should

be quite pleased. If you're confident that B is false, then $\mathrm{U}(A \vee B)$ should plausibly be close to $\mathrm{U}(A)$. If you're confident that A is false, then $\mathrm{U}(A \vee B)$ should be near $\mathrm{U}(B)$.

Your utilities depend on your beliefs! On reflection, this should not come as a surprise. A lot of the things we desire we only desire because we have certain beliefs. If you want to buy a hammer to hang up a picture, then your desire for the hammer is based (in part) on your belief that the hammer will allow you to hang up the picture.

Here is the general rule for $\mathrm{U}(A \vee B)$, assuming A and B are incompatible. The rule was discovered by Richard Jeffrey in the 1960s and is our *only* basic rule of utility, apart from the assumption that logically equivalent propositions have the same utility.

Jeffrey's Axiom

If A and B are logically incompatible and $\mathrm{Cr}(A \vee B) > 0$ then

$$\mathrm{U}(A \vee B) = \mathrm{U}(A) \cdot \mathrm{Cr}(A/A \vee B) + \mathrm{U}(B) \cdot \mathrm{Cr}(B/A \vee B).$$

In words: the utility of $A \vee B$ is the weighted average of the utility of A and the utility of B, weighted by the probability of the two disjuncts, conditional on $A \vee B$.

Why 'conditional on $A \vee B$'? Why don't we simply weigh the utility of A and B by their unconditional probability? Because then highly unlikely propositions would automatically have a utility near 0. If you are almost certain that the picnic will take place in Cedar Park, both $\mathrm{Cr}(\textit{Alder Park})$ and $\mathrm{Cr}(\textit{Buckeye Park})$ will be close to 0. But the mere fact that a proposition is unlikely does not make it undesirable. To evaluate the desirability of a proposition, we should bracket its probability. That's why Jeffrey's Axiom defines $\mathrm{U}(A \vee B)$ as the probability-weighted average of $\mathrm{U}(A)$ and $\mathrm{U}(B)$ *on the supposition that* $A \vee B$ *is true*.

Exercise 5.4 ††

You would like to win the lottery because that would allow you to travel the world, which you always wanted to do. Let *Win* be the proposition that you win the lottery, and *Travel* the proposition that you travel the world. Note that *Win* is logically equivalent to $(\textit{Win} \wedge \textit{Travel}) \vee (\textit{Win} \wedge \neg \textit{Travel})$, and thus has

the same utility. Suppose $\mathrm{U}(\mathit{Win} \wedge \mathit{Travel}) = 10$, $\mathrm{U}(\mathit{Win} \wedge \neg \mathit{Travel}) = 0$, and your credence that you will travel the world on the supposition that you will win the lottery is 0.9. By Jeffrey's axiom, what is $\mathrm{U}(\mathit{Win})$?

Exercise 5.5 ††

At the beginning of this section, I argued that if $\mathrm{U}(\neg A)$ is high, then $\mathrm{U}(A)$ should be low, and vice versa. Let's use the utility of the tautology $A \vee \neg A$ as a neutral point of reference, so that $\mathrm{U}(A \vee \neg A) = 0$. From this assumption, and Jeffrey's axiom, it follows that $\mathrm{U}(\neg A) > 0$ just in case $\mathrm{U}(A) < 0$. More precisely, it follows that

$$\mathrm{U}(A)\mathrm{Cr}(A) = -\mathrm{U}(\neg A)\mathrm{Cr}(\neg A).$$

Can you show how this follows? (It's not as hard as it looks. Hint: A is logically incompatible with $\neg A$.)

Exercise 5.6 †††

Derive the following rule from Jeffrey's axiom and the rules of probability: if A, B, and C are incompatible with one another and $\mathrm{Cr}(A \vee B \vee C) > 0$, then

$$\begin{aligned}\mathrm{U}(A \vee B \vee C) = \mathrm{U}(A) \cdot \mathrm{Cr}(A/A \vee B \vee C) + \mathrm{U}(B) \cdot \mathrm{Cr}(B/A \vee B \vee C) \\ + \mathrm{U}(C) \cdot \mathrm{Cr}(C/A \vee B \vee C).\end{aligned}$$

Exercise 5.6 shows that Jeffrey's axiom holds not only for two, but also for three propositions. We can similarly extend it to four, five, six, or any other (finite) number of propositions. It follows that if a proposition A divides into finitely many "possible worlds", then the utility of A is the weighted average of the utility of these worlds, weighted by their probability conditional A.

Exercise 5.7 ††

Let's say that an agent *desires* a proposition A iff $\mathrm{U}(A) > \mathrm{U}(\neg A)$. (Equivalently, iff $\mathrm{U}(A) > \mathrm{U}(A \vee \neg A)$.) One might have thought that whenever a rational agent desires a conjunction $A \wedge B$, then she also desires each conjunct. On the present analysis of desire, however, this is false. For example, if A is the proposition that I will break my leg in an accident today, and B is the proposition that I will get a billion pounds compensation, then I desire $A \wedge B$, but I do not desire A. Give similar counterexamples to the following hypotheses.

(a) Whenever an agent desires A, then she desires $A \vee B$.
(b) Whenever an agent desires A and desires B, then she desires $A \wedge B$.

Exercise 5.8 ††

Hans is convinced that there is a ghost in his attic. He fears that the ghost will keep him awake tonight. The following statement is intuitively true: 'Hans desires that the ghost in his attic will be quiet tonight'. Some have found this puzzling because Hans would much prefer that there is no ghost in his attic. If he had full control over what happens in his attic, he would see to it that it is free of ghosts, not that it is occupied by a quiet ghost. Explain why Hans nonetheless desires that there is a quiet ghost in his attic, assuming the analysis of 'desire' from the previous exercise.

5.4 Basic desire

I have presented Jeffrey's axiom as the sole formal requirement on rational utility. Even that much is controversial. Many economists and philosophers hold that rationality imposes no constraints at all on an agent's desires. (In a way, this is the opposite extreme of the hedonist doctrine that rational agents desire nothing but their own pleasure.) The idea was memorably expressed by David Hume in his *Treatise of Human Nature*:

> 'tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse

> my total ruin, to prevent the least uneasiness of an Indian or person unknown to me.

Hume held that our basic desires are not responsive to evidence, reason, or argument. If your ultimate goal is to help some distant stranger, there is no non-circular argument that could prove your goal to be wrong, nor could we fault you for not taking into account any relevant evidence. Whatever facts you might find out about the world, you could coherently retain your ultimate goal of helping the stranger.

For Hume, beliefs and desires are in principle independent. What you believe is one thing, what you desire is another. Beliefs try to answer the question: what is the world like? Desires answer an entirely different question: what do you want the world to be like? On the face of it, these two questions really appear to be logically independent. Two agents could in principle give the same answer to the first question and different answers to second, or the other way around.

What we have seen in the previous section seems to contradict these intuitions. We have seen that an agent's utilities are thoroughly entangled with her credences. Indeed, we can read off an agent's credence in any proposition A from her utilities, assuming the utilities obey Jeffrey's axiom, the credences obey the probability axioms, and the agent is not disinterested in A. Here is how.

By Jeffrey's axiom,

$$\mathrm{U}(A \vee \neg A) = \mathrm{U}(A) \cdot \mathrm{Cr}(A) + \mathrm{U}(\neg A) \cdot \mathrm{Cr}(\neg A).$$

By the Negation Rule, we can replace $\mathrm{Cr}(\neg A)$ by $1 - \mathrm{Cr}(A)$. Multiplying out, we get

$$\mathrm{U}(A \vee \neg A) = \mathrm{U}(A) \cdot \mathrm{Cr}(A) + \mathrm{U}(\neg A) - \mathrm{U}(\neg A) \cdot \mathrm{Cr}(A).$$

Now we solve for $\mathrm{Cr}(A)$:

$$\mathrm{Cr}(A) = \frac{\mathrm{U}(A \vee \neg A) - \mathrm{U}(\neg A)}{\mathrm{U}(A) - \mathrm{U}(\neg A)}.$$

The ratio on the right-hand side is defined whenever $\mathrm{U}(A) \neq \mathrm{U}(\neg A)$, which I meant by the agent being "not disinterested" in A.

What is going on here? Have we refuted Hume? Have we shown that an agent's beliefs are *part of her desires*?

Of course not – or not in any interesting sense. We need to distinguish **basic desires** from **derived desires**. If you are looking for a hammer to hang up a picture, your desire to find the hammer is not a basic desire. It is derived from your desire to hang up the picture and your belief that you need a hammer to achieve that goal. By contrast, a desire to be free from pain is typically basic. If you want a headache to go away, this is usually not (or not only) because you think having no headache is associated with other things you desire. You simply don't want to have a headache, and that's the end of the story.

When Hume claimed that desires are independent of beliefs, he was talking about basic desires.

How are basic desires related to an agent's utility function?

Imagine an agent whose *only* basic desire is to be free from pain, and let's pretend this is an all-or-nothing matter. The utility this agent gives to being free from pain then does not depend on her beliefs. Moreover, all states in which she is free from pain are equally good, equally desirable, no matter what she believes. Being pain-free *and rich*, for example, is equally desirable as being pain-free *and poor*. Both have the same utility as being pain-free itself.

Let's say that a proposition has *uniform utility* if the agent does not care how the proposition is realized: all subsets of the proposition (understood as a set of possible worlds) have equal utility.

Next, imagine an agent with two basic desires: being pain-free and being rich. These are logically independent, so there are four combinations: (1) being pain-free and rich, (2) being pain-free and poor, (3) being in pain and rich, and (4) being in pain and poor. Being pain-free no longer has uniform utility, since the worlds where the agent is pain-free divide into (better) worlds where the agent is pain-free and rich and (worse) worlds where the agent is pain-free and poor. As a consequence, the utility of being pain-free now depends on the agent's beliefs: the stronger she believes that she is rich if she is pain-free, the more she desires being pain-free.

The four combinations of being pain-free and being rich, however, have uniform utility. All worlds in which the agent is, say, pain-free and poor are equally desirable (pretending these are all-or-nothing matters). I'll say that these combinations are the agent's **concern**. Intuitively, a concern is a proposition that specifies everything the agent ultimately cares about. Formally, a concern is simply a proposition with uniform utility.

An agent's basic desires are reflected in the utility she attaches to her concerns.

The choice of concerns, and the utilities they get, is independent of the agent's beliefs. An agent's credence function is only needed to determine the ("derived") utility of propositions that are not among the agent's concerns.

Exercise 5.9 †

There's a party, and at first you want to be invited. Then you hear that Bob will be there, and you no longer want to be invited. Then you hear that there will be free beer, and you want to be invited again. Your desire seems to change back and forth. Nonetheless, your basic desires may have remained the same throughout. Explain how your fluctuating attitude might have come about without any change in basic desires.

Exercise 5.10 ††

Assume there are finitely many possible worlds. Explain why an agent's basic desires can be represented by the utility she assigns to individual worlds. (This assignment is sometimes called the agent's *value function*.)

Exercise 5.11 †††

Assume an agent's basic desires (her concerns and their utility) remain the same while she conditionalizing on an undesirable proposition E (with $\mathrm{U}_{\mathrm{old}}(E) < \mathrm{U}_{\mathrm{old}}(\neg E)$). How does this affect the utility of the logically necessary proposition? Explain your answer.

Essay Question 5.1

Do you agree with Hume that there are no rational constraints on basic desires? If so, try to defend this view. If not, try to argue against it.

Sources and Further Reading

Chapter 6 (“Game Theory and Rational Choice”) of Simon Blackburn, *Ruling Passions* (1998) eloquently defends the idea that one shouldn’t constrain what rational agents may care about in the theory of practical rationality. John Broome, “ ‘Utility’ ” (1991) provides some more background and details on the two conceptions of utility.

The formal theory of utility from section 5.3 comes from chapter 5 of Richard Jeffrey, *The Logic of Decision* (1965/1983).

The assumption that the objects of utility are the same kinds of things (propositions) as the objects of credence is common in philosophy, but not in other disciplines, where utilities are often assumed to pertain to “outcomes” that are distinct from the “states” to which credences pertain.

My distinction between basic desires and derived desires resembles a common distinction in economics between “direct utility” and “indirect utility”. It also resembles the popular distinction between “intrinsic” and “instrumental” desire. But note that if A and B are concerns, then a desire for their disjunction $A \vee B$ is derived, although a disjunction is not intuitively instrumental to its disjuncts. The label ‘concern’ is mine.

The puzzle of Hans and the ghost in his attic is from Paul Elbourne, “The existence entailments of definite descriptions” (2010).