principle of expert deference makes no reference to time or to personal identity and subsumes Modified Reflection as a special case.

4.2 Reflection for Preferences

I have so far focused on a reflection principle which applies to beliefs or credences. But insofar as such a principle is attractive, it is natural to expect a similar principle to hold for conative attitudes like desires and preferences.[6] Such a principle for preferences would say:

Preference Reflection
It is a requirement of rationality that if you believe that you will later prefer *A* to *B*, then you now prefer *A* to *B*, unless you believe that you might be irrational in the future or have lost evidence.[7]

Preference Reflection has some initial plausibility. First, as already noted, many philosophers have accepted some sort of reflection principle for beliefs, and so there is some prima facie reason to think there would be an analogous principle for desires.

Second, this principle gives intuitively correct results in many cases (though not all, as I will explain below). When I wake up feeling groggy and think about whether to get up and go to the gym or instead catch another hour of sleep, I often motivate myself by reflecting on the fact that if I go to the gym, I'll be glad I did so, whereas if I sleep in, I will regret it. Preference Reflection vindicates this sort of "I'll be glad I did it" reasoning.[8]

Third, and relatedly, one might think, with Nagel (1970), that something like Preference Reflection is necessary to underpin the rationality of prudence, understood as something like practical foresight. Nagel argues that "there is reason to do not only what will promote that for which there is presently a reason, but also that for which it is expected that there *will* be a reason" (1970, 36). Nagel is making about a claim about reasons in general, but he is clear that this thesis applies in particular to reasons stemming from desires, so that if you believe that you will

[6] See Arntzenius (2008) and Harman (2009) for discussion of reflection principles for conative attitudes.

[7] To make this principle more closely parallel to van Fraassen's Reflection principle, one might want to express it in terms of credences and utilities rather than beliefs and preferences. But I will argue in Section 4.2.5 that this cannot be done unless you are certain about what preferences you will have if you are rational in the future.

[8] See Harman (2009) for further discussion, though note that Harman is ultimately arguing against Preference Reflection, in part due to the problems with bootstrapping discussed below.

later have some desire which will provide a reason for you to act in a certain way, then you now have a reason to act in that way.[9]

4.2.1 Personal Identity

Despite some prima facie plausibility, Preference Reflection faces a series of devastating problems. First, it should by now be obvious that Preference Reflection will face problems with puzzle cases for personal identity over time. It seems that in these myriad puzzle cases, neither the agent nor the theorist should have to settle the metaphysical facts about identity in order to determine what the agent in the scenario ought to desire. Take Double Teletransportation. If Preference Reflection is a requirement of rationality, then whether Pre must defer to the preferences she expects Lefty (or Righty) to have depends on whether she is identical to one, to the other, to both, or to neither. But plausibly, we don't need to settle these identity facts in order to settle what Pre ought to prefer. Insofar as it is permissible to have a distinctive sort of self-concern, then Pre has reason to care about the preferences of Lefty and Righty regardless of whether she bears the relation of personal identity to either or both of them. So plausibly, Pre's reasons to promote the desires she expects Lefty or Righty to have should fall out of more general principles which, unlike Preference Reflection, will make no reference to the relation of personal identity over time. Of course, we could get around this particular problem by modifying Preference Reflection so that it says that you ought to defer to the preferences that you believe future psychological continuants of you will have. I have previously argued that this strategy of replacing personal identity with psychological continuity or R-relatedness is problematic and unmotivated in the context of epistemological principles like Conditionalization and van Fraassen's Reflection, but it actually seems fairly natural in the context of principles for preferences. So I would not want to rest my case against Preference Reflection (or related principles) on considerations of personal identity puzzle cases alone. But fortunately for my purposes, the case against Preference Reflection is overdetermined, for it faces a number of other devastating problems.[10]

[9] See Bratman (2014) for a defense of a relation practical reflection principle, which he calls *Standpoint Reflection*. It is a reflection principle not for a single kind of practical attitude like preferences or intentions, but for the agent's practical standpoint as a whole, which incorporates many elements of an agent's psychology. I suspect that Bratman's principle will face some of the problems I raise below, but not all of them (for instance, as Bratman (p.c.) pointed out, it may not face a version of the "No Formal Analogue" problem I raise below). Note that Bratman does not defend the Standpoint Reflection principle as a norm of rationality, but rather as a "structural principle that is part of the metaphysics of planning agency" which "says how the contours of a planning agent's present standpoint are potentially shaped by certain expectations of future attitudes" (15).

[10] Thanks to Dan Greco for helpful discussion of this issue.

4.2.2 Bootstrapping

Preference Reflection yields unattractive consequences in cases where you believe that what you will later prefer depends on what you will do now. Suppose that you are deliberating about whether to travel to Argentina or to Brazil. You believe that whatever you do, you will be very glad that you did that thing and not the other. If you go to Argentina, then while you are hiking in Patagonia and reminiscing about the fantastic culture of Buenos Aires, you will be very glad you went to Argentina rather than Brazil. You will have a preference for having gone to Argentina over having gone to Brazil. And you believe that if you go to Brazil, then while you're sitting in the Sambadrome at Carnaval and remembering the brilliant wildlife of the Amazon, you will be glad you chose Brazil over Argentina.

Preference Reflection says that which trip you ought to prefer now depends on which trip you believe you will take. If you find yourself leaning toward Argentina and hence believe that you will choose Argentina over Brazil, then you believe that you will later prefer Argentina over Brazil, and so by Preference Reflection you ought now to prefer Argentina over Brazil. And if you think you will wind up going to Brazil, then you ought to believe that you will later prefer Brazil over Argentina, and hence by Preference Reflection you ought now to prefer Brazil over Argentina. It seems, then, that Preference Reflection endorses the following sort of bootstrapping reasoning: "I believe I will go to Argentina (Brazil), and so I ought to go to Argentina (Brazil)." But intuitively, this kind of bootstrapping is irrational, for you know that whichever trip you take, you will find yourself happy and glad you did it.[11]

Worse, suppose that you believe that whatever you do, you will wish you did the other thing. If you go to Argentina, then you will think fondly of the amazing time you could have had in Brazil and wish you had gone there instead. And if you go to Brazil, you will wish you had gone to Argentina. Then, if you believe you will go to Argentina, then you ought to believe you will later prefer Brazil over Argentina,

[11] Harman (2009) discusses a case with just this structure. A mother gives birth to a child who is deaf. The doctors tell her that it is possible to cure the child's deafness with a cochlear implant, and the mother must decide whether to cure the child's deafness or not. She knows that if she cures the baby's deafness, she will be glad she did it. She will later rationally prefer having cured the baby to having refused the treatment. But she also knows that if she refuses the treatment, she will be glad she did so. The child will very likely grow up to be a happy deaf adult and will have numerous close relationships in the deaf community. The child might even be glad to be deaf (as evidence suggests many deaf adults are). And, of course, she will love the child as he or she is. Now, Preference Reflection entails that whether she should now prefer that she cure the child's deafness or not depends on what she believes she will do. If she starts out believing that she will cure the child's deafness, she should prefer that she do so, whereas if she starts out believing that she will refuse the treatment, she should prefer that she do that. Again, this seems like a bad result.

so by Preference Reflection you ought now prefer Brazil over Argentina. And similarly, if you believe you will go to Brazil, then you ought now prefer Argentina over Brazil. So Preference Reflection endorses reasoning as follows: "I believe I will go to Argentina (Brazil), and so I ought to go to Brazil (Argentina)." Once again, this sort of reasoning seems crazy.[12]

One might attempt to defend Preference Reflection by saying that in these cases, the anticipated preferences are irrational. This may be right, but it's worth noting that it does rely on the thought that it's irrational to care non-instrumentally about things other than pleasure and pain. If not, then when you have a wonderful time in Argentina, it may be rational to value *those very experiences*, which you wouldn't have had, had you gone to Brazil (though of course you would have had different memorable experiences). I value the experiences that I have had despite recognizing that I could have had other equally pleasant experiences. I am glad that I have lived *my* life, rather than any number of other lives which contain the same amounts of pleasure and pain. If this is right, then Preference Reflection cannot avoid highly counterintuitive implications in cases where you believe that what you will later desire depends on what you will now do.

4.2.3 Time-Bias

As we saw in the previous chapter, being biased toward the future results in shifts in your preferences, and so knowing that you are time-biased is incompatible with Preference Reflection. To see this, consider the case discussed in the previous chapter, in which you are ignorant of which of two courses of surgery you will have to undergo:

The Early Course
You will have 4 hours of painful surgery on Tuesday and 1 hour of painful surgery on Thursday.

The Late Course
You will have no surgery on Tuesday and 3 hours of painful surgery on Thursday.

On Monday, you prefer the Late Course over the Early Course, but you know that, being biased toward the future, you will on Wednesday prefer the Early Course over the Late Course. Hence, you violate Preference Reflection.

Similarly, if you are biased toward the near, your preferences will shift (unless, as explained in the previous chapter, you discount exponentially). And so, if you are biased toward the near and recognize this fact about yourself, you expect to

[12] See Hare and Hedden (forthcoming) for further discussion of these sorts of cases.

have preferences in the future which you do not have now. Hence, you violate Preference Reflection.

Insofar as we think that some form of non-exponential time-bias is at least rationally permissible, we must reject Preference Reflection.

4.2.4 Arbitrary Asymmetries

Preference Reflection, like Reflection for credences, is asymmetric. It privileges the preferences of your future selves over the preferences of other people, and it privileges your future preferences over your past preferences. Thus, as with van Fraasen's Reflection, Preference Reflection is arbitrarily asymmetric with respect to both the you/other distinction and the past/future distinction.

Parfit (1984, 187) makes this point while considering an argument against the rationality of time-bias. He imagines the following accusation against one who is biased toward the near:

> You do not *now* regret your bias towards the near. But you *will*. When you pay the price–when you suffer the pain that you postponed at the cost of making it worse–you will wish that you did not care more about your nearer future. You will regret that you have this bias. It is irrational to do what you know that you will regret.

This objection is grounded in something like Preference Reflection, for it suggests that if you know you will prefer one thing to another (so that if you did the other, you would regret it), then you ought now prefer the one to the other. But Parfit rejects this argument. He writes of the agent who is biased toward the near that:

> he may regret that in the past he had his bias towards the near. But this does not show that he must regret having this bias now. A similar claim applies to those who are self-interested. When a self-interested man pays the price imposed on him by the self-interested acts of others, he regrets the fact that these other people are self-interested. He regrets their bias in their own favour. But this does not lead him to regret this bias in himself.

Parfit is in effect defending time-bias by arguing that Preference Reflection is arbitrary, since it applies only with respect to desires that you anticipate *you* will later have. If we reject an analogous principle which applies in the in*ter*personal case, telling you to have the desires or preferences that you believe *others* have, we should likewise reject Preference Reflection.

Preference Reflection is also aymmetric in virtue of applying only with respect to desires you expect you will *later* have. If you ought to adopt as your own the desires you expect to have in the future, ought you also to adopt the desires you believe you had in the past? It seems doubtful. Parfit (1984, 157) imagines that in his youth he wanted more than anything to be a poet. And it was not as if he wanted to be a poet only if he would still want this in the future, when the time for

writing poetry arrived. But he no longer desires to be a poet. It is not that his value judgments have changed; he has not decided that poetry is frivolous or pretentious and so he does not regard his youthful desire as irrational in any way. Parfit thinks that in such a case he has no reason to pursue poetry, not even a reason which is overridden by other considerations. The mere fact that he believes he once desired to become a poet gives him no reason now to have any desire to be a poet. If this is right, it casts doubt on Preference Reflection. If you have no reason to defer to the desires you believe you once had, why should it be that you ought to defer to the desires you believe you will have in the future? So as with van Fraassen's Reflection, Preference Reflection is doubly asymmetric, and problematic for that reason.

4.2.5 No Fine-Grained Analog

In the case of reflection principles for doxastic attitudes, there was an analog for fine-grained attitudes (credences) of a more easily-statable principle for coarse-grained attitudes (binary beliefs). The coarse-grained principle stated that you ought to be such that if you believe you will later believe H, then you now believe H, and the fine-grained principle stated that you ought to be such that $P_0(H \mid P_1(H) = n) = n$. The principle of Preference Reflection, put in terms of coarse-grained attitudes, states that you ought to be such that if you believe you will later prefer that H, then you now prefer that H. Is there an analog of Preference Reflection which is put in terms of fine-grained attitudes like credences and utilities?

The natural formal analog would state that your utility for A, conditional on the claim that you will later have utility function U_i, should equal $U_i(A)$. Where the A_j form a partition of A, your conditional utility for A given E is defined thus:

$$U(A \mid E) = \sum_j U(A_j) \times P(A_j \mid A \wedge E).$$

Utility Reflection
It is a requirement of rationality that, unless you believe you might be irrational or have lost evidence in the future, then for all A,

$$U_{now}(A \mid U_{later} = U) = U(A)$$

Utility Reflection entails that if you do not believe that you might be irrational or have lost evidence in the future, your current utility for A should equal your expectation of your future utility for A:

$$U_{now}(H) = \sum_i U_i(H) \times P(U_{later} = U_i)$$

Unfortunately, Utility Reflection is unworkable unless it is never rationally permissible to change your ultimate preferences; that is, unless Utility Conditionalization is true. For unless this is the case, Utility Reflection will fall prey to a version of the

problem of interpersonal comparisons of utility.[13] I will show how this problem arises in the present context and argue that, while other instances of the problem may be more easily dealt with, in this case it is insoluble. This is because it assumes that we can talk about *the* utility you will later assign to a proposition, or about *the* utility function you will later have. But your preferences at a time do not determine a unique utility function. Rather, they determine a utility function which is unique at most up to positive affine transformation. Utility function U represents your preferences if and only if $U' = aU + b$ (where $a > 0$) represents your preferences as well. More intuitively, the zero point and the scale of a utility function are arbitrary; you can move the zero point and change the scale without changing which preferences the utility function represents.

Changing the zero point and scale of a utility function does not change anything *when only one utility function is in play*. If you rank options by expected utility relative to a credence function P and a utility function U, the ranking will be exactly the same if you calculate expected utilities relative to credence function P and utility function $U' = aU + b$ (where $a > 0$).

But changing the zero point and scale of a utility function *does* matter when multiple utility functions are in play, as when you are aggregating utilities. To see this in the context of Utility Reflection, suppose that you are 0.5 confident that you will later have preferences which are representable by utility function U_1 and 0.5 confident that you will later have preferences which are representable by utility function U_2. Utility Reflection says that your current utility function should be $U_a = 0.5 \times U_1 + 0.5 \times U_2$. But if we change the scale of U_1, say by multiplying it by 50, to arrive at a different utility function U_1' which represents the same preferences, Utility Reflection now gives a very different answer about what your current utility function ought to be. And the utility function it now tells you to have will *not* be a positive affine transformation of the utility function we got using U_1. We can see this as follows:

Let $U_a = 0.5 \times U_1 + 0.5 \times U_2$.
Then, $U_a \times 2 = U_1 + U_2$ is a positive affine transformation of U_a.
Let $U_b = 0.5 \times U_1' + 0.5 \times U_2 = 0.5 \times (50 \times U_1) + 0.5 \times U_2$.
Then, $U_b \times 2 = 50 \times U_1 + U_2$ is a positive affine transformation of U_b.
But $50 \times U_1 + U_2$ is not a positive affine transformation of $U_1 + U_2$.

So plugging U_1' into Utility Reflection results in a recommended utility function which is not a positive affine transformation of the utility function that Utility

[13] Arntzenius (2008) is also aware of this problem in his discussion of reflection principles for desire-like attitudes, but he does not address it in detail.

Reflection recommends if we plug in U_1 instead. Therefore, plugging U_1' rather than U_1 into Utility Reflection results in different recommendations about what preferences you ought to have now! But the choice of U_1 over U_1', or vice versa is completely arbitrary; there are no grounds for picking out *one* of the utility functions that represents the preferences you might later have over any other of the infinitely many utility functions which represents those same preferences. For this reason, Utility Reflection gives inconsistent recommendations about what your current preferences ought to be.

Of course, this problem will not arise if you must have the same set of ultimate preferences in the future in any case in which you are rational. This would be the case if rational agents never change their ultimate preferences, as Utility Conditionalization entails. For then you know that if you are rational in the future, you will have the same ultimate preferences that you have now. Then it is natural to stipulate that if you have the same ultimate preferences in two possible cases, your preferences in those two cases should be represented by utility functions which agree on the values they assign to maximally specific possibilities. That is, if you have the same ultimate preferences in case 1 and case 2, then it is natural to stipulate that once we arbitrarily fix on one particular utility function U_1 out of the set of utility functions that represent your case 1 preferences, we must choose a utility function U_2 out of the set of utility functions representing your case 2 preferences, where for all possible worlds w_i, $U_1(w_i) = U_2(w_i)$. Same ultimate preferences mean same utilities for possible worlds. It is worth emphasizing that this is a stipulation, and not something that follows from the decision-theoretic framework itself. Still, if we make this stipulation, then Utility Reflection is in good shape provided that Utility Conditionalization (or some more general principle that entails it) is true. Indeed, given Utility Conditionalization and the stipulation that same ultimate preferences mean same utilities for possible worlds, Utility Reflection will follow trivially from Modified (Belief) Reflection.

But if Utility Conditionalization is false, and you can have different ultimate preferences in the future without being irrational, then Utility Reflection faces the aforementioned problem. This problem is a version of the problem of interpersonal comparisons of utility. If utility functions are thought of as representations of preferences, there is no sense to be made of whether my utility for H is greater than yours. This, of course, is problematic for versions of consequentialism which say that you ought to maximize total well-being in the world, if well-being is understood as preference satisfaction, for different choices of utility functions to represent people's preferences will result in different conclusions about what you ought to do.

The remainder of this section looks at possible solutions to the problem of interpersonal comparisons of utility and concludes that there is no solution to that problem, if utilities are understood as representations of preferences (which is how utilities must be understood in order for Utility Reflection to be a fine-grained analog of Preference Reflection). As the discussion is somewhat long and involved, impatient readers are welcome to skip the remainder of this section.

The problem of interpersonal comparisons of utility may be less problematic if we think of utility functions as representing something like levels of happiness or as representing betterness relations. If utilities represent levels of happiness, then intrapersonal comparisons of utility (comparisons of utility between the same person at different times) may be easier to make than interpersonal comparisons, since you have better access (in particular, through memory) to facts about how happy something made you in the past than to facts about how happy that thing makes someone else. But even if interpersonal comparisons of levels of happiness are empirically more difficult, or even impossible, to make, this does not mean to entail that they are meaningless. It may be difficult or impossible to determine whether the pleasure or happiness I get from eating chocolate ice-cream is more intense than the pleasure or happiness you get from eating chocolate ice-cream, but this does not mean that there is no fact of the matter about whose pleasure or happiness is more intense. To say this is just to reject verificationism, as most philosophers do nowadays. Similarly, if utilities represent facts about betterness, then there is no reason to doubt that interpersonal comparisons of utility are meaningless. There may be facts about whether one state of affairs is better for one person than another state of affairs is for a different person, even if such facts are difficult to determine. In sum, if utility functions represent phenomenological states like levels of happiness, or alternatively if they represent facts about betterness, then both the problem of interpersonal comparisons of utility and the problem of intrapersonal comparisons of utility can be solved (though the latter may be empirically easier).

But crucially, because Utility Reflection is meant to be a fine-grained analog of Preference Reflection, the utility functions involved in Utility Reflection must be interpreted as representing possible future preferences, rather than as representing levels of happiness or goodness. Of course, if one thought that the strength of a preference was something phenomenological (e.g. the warm feelings one gets when one contemplates the prospect of the preference being satisfied), then one could solve the problem of interpersonal comparisons of utility in exactly the same way that it can be solved if we interpret utilities as representing levels of happiness or pleasure. There would be some fact of the matter, having to do with intensities of warm fuzzy feelings, about whether my desire for chocolate

ice-cream is stronger than your desire for chocolate ice-cream, or stronger than my past desire for chocolate ice-cream, even if it is difficult or impossible to verify this. But it is doubtful whether strengths of preferences can be interpreted as grounded in phenomenology. Compare Ramsey's (1931, 170) discussion of degrees of belief:

> It could well be held that the difference between believing and not believing lies in the presence or absence of introspectible feelings. But when we seek to know what is the difference between believing more firmly and believing less firmly, we can no longer regard it as consisting in having more or less of certain observable feelings; at least I personally cannot recognize any such feelings. The difference seems to me to lie in how far we should act on these beliefs: this may depend on the degree of some feeling or feelings, but I do not know exactly what feelings and I do not see that it is indispensable that we should know.

Similarly, I find it doubtful that my own preferences or strengths of desires involve introspectable phenomenological features. For instance, I very strongly desire not to die tomorrow, but I do not get any warm fuzzy feeling when I contemplate the prospect of surviving the week. Indeed, if anything, I get more of a warm fuzzy feeling when I think about chocolate ice-cream than when I think about surviving the week, even though I have a much stronger desire for survival than for ice-cream. Nor is it the case that strengths of preferences or desires always track the extent to which you would feel disappointment or suffering. I very strongly desire to survive the week, but as Epicurus observed, I won't feel anything if that desire is frustrated.

If utilities are interpreted as representing strengths of preferences (as opposed to levels of pleasure or happiness), and preferences are understood dispositionally rather than phenomenologically, then I think that the problem of inter- and intra-personal comparisons of utility may be in principle insoluble. I think that once we grant that preferences are not interpreted phenomenologically, we should think that while preferences and ordinal preferences are real, cardinal utilities are merely a device to represent your preferences over not just maximally specific possibilities, but over gambles as well. Suppose you prefer A to B to C. The fact that the difference between your utility for A and your utility for B is equal to the difference between your utility for B and your utility for C just amounts to something like the fact that you would be indifferent between getting B for certain and a gamble with a 50 per cent chance of yielding A and a 50 per cent chance of yielding C. On this view, desires, ordinal preferences, and even ratios of utility differences may be real, but the choice of the zero point and scale of your utility function is arbitrary and not determined by anything in your psychology. Note that on this interpretation of utility scales, which I take to be a fairly standard one in decision theory, both the problem of interpersonal comparisons of utility and the problem of intrapersonal but intertemporal comparisons are in principle

insoluble, since there is no fact of the matter what any given time-slice's zero point and scale are.[14]

The reason why there is no corresponding problem of inter- or intra-personal comparison of degrees of belief is that there are clear upper and lower bounds—certainty of truth and certainty of falsehood—to how confident any person can be in a proposition. The choice of 1 and 0 as numbers to represent these extremal degrees of belief is largely conventional.[15] And so whenever person A is certain of P and person B is certain of Q, we just set A's degree of belief for P and B's degree of belief in Q at 1. But while it is part of the concept of belief that there are upper and lower limits on how confident you can be in a proposition, it is not part of the concept of desire that there are upper and lower limits on how strongly you can desire something. If there were, then we could take anything that any person "maximally desires" and assign it some arbitrary utility (100, say) as the conventional upper bound, take anything that any person "maximally disprefers" and assign it some arbitrary utility (-100, say) as the conventional lower bound, and fill in everything else accordingly.[16] But desire isn't like that. There is no motivation (apart from wanting to solve the problem of inter- and intra-personal comparisons of utility) for thinking that there are upper and lower bounds for how strongly a person can desire something, nor for thinking that such upper and lower bounds should be the same for everyone. To take just one example, it seems that I could prefer more days in heaven to fewer, such that I have no diminishing

[14] This is admittedly a radical conclusion. Is there really no sense in which my preference for living for another week is stronger than your preference for having a cup of coffee? I am tempted to think any truth in this comparative claim derives not from some fact about your mental state and mine, but rather from a normative judgment that in ethical decision-making, we should give your preference for having a cup of coffee less weight than my preference for living another week. So perhaps I can account for our inclination to make comparative claims about strengths of preferences by interpreting them as proposals for how to weight our competing interests in determining what is morally required of us.

[15] Of course, there are good mathematical reasons for choosing 1 and 0 as the upper and lower bounds. In particular, the axiom of finite additivity and the definition of probabilistic independence seem to depend on using the $[0,1]$ scale. Standardly, your degree of belief in the disjunction of two mutually exclusive propositions should be the sum of your degrees of belief in each of the propositions, and your degree of belief in the conjunction of two independent propositions should be the product of your degrees of belief in each proposition. But if the scale were $[-1,1]$, for instance, then your degrees of belief in H and $\neg H$ could each be 0 (the midpoint of the scale), but by finite additivity, your degree of belief in the necessary disjunction $H \vee \neg H$ would also be 0. And if the scale were, say, $[0,2]$, you could have two independent propositions P and Q, each with degree of belief 1 (the midpoint of the scale), but by the standard definition of independence, your degree of belief of the conjunction $P \wedge Q$ should still be 1. So if we were to choose a different scale to represent probabilities or degrees of belief, we would have to replace finite additivity and the standard definition of independent with alternative principles which would almost certainly be far less elegant. So, while the $[0,1]$ scale is conventional, there are good reasons for choosing this convention over alternatives.

[16] This would be a version of the so-called "zero-one rule." See Hausman (1995) for discussion.

marginal utility for days spent in heaven. I prefer 1,001 days to 1,000 days just as strongly as I prefer 101 days to 100 days. If I can have preferences like that (and it certainly seems that I could), then my utility function must be unbounded.[17]

Another interesting but ultimately mistaken proposal for solving the problem of inter- and intra-personal comparisons of utility (again, under a preference interpretation of utility) is due to Harsanyi (1955, 1977). Harsanyi introduces the notion of an *extended alternative*, which is a pair consisting of a state of affairs and a set of personal characteristics. So, an extended alternative could be something like *being a runner with a personal dislike of lactic acid* (bad) or *being a philosopher while possessing great clarity of thought and an appreciation of solitude* (good). Harsanyi thinks that everyone has the same extended preferences and that this allows us to make interpersonal comparisons of utility, even under a preference interpretation of utilities, by stipulating that everyone's utility function assigns the same utilities to particular extended alternatives, and filling everything else out accordingly.

But there is every reason to think that people do not in fact have the same extended preferences.[18] Broome (1993, 63) makes this point clearly:

> I myself prefer to live the life of an academic, with my own academic characteristics, even in the conditions allotted to academics in contemporary Britain, to being a financial adviser living in the conditions allotted to financial advisers. I would expect a financial adviser,

[17] One might make the following proposal for fixing the zero points and scales of utility functions, suggested by Sepielli (2009) in a different context. Start with the fact that ratios of utility differences are independent of the choice of zero point and scale. Then, the idea is to find three propositions A, B, and C such that the two people *a* and *b* (they could be two time-slices of the same person) both prefer A to B to C and are the same with respect to the ratio of the difference between the utility of A and the utility of B, and the difference between the utility of B and the utility of C (that is, the value of $[U(A) - U(B)]/[U(B) - U(C)]$ is the same for each person). Then, the proposal is to arbitrarily pick a number x as the utility each assigns to A and fill everything else out accordingly. That is, we find three propositions such that the two agents have the same ordinal preferences among them and agree on the ratios of the utility differences between them, and then assign to *a* and *b* utility functions which assign the same numbers to these three propositions, so that $U_a(A) = U_b(A)$, $U_a(B) = U_b(B)$, and $U_a(C) = U_b(C)$.

But unfortunately, this solution is not only unmotivated, but also inconsistent. Suppose that agents *a* and *b* agree not only on the ordinal ranking and ratio of utility differences with respect to A, B, and C, but also on the ordinal ranking and ratio of utility differences with respect to D, E, and F. And suppose that for *a*, the utility difference between A and B is very large relative to that between D and E, while for *b*, the utility difference between D and E is very large relative to that between A and B. Then, if we run Sepielli's procedure using A, B, and C as our privileged triplet of propositions (so that $U_a(A) = U_b(A)$, $U_a(B) = U_b(B)$, and $U_a(C) = U_b(C)$), we will get very different results than if we run it using D, E, and F as the privileged triplet (so that $U_a(D) = U_b(D)$, $U_a(E) = U_b(E)$, and $U_a(F) = U_b(F)$).

[18] Note that denying that people have the same preferences over extended alternatives does not entail that people differ with respect to how good a given extended alternative would be for them, unless we are committed to a preference-satisfaction theory of well-being.

with her different values, to have the opposite preference. So her extended preferences are different from mine. The reason I have mine is that an academic has some slight chance of making a worthwhile contribution to knowledge. I recognize that, if I were a financial adviser, with all the characteristics of a financial adviser, I would not then value knowledge as I do now. Nevertheless, I do value knowledge, and that is why I prefer to be an academic.

Harsanyi seems to defend his claim that everyone has the same extended preferences by appeal to the fact that "different individuals' behavior and preferences are at least governed by the *same basic psychological laws*" (Harsanyi (1977, 58)). Let Φ be a variable which takes as possible values all of the relevant causal factors which result in agents having the particular preferences that they do, things like upbringing, genes, age, physical and psychological abilities, and the like. Now, let $U_i(-) = V_i(-;\Phi)$ be the utility function (or rather, one of the family of utility functions) that individual i would have had if causal factors Φ had obtained. And as Harsanyi notes, if the same basic psychological laws govern everyone's preferences and behavior (so that if two individuals were subject to exactly the same causal factors, they would have the same preferences and behavior), then all differences between their utility functions $U_i(-) = V_i(-;\Phi_i)$ and $U_j(-) = V_j(-;\Phi_j)$ are due to the differences between the causal factors Φ_i and Φ_j, "and not to differences between the mathematical form of the two functions V_i and V_j." Put briefly: everyone has the same function V_i, since everyone's preferences are governed by the same psychological laws.

But it would be illegitimate to leap, as Harsanyi seems to do, from the true claim that everyone has (or is subject to) the same function V_i from causes to preferences to the claim that everyone has the same extended preferences. As Broome (1993) argues, this leap rests on a confusion between objects of preference and causes of preference. In the function $V_i(-;-)$, it is important to emphasize that what comes after the semicolon is a slot for causes of preferences, not objects of preferences (though of course one can also have preferences over factors that are causally relevant to one's preferences). The fact that $V_i(A;\Phi_j) > V_i(B;\Phi_k)$ does not mean that everyone prefers A's obtaining when causal factors Φ_j obtain over B's obtaining when causal factors Φ_k obtain. For instance, let Φ refer to the causal condition of having been turned into a zombie. Let x refer to feasting on the flesh of the living and y refer to refraining from doing so. Since zombies desire nothing more than to feast on the flesh of the living, $V(x;\Phi) > V(y;\Phi)$. Being subject to the causal condition of having been zombified would result in one's preferring feasting on flesh to not doing so. But the non-zombies among us generally prefer the extended alternative $< y, \Phi >$ to the extended alternative $< x, \Phi >$. As a non-zombie, I prefer that, were I to become a zombie, I refrain from feasting on the living, even though, were I to become a zombie, I would prefer to give in

to feasting on the living. So Harsanyi's function $V_i(-;-)$ may be the same for everyone, but this is irrelevant, since it does not represent anyone's (extended) preferences; instead it represents the way in which their differing preferences are governed by causal factors.

In sum, Harsanyi's appeal to extended preferences fails to solve the problem of intertheoretic comparisons of utility (again on a preference interpretation of utility). We do not all have the same extended preferences, even though we are the same with respect to how causal factors determine our extended preferences.

No doubt there are other attempts to solve the problem, but an exhaustive survey would go far beyond the scope of this book. Nonetheless, having looked at the most prominent proposed solutions and found them wanting, I conclude that it is insoluble on a preference interpretation of utility functions which is relevant here. Interpreted this way, it is natural to think that while desires and ordinal preferences may be "psychologically real," the zero points and scales of utility functions have no psychological reality; there is no fact of the matter whether the agent's utility function is *really U* as opposed to $aU + b$ $(a > 0)$. And ultimately, it is this fact that means that the problem of inter- or intra-personal comparisons of utility cannot be solved.

If I am right, then Utility Reflection is unworkable unless rational agents must always have the same ultimate preferences. There is no analog of Preference Reflection that deals with fine-grained attitudes like credences and utilities instead of beliefs and preferences.[19]

Where does this leave us? In Chapter 9, I will suggest that Utility Reflection can be salvaged if we adopt a strong uniqueness thesis for preferences, according to which everyone is rationally required to have the same ultimate preferences, preferring one thing to another just in case the one is better than the other. As

[19] What about a formalization of Preference Reflection that replaces talk of beliefs with talk of credences, but does not likewise replace talk of preferences with talk of utilities? There are two potential problems. First, the spirit behind Preference Reflection is such that whether you ought now to prefer A to B or B to A depends not only on your credence that you'll prefer A to B and your credence that you'll prefer B to A, but also on *how strongly* you think you'll later prefer A to B or B to A. But talk of strengths of preferences, I have suggested, makes sense only in the context of a utility function. Second, suppose we ignore strengths of preferences and say only that if your credence that you'll later prefer A to B is above some threshold n, then you ought now prefer A to B. This sort of proposal threatens to yield intransitive preferences. For instance, suppose that you have credence 1/3 that you'll prefer A to B to C, 1/3 that you'll prefer B to C to A, and 1/3 that you'll prefer C to A to B. Then, you have credence 2/3 that you'll prefer A to B, 2/3 that you'll prefer B to C, and 2/3 that you'll prefer C to A, and so if our threshold n is below 2/3, you will be required to have intransitive preferences and prefer A to B, B to C, and C to A. Nor can we avoid the problem altogether by raising the threshold n, for more complicated cases involving preferences over more propositions will still yield intransitivity. This is an instance of the problem of judgment aggregation. See List and Pettit (2011) for discussion.

I noted above, if you are certain that all possibilities in which your future self is rational involve having the same ultimate preferences, then we can solve the problem of interpersonal comparisons of utility by requiring that each of your possible future preference orderings be represented by a utility function, all of which assign the same utility to maximally specific possibilities. Interestingly, such a uniqueness thesis for preferences would yield Utility Conditionalization and Utility Reflection as instances of more general, impersonal, and synchronic which require that everyone, at all times, has the same ultimate preferences, namely the uniquely rational ones.

Of course, this is a very strong claim about the rationality of preferences. I will give my best attempt at motivating it in Chapter 8 and will tentatively endorse it. If you are not convinced, then Utility Conditionalization and Utility Reflection should be rejected outright, but if you are sold, then they can be subsumed under more general time-slice-centric principles.