Interpersonal Comparisons with Preferences and Desires

Jacob Barrett

1. Introduction

Most moral and political theories are concerned, at least in part, with welfare. They deem that the effect that actions and policies have on individuals' welfare is relevant to their evaluation, even if other things—such as rights and desert—matter, too. At a minimum, all such theories require us to make *intra*personal comparisons of welfare, like, "individual *i* is better off in outcome *x* than outcome *y*." These comparisons allow us to apply the Pareto criterion, according to which one outcome is superior to another if the first is better for some individuals and worse for none. But the Pareto criterion is notoriously unhelpful in most real world moral and political contexts, since nearly every action or policy leaves at least one individual worse off than some feasible alternative. So most moral and political theories extend their concern for welfare beyond Pareto by considering not just who gains and loses, but also the relative size of these gains and losses, and how well off individuals are in relation to one another. These *inter*personal comparisons of welfare are necessary if we are to evaluate actions and policies according to efficiency criteria like, "which confers a greater balance of welfare gains over losses?," distributive criteria like "which promotes a more equal distribution of welfare?," and even non-welfarist criteria like "which provides greater gains to those who deserve it more?" Though few theories endorse each of these criteria—utilitarianism, for example, recognizes only the first—most accept at least one. Thus, most moral and political theories require us to make interpersonal comparisons of welfare.

According to one popular theory of welfare, an individual is well off to the extent that her (suitably idealized and restricted) preferences or desires are satisfied.[1] This theory is especially interesting in this context, because while there are clear epistemic difficulties involved in making interpersonal comparisons of pleasure and of various objective goods—as other theories of welfare require—interpersonal comparisons of preference or desire satisfaction are widely thought to raise an additional conceptual problem, and purported solutions to this problem to lead to a hopeless subjectivism about these comparisons.[2] This has led many to accept that preference or desire satisfaction theories of welfare do not allow us to make interpersonal comparisons of welfare, at least in the objective way our moral and political theories require. The challenge for preference or desire satisfaction theories of

[1] By "suitably idealized and restricted," I mean, for example, that the preferences or desires might be idealized to those one would have in the face of complete information, and restricted to exclude concern for others' desire or preference satisfaction. Such idealizations and restrictions make little difference to the problem at hand, so I ignore them in what

[2] Although he later changed his view (adopting a version of the "extended preference" solution that I discuss in section 3), Arrow, for example, influentially denied that such comparisons have any "meaning" (1963: 9). This position had already been in vogue among economists since the 1930s, in no small part due to the influence of Robbins (1932). But Robbins himself was more worried about the objectivity of interpersonal comparisons than their meaningfulness. As Roberts explains: "For Robbins… the problem is that there may be many different views or subjective comparisons but there is no 'scientific' way of choosing between them. Thus, the fundamental question is how to proceed when there are many different views and, in particular, how can a set of *objective* interpersonal comparisons be found in such a situation?" (1997: 73).

welfare is clear: if they cannot support objective interpersonal comparisons of welfare, then they are unsuitable for most moral and political theories. We would seem better off rejecting such theories of welfare than giving up on the possibility of objective interpersonal comparisons.

In this paper, I argue that the key to meeting this challenge lies in distinguishing preferences from desires, and preference satisfaction from desire satisfaction theories of welfare. More specifically, I defend the following three conclusions. First (Section 2), interpersonal comparisons of preference satisfaction do raise a serious conceptual problem, but this same problem does not arise for interpersonal comparisons of desire satisfaction. Second (Section 3), none of the existing solutions to this problem are satisfactory, since none explain how we can make interpersonal comparisons of preference satisfaction objectively. Third (Section 4), we can at least make a limited range of objective interpersonal comparisons of desire satisfaction, and there are reasons to be optimistic about the possibility of making a wider range of such comparisons. But whether this optimism ultimately proves warranted turns not on any formal issue, but on our substantive theory of desire along with various further empirical considerations and questions in the philosophy of mind.

2. Preferences and desires

To begin, preference and desire satisfaction must be distinguished from the feelings of satisfaction that typically accompany the belief that one's preferences or desires have become better satisfied. Instead, one's preferences are better satisfied when the world is as one prefers it to be, and one's desires are better satisfied when the world is as one more strongly desires it to be—regardless of whether one believes that this is the case. In this respect,

preference and desire satisfaction are similar, and preference and desire satisfaction theories of welfare are similarly motivated. Both allow, first, that a person's welfare depends not on her achievement of certain objectively valuable goods, but on what she herself cares about (as determined by her preferences or desires). And both allow, second, that an individual may care more than just about her experiential states (such as pleasure and pain), and so may be benefited or harmed by events outside of her awareness. Thus, both are allied against objective list theories of welfare on the one hand, and hedonism on the other. Together, they constitute a third major approach to understanding human welfare.

Given their similarity, it is perhaps unsurprising that few have been careful to distinguish preference and desire satisfaction theories of welfare. But distinguish them we must, and we may do so by noting that preferences and desires have different formal structures. The preference relation holds between a person *i* and two outcomes *x* and *y*: *i* prefers *x* to *y*, or xP_iy.[3] The desire relation holds between a person *i*, an outcome *x*, and a strength *r*: *i* desires *x* with strength *r*, or $D_i(x)=r$.[4] So while *i*'s preferences are "better satisfied" in *x* than *y* just in case *i* prefers *x* to *y*, or xP_iy, *i*'s desires are "better satisfied" in *x* than *y* just in case the strength of *i*'s desire for *x* is greater than the strength of *i*'s desire for *y*, or $D_i(x)>D_i(y)$. This formal difference between preferences and desires may seem slight. And

[3] For now, I assume without argument that preferences and desires range over outcomes (complete descriptions of states of affairs) rather than features of outcomes (partial descriptions of states of affairs). Later, I consider this assumption in some detail.

[4] I represent the desire relation as a function from individuals and outcome to strengths, rather than as, say, $D_i(x, r)$, purely for the sake of convenience. Doing so allows me to refer to the strength of *i*'s desire for *x* as "$D_i(x)$," rather than more cumbersomely as "the value of *r* in the relation $D_i(x, r)$." Nothing else turns on this.

it is, in these intrapersonal cases. But it is exactly this difference that makes interpersonal comparisons of preference satisfaction conceptually problematic in a way that interpersonal comparisons of desire satisfaction are not.

Take first the less problematic case. Suppose that there is some other person, *k*, and we want to know whether *i*'s desires are better satisfied in *x* than *k*'s. Then the generalization from the intrapersonal case is straightforward: *i*'s desires are better satisfied in *x* than *k*'s just in case the strength of *i*'s desire for *x* is greater than the strength of *k*'s desire for *x*, or $D_i(x)>D_k(x)$. This is straightforward because the "better satisfied" relation in both the intrapersonal and interpersonal case is a "greater than" relation that holds between two quantities of strength, and quantities may be greater or lesser than each other regardless of whose they are—there is, for example, no difficulty comparing the quantity of my wealth or height to the quantity of yours. Of course, this assumes that one person's desire strength really is measurable on the same scale as another's, and perhaps that isn't so. But the point is that, if there is a problem here, it doesn't derive from the formal structure of the desire relation. Whether we can make interpersonal comparisons of desire satisfaction turns on what desire strength is, so to respond to the skeptic about such comparisons, the way forward is clear enough: one must offer a theory on which different people's desires have strengths that are—like different people's wealth and height—measurable on the same scale. The skeptic may in turn reject this theory, but this is the ground on which the battle must be fought.

Suppose, however, that we want to know whether *i*'s preferences are better satisfied in *x* than *k*'s. Now we do run into a conceptual problem, and one that derives from the formal structure of the preference relation. For *i*'s preferences to be better satisfied in *x* than *y* is for *i* to prefer *x* to *y*: xP_iy. But what is it for *i*'s preferences to be better satisfied in *x* than

k's? There is no way to generalize from the intrapersonal case. When it comes to preferences, the "better satisfied" relation is a "preference" relation that holds between a person and two outcomes, rather than a "greater than" relation that holds between two quantities of strength. It tells us that, of two outcomes, some given person prefers one to the other. But we cannot extend this relation to hold between two people—between both an "*i*" and a "*k*." It simply doesn't have the argument places for that.

One might protest that this problem is just an artifact of my notation, since it is, after all, possible to represent preferences as "greater than" relations between quantities of utility. But the problem does not go away that easily. To explain: a utility function is a numerical representation of an individual's preferences over outcomes. It assigns to that individual a value or "utility" at each outcome, where, for example, $U_i(x)=10$ represents that *i*'s utility in *x* is 10. If one's preferences form an ordering (if they are transitive and complete), they may be represented by an ordinal utility function that assigns utilities according to a single rule: if *i* prefers *x* to *y*, then $U_i(x)>U_i(y)$.[5] So, for example, if $U_i(x)=10$ and $U_i(y)=5$, this represents that xP_iy, or, equivalently, that *i*'s preferences are "better satisfied" in *x* than in *y*. But this is all that the values "10" and "5" represent: assigning any other values such that $U_i(x)>U_i(y)$ would represent precisely the same thing.[6] So suppose xP_iy and yP_kx. Then, we might represent this by assigning utilities in one of the following ways:

(1) $U_i(x)=10$, $U_i(y)=5$; $U_k(x)=20$, $U_k(y)=100$

(2) $U_i(x)=100$, $U_i(y)=1$; $U_k(x)=20$, $U_k(y)=21$

[5] Furthermore, if *i* is indifferent between *x* and *y*, then $U_i(x)=U_i(y)$. I ignore cases of indifference in what follows.

[6] Formally: an ordinal utility function is unique up to an increasing monotonic transformation.

Here, (1) and (2) represent exactly the same information: that $x\mathrm{P}_i y$ and that $y\mathrm{P}_k x$. But in (1) $\mathrm{U}_i(x)<\mathrm{U}_k(x)$ and in (2) $\mathrm{U}_i(x)>\mathrm{U}_k(x)$. So there can't be any difference between the information that $\mathrm{U}_i(x)<\mathrm{U}_k(x)$ and $\mathrm{U}_i(x)>\mathrm{U}_k(x)$ represent: when they are placed between different individuals' utilities, the "greater than" and "lesser than" relations don't represent anything at all. Even though representing preferences with utility functions allows us to write down "$\mathrm{U}_i(x)>\mathrm{U}_k(x)$," we can't understand this to represent in any sense that *i* has more utility in *x* than *k* does, or that *i*'s preferences are better satisfied in *x* than *k*'s. So representing people's preferences with utility functions doesn't solve the problem. We still can't meaningfully place two individuals on different sides of the preference relation: doing so doesn't represent anything at all.

But perhaps this conclusion is premature. For while ordinal utility functions don't represent preference strength, cardinal utility functions arguably do. And if one's preferences have strengths, then can't we just compare them in the same way I have suggested we may be able to compare desire strengths? Unfortunately, we cannot. A cardinal utility function represents preference strength in the sense that it assigns utilities on an interval scale that carries information about the ratio of the differences between its values.[7] The basic idea behind this representation is that, so long as an individual's preferences meet certain further axioms, it is possible to consider her preferences over not just outcomes but prospects (probability distributions of outcomes) and then understand the relative strength of her preferences over outcomes in terms of the different tradeoffs she would make between prospects.[8] For example, if *i* prefers *x* to *y* to *z*, and is indifferent between *y* and a 50/50

[7] Formally: a cardinal utility function is unique up to an increasing linear transformation.

[8] See von Neumann and Morgenstern (1944: ch. 3) and, for a helpful explanation aimed at non-specialists, Gauthier (1986: ch. 2).

chance of *x* and *z*, then we may assign her a cardinal utility function according to which her preference for *x* over *y* is just as strong as her preference for *y* over *z*, such that $U_i(x)$-$U_i(y)$=$U_i(y)$-$U_i(z)$. And if we do assign utilities in this way, then

(3) $U_i(x)$=45, $U_i(y)$=25, $U_i(z)$=5

represents the same information as

(4) $U_i(x)$=10, $U_i(y)$=5, $U_i(z)$=0

but not the same information as

(5) $U_i(x)$=40, $U_i(y)$=20, $U_i(z)$=10

because the ratio between $U_i(x)$–$U_i(y)$ and $U_i(y)$–$U_i(z)$ is the same—1:1—in both (3) and (4), but different—2:1—in (5).

Unlike ordinal utility functions, cardinal utility functions therefore allow us to make "unit" intrapersonal comparisons, of the form $U_i(x)$–$U_i(y)$=n($U_i(y)$–$U_i(z)$) ("the strength of *i*'s preference for *x* over y is *n* times the strength of *i*'s preference for *y* over *z*"), rather than merely the sort of "level" intrapersonal comparisons, of the form $U_i(x)$>$U_i(y)$ ("*i*'s preferences are better satisfied in *x* than *y*"), that we have been considering so far. That, and that alone, is the sense in which they represent preference strength. But note that cardinal utility functions still don't allow us to make level or unit *inter*personal comparisons.[9] For suppose that we represent *k*'s preferences like this:

(6) $U_k(x)$=30, $U_k(y)$=10, $U_k(z)$=0

Then, if we represent *i*'s preferences as in (3) $U_i(x)$>$U_k(x)$ and $U_i(y)$-$U_i(z)$=2($U_k(y)$-$U_k(z)$), but if we represent them as in (4), $U_i(x)$<$U_k(x)$ and $U_i(y)$-$U_i(z)$=1/2($U_k(y)$-$U_k(z)$)—even though

[9] For a classic explanation of these different sorts of interpersonal comparability, and the way that different moral and political theories make use of them, see Sen (1977). For a more comprehensive and up to date discussion, see Bossert and Weymark (2004).

(3) and (4) represent exactly the same information. So, again, statements like "$U_i(x)>U_k(x)$" and "$U_i(x)-U_i(y)=n(U_k(y)-U_k(z))$" don't represent anything, and we can't use cardinal utility functions to make level or unit interpersonal comparisons of preference satisfaction.

The moral of the story is that representing people's preferences with utility functions cannot change the fact that the preference relation cannot hold between two people. It remains a relation between one person and two outcomes (or prospects)—so the problem remains, too. To make a level intrapersonal comparison like "*i*'s preferences are better satisfied in *x* than *y*" is just to judge that "*i* prefers *x* to *y*," and since this relation cannot hold between two people, it does not allow us to make level interpersonal comparisons. And though it is possible to give a sensible interpretation to "preference strength" via the construction of a cardinal utility function, claims about the strength of *i*'s preferences ultimately reduce to claims about *i*'s preferences over prospects, and so—as we have just seen—cannot serve as the basis for either level or unit interpersonal comparisons either. When it comes to desires, however, there is nothing in the formal structure of the desire relation that prevents us from comparing desire strengths in the same way in both intrapersonal and interpersonal case. Desire strength may turn out not to be the sort of thing that is interpersonally comparable, but whether this is so depends not on any formal issue, but on our substantive account of desire strength. The challenge facing the proponent of interpersonal comparisons of desire satisfaction is therefore to provide an account of desire strength that allows us to make interpersonal comparisons of it, and we will turn to this challenge in section 4. But the challenge facing the proponent of interpersonal comparisons of preference satisfaction is worse: it is to explain what it could even be to make an interpersonal comparison of preference satisfaction, given that the formal structure of the preference relation doesn't seem to allow them. This is the conceptual problem of

interpersonal comparisons of preference satisfaction, and one that, we will now see, resists easy dissolution.

3. Interpersonal comparisons with preferences

There is, by now, a rather large literature on attempted solutions to this problem. The most popular derives from Harsanyi (1955, 1977a: 638-642, 1977b: ch. 4; compare Arrow, 1977; Goldman, 1995; Hare, 1981: ch. 7; Sen, 1979). On Harsanyi's view, we make interpersonal comparisons by considering not just a person's "personal preferences" over outcomes (or prospects), but instead her "extended preferences" over ordered pairs of individuals and outcomes (or prospects)—for example, *i*'s extended preference for being *i* in *x* rather than *k* in *y*, which we may represent as $x_i P_i y_k$. Then, we interpret judgments that one person's preferences are better satisfied than another's as reports of extended preferences. So when *i* judges that her preferences are better satisfied in *x* than *k*'s are in *y*, what she is really doing is reporting that she would prefer to be herself in *x* than to be *k* in *y*: that $x_i P_i y_k$. In this way, then, our conceptual problem appears to be solved. We now have the necessary argument places to place both an "*i*" and a "*k*" on different sides of the preference relation.[10]

One foundational question for the extended preference framework is whether extended preferences raise their own conceptual difficulties: does it really make sense to attribute people not only personal preferences over what outcomes they are in, but also extended preferences over which people they are in those outcomes (Adler 2014, Greaves and Lederman, 2018: 642-650)? But let us set this worry to the side, and grant that extended

[10] Harsanyi (1977b: ch. 4) furthermore suggests that we make interpersonal unit comparisons by assigning individuals cardinal extended utility functions on the basis of their preferences over probability distributions of ordered pairs of individuals and outcomes. But the problem with his approach arises even before this step, so we need not go into this here.

preferences provide us with the conceptual resources to explain what we are doing when we make interpersonal comparisons of preference satisfaction. Still, it is one thing to make such comparisons, and another to make them objectively, and the extended preferences framework does not yet explain how we can do the latter—how, that is, we can make interpersonal comparisons of preference satisfaction that are *comparer-independent*, in the sense that their truth or falsity does not depend on who is doing the comparing. The trouble arises when we consider that individuals might differ in their extended preferences. Suppose, for example, that I prefer to be me in my circumstances, while you prefer to be you in yours. On the stated account, it follows that I can truly judge that my preferences are better satisfied than yours while you can truly judge the reverse, since both are accurate reports of our extended preferences. How are we to resolve this disagreement?

Here, one option is simply to embrace a subjectivist interpretation of interpersonal comparisons, on which, in the above sort of case, "there is nothing to 'resolve'" (Sen, 1979: 189). But as Harsanyi puts it, "most people actually making such comparisons would hardly engage in this activity if they did not expect that their judgments concerning the relative magnitude of two different individuals' utility levels would have some degree of objective validity" (1977b: 57-58). We do not just take ourselves to be reporting our own preferences when we make interpersonal comparisons of welfare, and if this were all that a preference-satisfaction theory of welfare allowed, then interpersonal comparisons of welfare could hardly play the role that we require of them in our moral or political decision-making—such a view would allow, for example, that I might truly judge that an action or policy resulted in a net gain in welfare, while you truly judge that the same action or policy resulted in a net loss. This is the sense in which Harsanyi's solution to the conceptual problem of interpersonal comparisons leads to a hopeless subjectivism. It does not allow us to make the

sort of objective (comparer-independent) interpersonal comparisons that our moral and political theories require.

A second option, then, is to follow Harsanyi in arguing that, as a matter of fact, we do all share the same extended preferences (1977a: 639, 1977b: 58-59). At least since Broome's (1993) incisive critique, however, Harsanyi's argument has been widely regarded as a failure (see also Greaves and Lederman 2018: 650-652). Without going into the details, we may note that Harsanyi's position is subject to straightforward counter-example. Suppose that, in x, i is a political activist in jail and k is a corrupt politician at a fancy resort. i might personally prefer to be at the resort than in jail, but loathe k and everything k stands for so much that i would rather be herself in jail than k at the resort. Similarly, k might feel the same way about i, and prefer to be himself at the resort than k in jail. In that case, both $x_i P_i x_k$ and $x_k P_k x_i$: i and k have conflicting extended preferences, so, on the stated account, each can truly judge that her preferences are better satisfied than the other's. From this, and many other similar examples, it follows that we do not all have the same extended preferences. The problem of subjectivism remains.

In light of this difficulty, the contemporary literature has turned, following Adler, (2012: 220-22), to an attempt to rescue the extended preference framework by aggregating individuals' diverse extended preferences into a single set of objective interpersonal comparisons (for an earlier, analogous strategy, see Roberts, 1997). This is a clever move, but as Greaves and Lederman have argued, it stumbles as soon as we consider which aggregation rule to use. Not only do we run into problems relating to Arrow's theorem (Greaves and Lederman, 2017), but it moreover turns out that unless we put very strong constraints on individuals' admissible extended preferences, no aggregation rule can satisfy both (a) the seemingly foundational aggregative principle that i's preferences are better satisfied than k's

at least when *everyone* would (extendedly) prefer to be *i* than *k*, and (b) the conceptual truth that *i*'s preferences are better satisfied in *x* than *y* when she (personally) prefers to be (herself) in *x* than (herself) in *y* (Greaves and Lederman, 2018). So far, then, the prospects for this rescue mission seem bleak.

After completing their illuminating critique of the extended preference framework, Greaves and Lederman go on to suggest two more promising approaches available to the defenders of objective interpersonal comparisons (2018: 664). The first is to appeal to a normalization criterion such as the "zero-one rule," which assigns a 0 to each individual's least preferred outcome and a 1 to each individual's most preferred outcome, and then use that criterion to make interpersonal comparisons accordingly: for example, if *i* is halfway up her preference ranking in *x*, and *k* only a quarter way up in *y*, then according to the zero-one rule, $U_i(x)=0.5$ and $U_k(y)=0.25$, and we may judge that *i*'s preferences are better satisfied in *x* than *k*'s in *y* (e.g., Hausman, 1995, Schick, 1971). Though a full refutation of this approach would require a more careful examination of it than I can give it here, a well-known problem for it is that there are many different ways to normalize, and the choice of any particular criterion seems arbitrary. For example, instead of the zero-one rule, we might adopt a rule of "assigning 0 to the worst alternative and the value 1 to the *sum* of utilities from all alternatives" (Sen, 1970: 98). Or we might assign a 0 either to the point at which an individual is "indifferent between [an] outcome… and a possible world in which he never comes into existence" (Adler, 2012: 219) or "indifferent between [a] prospect and its contrary" (Bradley, 2008: 95), and then complete our normalization some other way. The viability of the normalization approach therefore turns on whether any particular normalization criterion can be defended as objectively correct—as the "one right way" to make interpersonal comparisons (Hausman, 1995: 480). So far, no such defense has met

with much applause.[11]

The final approach suggested by Greaves and Lederman is to shift our focus from preferences, extended preferences, and normalization criteria, to desires and desire strengths. Without claiming to have definitively shown that no objectivist solution to the conceptual problem of interpersonal comparisons of preference satisfaction is possible, I would now like to suggest that this last approach—given that it is able to bypass this conceptual

[11] The most influential defense is due to Hausman (1995). Yet while there appears a consensus in the literature that his argument fails, it is difficult to find a careful discussion of why. My own diagnosis runs as follows. Hausman argues that we must accept the zero-one rule because we are otherwise committed to the possibility of the following "impossible" situation: a person's preferences change, she remains at the top (or bottom) of both her "before" and "after" preference ordering, but her overall level of preference satisfaction alters (Hausman 1995: 481-482). The problem is that Hausman provides no argument that this situation is indeed impossible, and once we recognize (as Hausman 1995: 480 does) that the zero-one rule simply states that we must assign the same level of preference satisfaction to any two individuals who are both at the top or bottom of their preference orderings, we must also recognize (as Hausman seems not to) that his assumption that this situation is impossible just is the assumption that we must use the zero-one rule when making intrapersonal comparisons between "before" and "after" orderings in cases of preference change. But intrapersonal comparisons between the same person's "before" and "after" preference orderings are structurally identical—and therefore raise precisely the same problem—as interpersonal comparisons between two different individual's preference orderings, and Hausman provides no argument for the use of the zero-one rule in the former case. His argument is therefore incomplete at best, question begging at worst.

problem altogether—merits far more attention than it has so far received. It is entirely understandable that Greaves and Lederman (2018: 664), who are concerned primarily with critiquing the extended preference framework, devote only a single paragraph to the view. Less understandable is why, with very few exceptions, there is no literature on interpersonal comparisons of desire satisfaction (as distinguished from preference satisfaction) at all.[12] My own hypothesis is that many theorists have wrongly assumed that the conceptual problem of interpersonal comparisons of preference satisfaction extends to interpersonal comparisons of desire satisfaction, and that is why I haven taken such pains (in Section 2) to clarify the conceptual problem and to show that this is not so. But whatever the reason, the literature on interpersonal comparisons of desire satisfaction stands in desperate need of a jumpstart. The rest of this paper attempts to provide one.

4. Interpersonal comparisons with desires

My discussion to this point has hardly required me to say anything about what preferences are. The conceptual problem of interpersonal comparisons derives entirely from the formal structure of the preference relation, and so arises regardless of any particular account of preference—regardless, that is, of whether for *i* to prefer *x* to *y* is for *i* to be disposed to choose *x* over *y*, to derive more pleasure from *x* than from *y*, or even to more strongly desire

[12] One exception is Weirich (1984: 297-305), who defends a turn to desires from certain objections but provides a positive argument only that it would be possible to make objective interpersonal comparisons of desire satisfaction between individuals with identical desires. A second is Griffin (1991: 55-56), who attributes the strategy of appealing to desire strength to Harsanyi himself, but expresses skepticism about the possibility of "speak[ing] coherently of relative 'strength' of desire" (56). I attempt to dispel this skepticism in what follows.

x than *y*.[13] Here, it is important to realize that even if we understand preferences in this last way, there is still an important difference between preferences and desires. We may derive preferences from desires by defining the "preferred to" relation as the "more strongly desired than" relation, but we will then lose information about desire strength, while gaining none, by switching from talk of desires to talk of preferences (compare Pollock, 2006: 26-28; Weirich, 2004: 8, 19-20). For example, if *i* desires *x* with strength 5 and *y* with strength 1, then—on this interpretation of what preferences are—we can infer that *i* prefers *x* to *y*; but from the fact that *i* prefers *x* to *y* we cannot infer anything about the strength of *i*'s desires for *x* or for *y*. This is analogous to the way that the "richer than" relation loses information about absolute wealth: if *i* has $1000 in *x*, and $500 in *y*, then we can infer that *i* is richer in *x* than in *y*, but from the fact that *i* is richer in *x* than in *y* we cannot infer anything about the extent of *i*'s wealth in either outcome. This point—that desires carry more information than preferences, and lose none—is important, because it means that if we were to adopt a desire satisfaction framework we could use it to do everything that we can do in a preference satisfaction framework, and perhaps more. Most notably, we might also be able to use it to make objective interpersonal comparisons.

Still, while interpersonal comparisons of desire satisfaction avoid the conceptual problem as interpersonal comparisons of preference satisfaction, it may turn out that such comparisons are impossible all the same. Just above, I referred to desires with strengths of "5" and "1," but what do these numbers represent, and what sorts of interpersonal comparisons can we make with them? That depends on what desires are, and on what desire strength is. With preferences, we do not need to address such questions to see that the

[13] For discussion of the various ways that economists and philosophers interpret preferences, see, for example, Sen (1997).

formal structure of the preference relation renders interpersonal comparisons problematic. But since the formal structure of the desire relation raises no similar problem, these are questions to which we must now turn.

4.1 Desires, limited interpersonal comparability, and wide interpersonal comparability

In my earlier contrast of preferences and desires, I relied on one explicit and two implicit assumptions about desires. The explicit assumption was that individuals have desires directed at single outcomes with particular strengths. This is built right into my formal representation of desires: $D_i(x)=r$ represents that individual i desires outcome x with strength r. The implicit assumptions both concern desire strength itself. First, I assumed that an individual's desires have strengths that are at least measurable on an interval scale. This means that an individual's desire strengths can—like cardinal utility—be represented by numbers that carry information about the ratio of the differences between them, such that we can make intrapersonal unit comparisons of the form: $D_i(x)-D_i(y)=n(D_i(y)-D_i(z))$. Second, I assumed that these facts about desire strength are not reducible to facts about preferences over prospects, but that they are psychologically basic. In other words, desires come with interval-scale measurable strengths, in the same way that beliefs plausibly come with interval-scale measurable credences, and people come with interval-scale measurable heights.[14]

These assumptions are controversial. But before I defend them, I would first like to add two further details to this picture, each of which implies the possibility of a certain type of objective interpersonal comparison. The first is that desire strength comes with a non-arbitrary zero-point—that is, a zero-point that isn't just the artifact of an arbitrarily chosen utility function or normalization criterion—because it is possible not only to have a desire of

[14] On the analogy between preference or desire strength (he does not distinguish them) and credence or belief strength, see Bradley (2008).

some strength toward an outcome, but to lack a desire of any strength toward an outcome: "[t]he limit of desire weakness is, of course, zero—not having a desire at all" (Schroeder, 2004: 13). This may seem trivial, but combined with the assumption that desire strength is interval-scale measurable, it entails that desire strength—again, like credence and height—is furthermore *ratio*-scale measurable. This means that we can represent desire strength with numbers that carry information about the ratios (and not just the ratios of the differences) between them, or, in other words, that we can make *intra*personal ratio comparisons like: $D_i(x)=n(D_i(y))$ (*"i* desires *x n* times as strongly as *i* desires *y*"). And from this, it follows in turn that we can at least make a limited sort of *inter*personal comparison. For on the basis of pairs of intrapersonal ratio comparisons like "*i* desires *x* twice as strongly as *y*" and "*k* desires *x* three times as strongly as *y*," we may make interpersonal comparisons of the percentage change that the transition from *x* to *y* would result in for each individual's total desire satisfaction: in this case, it would increase *i*'s desire satisfaction by 100% and increase *k*'s desire satisfaction by 200%, so it would result in twice as large a percentage change for *k* as for *i*. We may put this by saying that, given our assumptions to this point, desire satisfaction is at least "percentage change" interpersonally comparable.[15]

The next detail I would like to add to our picture is that non-zero desire strengths furthermore come with different valences. More specifically, *i* can have a positive "appetitive" desire for *x* ($D_i(x)>0$), a negative "aversive" desire for *x* ($D_i(x)<0$), or—as we have just seen—a neutral lack of desire for *x* ($D_i(x)=0$).[16] This is more controversial, but if desires can indeed be split into these three types, then this, too, implies a limited sort of

[15] See Bossert and Weymark (2004: 1120-1121) and Bradley (2008: 98-99) for a discussion of this sort of comparability.

[16] Among others, Schroeder (2004: 10, 13, 25-26) emphasizes this point.

interpersonal comparability. It allows us to partition individuals into three categories, and to judge that those who positively desire their outcome ($r>0$) have their desires better satisfied than those who lack a desire towards it ($r=0$), and that these latter individuals both have their desires better satisfied than those who are averse to their outcome ($r<0$) and just as well satisfied as all others who lack a desire towards it ($r=0$). We may put this by saying that if desires for outcomes can indeed be split into these three types, then desire satisfaction is "valence" interpersonally comparable.[17]

Altogether, then, we have reached the following account of desire:

(D_o) Individuals have desires directed at outcomes, with psychologically basic strengths that permit intrapersonal ratio comparisons and that have positive or negative valences.

And we have seen that, if D_o holds, then desire satisfaction is at least percentage change and valence interpersonally comparable. This, already, is a significant result. Percentage change interpersonal comparability allow us to make objective judgments like "giving $500 to *i* would increase *i*'s desire satisfaction by 100%, while giving it to *k* would increase *k*'s desire satisfaction by only 5%," and can ground a view on which, for example, one outcome is better than another if the *product* of each individual's desire satisfaction is higher in the first than in the second (see Kaneko and Nakamura, 1979 for a discussion of this view). And valence interpersonal comparability allows us to make objective judgments like "*i* (who has a positive desire satisfaction level) has her desires better satisfied than *k* (who has a negative desire satisfaction level)," and can ground a view on which, for example, we give absolute

[17] Bossert and Weymark (2004: 1121) have again noted this sort of interpersonal comparability.

priority to the promotion of the desire satisfaction of those who have a negative desire satisfaction level over those who have a positive level. Importantly, neither sort of comparison is undermined by anything like the conceptual problem of interpersonal comparisons of preferences satisfaction. The formal properties of desire entail the possibility of such comparisons; they do not undermine it.

Assuming that D_o is an accurate picture of desire, it is therefore worth investigating further the full range of options available to a moral or political theory that allows these, and only these, two forms of objective interpersonal comparability. But it is also worth investigating whether desires allow for a wider range of interpersonal comparisons as well. We are therefore left with two questions. First, is it plausible to think that we do have desires as described in D_o, such that we can at least make the limited range of interpersonal comparisons discussed so far? And, second, is it plausible to think that desires allow for a wider range of interpersonal comparisons (including level and unit interpersonal comparisons), too? In the remainder of this section, I consider each question in turn.

4.2 A defense of limited interpersonal comparability

Although D_o fits well with our everyday thinking about desires, one might complain that it is psychologically unrealistic. Earlier, I suggested the following relation between desires and preferences: if *i* has a primitive (that is, psychologically basic) desire for *x* that is stronger than *i*'s primitive desire for *y*, then *i* has a derived preference for *x* over *y*. But one might instead claim that preferences are primitive, and that the sort of claims we make in everyday life about desires can either be derived from claims about preferences or paraphrased away. After all, we can tell by introspection that we have preferences, and it is a mathematical truth that, on the basis of our preferences over prospects, we may derive an interval-scale measure of preference strength—but we don't seem to be able to introspect anything like primitive

desire strengths at all (compare von Neumann and Morgenstern, 1944: 16). Why, then, shouldn't we think that claims about desire strength should be reinterpreted as claims about preference strength derived from claims about preferences over prospects, and that a claim such as "*i* has a desire for *x*"—which, on its face, seems to require desires rather than preferences—should be paraphrased as reporting, say, that *i* prefers *x* to some contextually salient alternative? Adopting this strategy seems to account for our everyday thinking about desires, but doesn't require us to posit desires that come with psychologically basic, introspectively inaccessible strengths. So what justifies this posit?

This is an important challenge, but, conveniently enough for me, Pollock has provided a compelling response: that we must have primitive desires from which we derive our preferences, because the limited storage capacity of the human brain would otherwise render it impossible for us to have enough preferences. Pollock's argument begins with a distinction between outcomes—"'total' ways the world might be, i.e., complete possible worlds"—of the sort we have so far been assuming that preferences and desires range over, and features of outcomes—"partial descriptions of ways the world might be"—of the sort we perhaps more commonsensically think desires take as their objects (2006: 22). He then asks us to suppose—what is an enormous simplification—that every possible outcome can be characterized by some configuration of 300 two-valued parameters, each representing a feature that an outcome might either realize or fail to realize. This leaves 2^{300} outcomes, a number which is "12 orders of magnitude larger than the number of elementary particles in the universe," and which it would therefore clearly be impossible for an individual to have primitive preferences over (Pollock, 2006: 24). Indeed, even if we simplify further and assume that we could characterize every outcome using only 60 two-valued parameters, the number of primitive preferences an individual would need to form preferences over all these

outcomes would still exceed the estimated storage capacity of the human brain (Pollock, 2006: 25). So this is clearly impossible, too.

The upshot, then, is that either agents "will rarely have preferences between actual states of the world"—an untenable result for either side of this debate, which I will therefore assume to be false—or else preferences "are computed from something else that can be stored more compactly" (Pollock, 2006: 26). But, Pollock asks, what could this more compact form of storage be? The answer is not "desires for outcomes" (of the sort described by D_o), because if we assume we have one desire for every outcome then this, too, will exceed our storage capacity: if there are 300 two-valued parameters, we will need 2^{300} desires. So instead, Pollock argues that the answer must be: "desires for features." For if we have desires for features with psychologically basic strengths, and if we can add together the strengths of our desires for the various features that constitute an outcome to obtain an overall desire strength for that outcome, then we will need at most 600 desires of particular strengths (one for each parameter value) to derive desires of particular strengths for each of these 2^{300} outcomes—from which we can in turn derive preferences. As Pollock puts it, this "is the difference between the trivial and the impossible" (2004: 26).

Pollock's argument turns the objection to D_o on its head. The worry there was that it is more psychologically realistic to think that we have primitive preferences than that we have primitive desires with psychologically basic strengths. But Pollock shows that we must have desires like this, because it would otherwise be impossible for us to have a wide range of preferences over outcomes, let alone enough preferences over prospects to derive a cardinal utility function that represents preference strength. More precisely, Pollock's argument—combined with our earlier discussion—yields the following account of desire:

(D_f) Individuals have desires directed at features, with psychologically basic strengths

> that permit intrapersonal ratio comparisons and that have positive or negative valences. An individual's desire strength for an outcome is the sum of the strength of her desires for the features that constitute that outcome.

D_f entails D_o, and therefore that desire satisfaction allows for the limited sorts of interpersonal comparisons mentioned earlier. It does, however, require us to engage in some reinterpretation. In particular, if *i* has a positive or negative desire strength for outcome *x* (respectively, $D_i(x)>0$ or $D_i(x)<0$), we must now interpret this to mean, respectively, either that the sum of the strengths of *i*'s positive desires for the features in *x* is greater than the sum of the strength of *i*'s aversions to the features of *x* or that the reverse holds. And, similarly, if *i* has a neutral desire strength towards *x* ($D_i(x)=0$), we must now interpret this to mean either that *i* lacks desires for any of the features of *x*, or that the strength of *i*'s positive desires for the features of *x* and the strength of *i*'s aversions for *x* exactly balance out.

But D_f faces two further challenges. The first is that, according to D_f our desires for features are independent, in the sense that, for any two features f_1 and f_2, the strength of one's desires for the conjunction of those features must equal the sum of the strength of one's desires for each feature. But this doesn't always seem to be the case. For example, I might positively desire to eat sardines, and positively desire to eat chocolate, but be averse to eating sardines and chocolate together (Weirich, 2004: 202). This is a perfectly ordinary phenomenon—Pollock's favorite example instead involves eating ice cream with a dill pickle (Pollock 2006: 26)—but it seems incompatible with D_f, which appears to predict that I should instead have a positive desire for eating both items together that is equal to the sum of my positive desires for eating each by itself. So something has to give.

One response to this challenge is to deny that, properly described, cases like these ever actually occur. Thus, Weirich defends D_f by arguing that the above case is not a real

counter-example to the independence of desires, because, strictly speaking, "the objects of [my desires] are the taste of sardines alone and the taste of chocolate alone. These [features] are not realized when both foods are eaten together" (Weirich, 2004: 202). Weirich similarly argues that cases where the strength of my desires for the conjunction of two features is greater than the sum of the strength of each desire on its own—say, a case where I desire pleasure at time t_1, and I desire pleasure at time t_2, but the strength of my desire for pleasure at both t_1 and t_2 is greater than the sum of the strengths of these first two desires—do not present true counter-examples to independence either. Instead, they may be redescribed as cases where I actually have one desire for each feature, and one for a third feature that the joint realization of the first two features entails: in this case, the third desire would be for "pleasure constancy" (Weirich, 2001: 56-57, 71). So they provide no problem for D_f, either.

Weirich's strategy of redescription can probably handle most purported counter-examples to the independence of desires. But it is important to realize that even if there are some cases where our desires for features are truly interdependent, this makes no real trouble for us. All we need to do is follow Pollock (2006: 67-71) in claiming that, whenever an individual desires a conjunction of features with a strength that diverges from the sum of her desires for each feature, we must impute to her a new desire that takes as its object this conjunction of features and reinterpret the former two desires as desires for each feature in the absence of the other. So, for example, suppose my desire strength for feature f_1 is 3, and my desire strength for feature f_2 is 5, but my desire strength for the conjunction of f_1 and f_2 is -3 rather than 8. In that case, we must simply impute to me the following three, more carefully regimented desires: a desire for f_1 in the absence of f_2 with strength 3, a desire for f_2 in the absence of f_1 with strength 5, and a desire for f_1&f_2 with strength -3.

Allowing for interdependent desires in this (or some similar) way may seem to

threaten the compactness of our account, thereby undermining its advantage over a view on which preferences are primitive: the more interdependencies there are between our desires for features, the more desires we will need to encode. But this will only be a problem if there are too many interdependencies. So long as our desires are, for the most part, independent, we will still be able to derive a huge number of preferences over outcomes from a relative handful of desires for features. What is important for us, then, is that if we did not have desires with psychologically basic strengths that were at least largely independent, it would be impossible for us to have preferences over a wide range of outcomes. Since we do have such preferences, it follows that we must have such desires.

There remains one characteristic of our picture of desire that stands in need of defense: that there is indeed a distinction between positive desires and negative aversions. This distinction may seem obvious. For example, it may seem obvious that my desire to go out to dinner is different in kind from my aversion to being fired from my job—that the former is a positive or "pro" attitude and the latter a negative or "con" one. But there is an alternative explanation available: that I only have one desire-type, and that what I actually have is a desire for going out to dinner alongside a desire for *not* being fired from my job. In the current context, this difference is important, because if to be averse to something were really just to positively desire its absence, then this would threaten the valence comparability of desire satisfaction: if all our desires were positive, then it would be impossible for an individual to have an on balance negative desire strength towards an outcome. The question, then, is whether there is indeed a genuine distinction between appetites and aversions. Is there, in other words, a difference between positively desiring that some feature doesn't obtain, and being averse towards it obtaining?

Although considerations of storage capacity cannot decide this issue, there are

nonetheless good reasons to hold onto this distinction. In the first place, it is present in our phenomenology. As Schroeder points out, with positive desires, "the primary phenomenology of desire satisfaction is itself positive—joy, or contentment, for example," but with negative aversions, "the primary phenomenology of desire satisfaction is emotionally flat, or a sensation of relief" (2004: 26). And, likewise, when one anticipates that one will fail to satisfy a positive desire, this causes disappointment, whereas when one anticipates that one will satisfy an aversion this leads to anxiety (2004: 132). This seems true to our experience. When I believe my boss won't let me get off work early to go to dinner I feel disappointment, which changes to joy when she grants me the night off; when I believe she will fire me I feel anxiety, which changes to relief when she reveals I still have my job. Furthermore, as Sinhababu notes, there is psychological research supporting a similar disanalogy between positive desires and aversions when it comes to attention: "Positive desires direct more attention towards desired things, with hungry people attending more to food than missing out on food. Aversions direct more attention towards things we desire to avoid, with arachnophobes attending more to spiders than to freedom from spiders" (2017: 40). So this, too, suggests, a difference between positive desires and aversions.

This brings us to the second source of support for this distinction: that it is supported by our current best neuroscience and empirical psychology. Although it is, of course, controversial exactly what makes a particular mental state a desire, it is uncontroversial that, at least in humans, desires have an important influence on motivation and decision-making, as well as some close connection to positive and negative feelings and to our patterns of attention. And as Schroeder explains, it is textbook neuroscience that desires therefore must be realized in the reward system, located in the ventral tegmental area and substantia nigra pars compacta of the brain (2004: 48-57): "The reward system is… a

normal cause of overt and covert action, of positive and negative feelings, and of the [attentional] effects most associated with intrinsic desires. The reward system is also the only thing in the brain that is a common cause of all of these effects" (Arpaly and Schroeder 2014: 127). Without going into the details, what is important for our purposes is that neuroscientists recognize a distinction between this reward system and a parallel "punishment" system—which is less well understood, but likely located in the dorsal raphe nucleus—that seems instead to realize negative aversions (Schroeder 2004: 54-57). And this same distinction is prevalent among experimental psychologists. As Elliot writes in his introduction to the *Handbook of Approach and Avoidance Motivation*, the distinction between positive "appetitive" desires and negative "aversive" desires "has been utilized in all of the major theoretical approaches that have been employed to scientifically explain behavior, regardless of how these approaches might be characterized" (2008: 5). While Elliot prefers to put the distinction in terms of approach and avoidance motivation, he notes that ""appetitive–aversive' and 'approach–avoidance' have been proffered and used in highly similar fashion in the literature" (2008: 10), and, moreover, that "positive or negative valence is construed as the conceptual core of the approach–avoidance distinction" (2008: 8).

So not only do we experience a distinction between positive desires and negative aversions in our mental life, but the distinction enjoys broad empirical support from both neuroscience and experimental psychology. And this, by the way, provides an independent argument that we have desires rather than merely preferences, since it would be impossible to draw this distinction if preferences were primitive: a preference for f_1 over f_2 cannot distinguish between cases where one positively desires f_1 more than f_2, positively desires f_1 and is averse to f_2, or is less averse to f_2 than f_1. We therefore have overlapping grounds for holding that we have desires as described in D_f and D_o, and, in turn, for holding that desire

satisfaction is both percentage change and valence interpersonally comparable. So desires at least allow for these limited sorts of interpersonal comparability, even if preferences do not.

4.3 A partial defense of wide interpersonal comparability

So far, we have seen that individuals have desires for features with psychologically basic strengths that can (at least for the most part) be added together to obtain psychologically basic desire strengths for outcomes, and that come with different valences. We have also seen that, as a result, desire satisfaction is percentage change and valence interpersonally comparable. But it remains to be seen whether desire satisfaction allows for a wider range of interpersonal comparisons, such as level and unit interpersonal comparisons, as well. Since we know (given Pollock's argument from storage capacity) that it must be possible to compare the strengths of two desires for features when those desires belong to the same individual, the relevant question is whether we can similarly compare the strengths of desires for features when those desires belong to different individuals. In particular, we want to know if the following holds:

> (D_f*) Individuals have desires directed at features, with psychologically basic strengths that permit *inter*personal ratio comparisons and that have positive or negative valences. An individual's desire strength for an outcome is the sum of the strength of her desires for the features that constitute that outcome.

For just as D_f implies D_o, D_f* implies:

> (D_o*) Individuals have desires directed at outcomes, with psychologically basic strengths that permit *inter*personal ratio comparisons and that have positive or negative valences.

So if desires for features have interpersonally rather than merely intrapersonally comparable strengths, desire satisfaction will be ratio interpersonally comparable, and—since ratio comparability implies level and unit comparability—it will therefore be level and unit interpersonally comparable as well.

At this point, we must finally get more precise about desire strength. Just above, I suggested that, in humans, desires have some important connection to motivation as well as to pleasure and pain. But there is still a further question of what it is that makes a particular mental state a desire—in humans or any other being—and, on this point, theorists of desire disagree. On what is perhaps the most popular view, what is essential to desires is that they dispose one to act; as Stalnaker puts it, "[t]o desire that *P* is to be disposed to act in ways that would tend to bring it about that *P* in a world in which one's beliefs, whatever they are, were true" (1984: 15; compare Smith, 1994: 115). On another popular view, what is essential to desires is instead their connection with pleasure and pain: desires dispose one to feel pleasure when one believes they are satisfied, and aversions dispose one to feel pain (e.g., Morillo, 1990; Strawson, 1994, ch. 9). And Arpaly and Schroeder (Schroeder, 2004: ch. 5; Arpaly and Schroeder, 2014: ch. 6) have recently developed a third, learning-based theory of desire, on which desires "*cause*, but are not *constituted by*, their familiar effects on motivation or pleasure"' (Arpaly and Schroeder, 2014: 126). Instead, to desire something "is for representations of [it] to tend to contribute to the production of a reinforcement signal… in the sense made clear by computational theories of…'reinforcement learning"' (Schroeder, 2014: 66; compare Railton, 2012 and Dretske, 1988: ch. 5).[18]

Here, it does not matter what *desires* are so much as what desire *strength* is, and each of

[18] See Schroeder (2004) and Arpaly and Schroeder (2014: ch. 6) for a thorough discussion of these three views.

these theories of desires comes with such a theory of desire strength—one that isn't strictly speaking the sole province of that theory, but that is most closely associated with it.[19] According to the first, a desire's strength is identical with its motivational strength: the stronger one's desire is, the greater the causal role it plays in producing action. According to the second, desire strength is identical with the intensity of pleasure or pain that a desire disposes one to experience: stronger desires tend to produce more pleasure when one believes they are satisfied, and stronger aversions to produce more pain. Finally, according to the third theory, a desire's strength is identical with the contribution it makes to reinforcement learning. Perhaps there are other theories of desire strength, but here I focus on these three. I consider them not with the aim of showing that we can indeed make a wide range of objective interpersonal comparisons of desire strength, but with the more modest goals of, first, suggesting some promising strategies for assuring this sort of interpersonal comparability, and, second, showing what further sorts of questions the viability of these strategies turn on.

Consider first the motivational theory of desire strength. Here, the best developed account has been provided by Mele, who argues that a desire's motivational strength depends on its physical basis:

> just as the fact that one vase is more fragile than another is grounded in differences in the respective physical bases of the fragility of the two vases (e.g., crystalline

[19] For example, Sinhababu holds that in order to be a desire, a mental state must have some connection both to motivation and to pleasure and pain (2017: 32) but that desire strength depends entirely on the former aspect (2017: 23). He therefore accepts a motivational theory of desire strength without accepting a (purely) motivational theory of desire.

> structures), the fact… that one member of a pair of… desires of mine is stronger than another is presumably grounded in differences in the respective physical bases of the strength of the two desires... It may be reasonably suggested that we can conceive of the relative strength of a human agent's… desires at a given time as analogous, in the respect mentioned, to the relative fragility of the vases stored in my son's kitchen and the relative elasticity of the various rubber bands in my desk drawer: there is a physical basis in each case, and comparative truths about the fragility, the elasticity, and the motivational strength of the relevant items are grounded in differences in the physical bases. (2003: 173)

Although Mele himself is only concerned with making intrapersonal comparisons of desire strength, such an approach seems to allow for interpersonal comparisons, too. For if the relative strength of our desires depend on their physical bases, then we can (at least in principle) compare the physical bases of my desires to yours, and so compare their motivational strengths accordingly. In other words, interpersonal comparisons of desire satisfaction may amount to comparisons of whatever it is in the brain that physically realizes different motivational tendencies; as Weirich puts a similar point, "[u]tilities may depend on agents' beliefs and desires in complex ways and yet be interpersonally comparable because they are ultimately reducible to agents' physical states, which are comparable" (2001: 83). Of course, whether this strategy is ultimately viable turns on various further questions in neuroscience (what is the physical basis of motivational strength?) and the philosophy of mind (can we identify motivational strength with its physical basis?). Further work, perhaps of an interdisciplinary nature, is therefore needed before anything more definitive can be said.

Let us turn, then, to the theory that desire strength is identical with the intensity of

pleasure or pain that a desire disposes one to experience. On this view, objective interpersonal comparisons of desire strength will be possible just in case it is possible to make objective interpersonal comparisons of pleasure or pain. Interestingly, most philosophers assume that such comparisons pose no special conceptual or metaphysical problem, and instead treat the problem posed by them as a purely epistemic one (e.g. Hausman, 1997: 99; Brandt, 1979: ch. 13). And such philosophers do have common sense on their side; as Railton puts it, "[i]t seems to me quite unconvincing to say that we simply have no idea whether, on a given occasion, the pain of McAllister's wasp sting is greater or less than the pain of McMurtry's mosquito bite…[or] to say that we typically have no idea whether the distress of McNeil at the loss of his job is worse than both together (1992: 44). But, if we are to get more precise, it is clear that a parallel strategy of appealing to the physical bases of pleasure and pain is available here. An additional strategy for assuring the interpersonal comparability of desire strength would be to identify the intensity of pleasure with its felt intensity (Charlton 1988: 127), and to argue that such feelings are interpersonally comparable. But whether one considers this last strategy viable will depend on one's position on the interpersonal comparability of feelings more generally, so exercising this strategy will once again require one to venture deep into the philosophy of mind.

Finally, consider the view that desire strength is identical with the contribution it makes to reinforcement learning. This requires some further explanation. Suppose, again, that I desire to go out for dinner. I ask my boss if I can leave work early, expecting her not to let me—but she surprises me and lets me go. This increases or "reinforces" my tendency to ask her again in the future: because asking her if I could leave work early led to a more desired outcome than I expected it to, I will be more likely to ask her in the future. More generally, when I perform an action that leads to a more desired than expected outcome, this

makes me more likely to perform that action in the future, and when I perform an action that leads to a less desired than expected outcome, this makes me less likely to perform it again. This is reinforcement learning. And, according to Schroeder, what makes one desire stronger than another is that it contributes more to reinforcement learning than another: everything else being equal, a strong desire to go to dinner will greatly increase my tendency to ask my boss to let me off in the future, while a weak desire to go to dinner will only slightly increase it (2004: 138-144).

According to this view, then, objective interpersonal comparisons of desire strength will be possible just in case it is possible to make objective interpersonal comparisons of the extent to which one's representations of different objects contribute to reinforcement learning. Here, we might once again turn to the relevant physical bases. But building on a suggestion of Schroeder's (personal correspondence), we might also adopt something like the following strategy. Take a case in which, in a given context, I have a tendency to perform some action with probability p, and in which, as a result of my performing that action, I represent that feature f_1 is realized (and that no other change occurs). If I didn't expect this, and there is no change to my tendency to perform that action, then assign a desire of strength 0 to my desire for f_1. Otherwise, assign desire strengths in proportion to the percentage change they made to p—for example, if p increased by 0.3, assign a desire strength of 0.3 to my desire for f_1, if it increased by 0.6, assign a desire strength of 0.6 to f_1, and so on. Finally, make interpersonal comparisons of desire strength by comparing these probability changes. For example, if neither of us expected f_1 to obtain, we both perform an action that results in f_1, and my probability of performing that action in the future increases by 0.6 while yours increases by 0.3, then your desire for f_1 is twice as strong as mine. The situation is more complicated, but not in an ultimately problematic way, in cases where

people do expect some change to occur or where multiple features change at once. Even then, we may interpret claims about desire strength as counterfactual claims about how people's probability of performing an action would change in the simpler sort of case just described.

This last approach is, I think, especially promising, but a full defense of it would require us to engage more with the relevant neuroscience and with the literature on reinforcement learning than I am able to go into here. So while more still needs to be said, what I have said will have to do for now. Desires are at least limited interpersonally comparable, and whether or not they allow for a wider range of objective interpersonal comparisons depends on various further questions in the theory of desire, the philosophy of mind, neuroscience, and the theory of reinforcement learning. These are questions that those working on the topic of interpersonal comparisons of welfare have, until now, all but ignored. They cannot afford to do so any longer.

5. **Conclusion**

While I cannot claim to have shown conclusively that desire strength permits a wide range of objective interpersonal comparisons—and while I have not even begun to address issues relating to the epistemology of interpersonal comparisons of desire satisfaction that have no doubt occurred to many readers—I hope to have shown at the very least that the sort of questions we must ask about the possibility of objective interpersonal comparisons of desire satisfaction are different in kind from those we must ask about interpersonal comparisons of preference satisfaction, and, furthermore, that what questions we must ask depends on what we think desire strength is. To address these questions in any more depth would require us to dive into the various further issues just mentioned, but at this stage, I see little reason to think that objective interpersonal comparisons of desire strength will prove impossible in the

way that interpersonal comparisons of preference satisfaction seem to be. Indeed, even if level, unit, and ratio interpersonal comparisons of desire satisfaction do prove impossible, I have shown that desire satisfaction remains at least percentage change and valence interpersonally comparable. These are therefore forms of interpersonal comparability that require more investigation than they typically receive.

In closing, then, I would like to urge philosophers and economists to worry less about interpersonal comparisons with preferences, and more about interpersonal comparisons with desires. This shift of focus is necessary if progress is to be made on the perennial problem of interpersonal comparisons of welfare. And progress on this topic is indeed necessary, given that nearly every moral or political theory requires us to make objective interpersonal comparisons of welfare, and given the widespread appeal of the view that welfare consists in the satisfaction of preferences or desires.

References

Adler MD (2012) *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. New York: Oxford University Press.

Adler MD (2014) Extended Preferences and Interpersonal Comparisons: A New Account. *Economics and Philosophy* 30(2): 123-162.

Arpaly N and Schroeder T (2014) *In Praise of Desire*. New York: Oxford University Press.

Arrow KJ (1963) *Social Choice and Individual Values*. Newhaven: Yale University Press.

Arrow KJ (1977) Extended Sympathy and the Possibility of Social Choice. *The American Economic Review* 67(1): 219-225.

Bradley B (2008). Comparing Evaluations. *Proceedings of the Aristotelian Society* 108: 85-100.

Brandt RB (1979) *A Theory of the Good and the Right*. New York: Oxford University Press.

Broome J (1993). A Cause of Preference Is Not an Object of Preference. *Social Choice and Welfare* 10(1): 57-68.

Bossert W and Weymark JA (2004) Utility in Social Choice. In: Barbera S, Hammond P, and Seidl C (eds.) *Handbook of Utility Theory: Volume 2 Extensions*. New York: Springer, pp. 1099-1177.

Charlton W (1988) *Weakness of Will*. New York: Basil Blackwell.

Dretske F (1988) *Explaining Behavior: Reasons in a World of Causes*. Cambridge: MIT Press.

Elliot AJ (2008). Approach and Avoidance Motivation. In: Elliot AJ (ed.) *Handbook of Approach and Avoidance Motivation*. New York: Psychology Press, pp. 3-14.

Gauthier D (1986) *Morals by Agreement*. New York: Oxford University Press.

Goldman A (1995) Simulation and Interpersonal Utility. *Ethics* 105(4): 709-726.

Greaves H and Harvey L (2017). Aggregating Extended Preferences. *Philosophical Studies* 174(5): 1163-1190.

Greaves H and Harvey L (2018) Extended Preferences and Interpersonal Comparisons of Well-Being. *Philosophy and Phenomenological Research* 96(3): 636-667.

Griffin J (1991) Against the Taste Model. In: Elster J and Roehmer JE (eds.), *Interpersonal Comparisons of Well-Being,* 45-69. New York: Cambridge University Press.

Hare RM (1981) *Moral Thinking: Its Method, Levels, and Point*. New York: Oxford University Press.

Harsanyi JC (1955) Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *The Journal of Political Economy* 63(4): 309-321.

Harsanyi JC (1977a) Morality and the Theory of Rational Behavior. *Social Research* 44(4): 623-656.

Harsanyi JC (1977b) *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. New York: Cambridge University Press.

Hausman DM (1995) The Impossibility of Interpersonal Utility Comparisons. *Mind* 104(415): 473-490.

Hausman DM (1997) The Impossibility of Interpersonal Utility Comparisons--A Reply. *Mind* 106(421): 99-100.

Kaneko M and Kenjiro N (1979) The Nash Social Welfare Function. *Econometrica* 47(2): 423-435.

Mele AR (2003) *Motivation and Agency*. New York: Oxford University Press.

Morillo C (1990) The Reward Event and Motivation. *Journal of Philosophy* 87(4): 169-186.

Pollock JL (2006) *Thinking About Acting: Logical Foundations for Rational Decision Making.* New York: Oxford University Press.

Railton P (2012) That Obscure Object, Desire. *Proceedings and Addresses of the American Philosophical Association* 86(2): 22-46.

Railton P (1991) Some Questions About the Justification of Morality. *Philosophical Perspectives* 6: 27-53.

Robbins L (1932) *An Essay on the Nature and Significance of Economic Science*. New York: Macmillan.

Roberts K (1997) Objective Interpersonal Comparisons. *Social Choice and Welfare* 14(1): 79-96.

Schick F (1971) Beyond Utilitarianism. *Journal of Philosophy* 68(20): 657-666.

Schroeder T (2004) *Three Faces of Desire*. New York: Oxford University Press.

Sen A (1970) *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Sen A (1977) On Weights and Measures: Informational Constraints in Social Welfare Analysis. *Econometrica* 45(7): 1539-1572.

Sen A (1979) Interpersonal Comparisons of Welfare. In: Boskin MJ (ed.), *Economics and Human Welfare: Essays in Honor of Tibor Scitovsky*. New York: Academic Press, pp. 182-201.

Sen A (1997) Individual Preference as the Basis of Social Choice. In: Arrow KJ, Sen A, and Suzumura K (eds.), *Social Choice Re-Examined*. New York: St. Martin's Press, pp. 15-37.

Sinhababu, N (2017) *Humean Nature: How Desire Explains Action, Thought, and Feeling*. New York: Oxford University Press.

Smith M (1994) *The Moral Problem*. Malden, MA: Blackwell Publishing.

Stalnaker R (1984) *Inquiry*. Cambridge: MIT Press.

Strawson G (1994) *Mental Reality*. Cambridge: MIT Press.

von Neumann J and Morgenstern O (1944) *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Weirich P (2001) *Decision Space: Multidimensional Utility Analysis*. Cambridge: Cambridge

University Press.

Weirich P (1984) Interpersonal Utility in Principles of Social Choice. *Erkenntis* 21(3): 295-317.

Weirich P (2004) *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances.* New York: Oxford University Press.