Abstract

In this essay I support, develop and apply a theory of hedonic value. These tasks are interwoven, but principally, I support the theory in chapters 1-4, develop it in chapters 5 and 6, and apply it to a challenging cluster of problems in chapter 7.

Sentient experience, I suggest in chapter 1, provides key evidence for founding ethics: a severely painful experience gives its subject evidence that it's bad in some way. Moreover, similar considerations, as well as analogies, support thinking that all unpleasures (unpleasant experiences) are bad in some way and all pleasures (pleasant experiences) are good in some way. But what type of value and disvalue do they have (if indeed this evidence is not outweighed)?

Experiences that are pleasant or unpleasant are intrinsically so (I argue in chapter 2); so, their normative import doesn't derive from extrinsic motivational or affective conditions that their pleasantness or unpleasantness might be thought to consist in. In chapter 3 I extend my argument in chapter 1 to the conclusion that pleasures and unpleasures have agent-neutral moral significance. Hence, I have just as much basic reason to promote your hedonic well-being as mine. And even the pleasure of potential persons matters, I argue in chapter 4; the fact that a person would feel pleasure is a reason to create her.

In chapter 5 I argue that *being hedonically better than* is not a transitive relation. With this result in hand, I offer several snippets of advice and a host of principles in chapter 6 for assessing the hedonic value of states of affairs. Along the way I argue against a higher/lower distinction for pleasures and for the thesis that pleasures have value when undeserved or taken in bad objects. Finally, in chapter 7 I show that the theory of hedonic value I've developed entails a viable set of solutions to the problems Derek Parfit poses for moral theory in *Reasons and Persons*.

Pleasures, I conclude, are intrinsically good and unpleasures are intrinsically bad.

Hedonic Value

by

Stuart Craig Rachels

B. A., Emory University, 1991
Philosophy and Politics, University of Oxford, 1993

Dissertation

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy to be awarded by Syracuse University in the field of Philosophy in August, 1998.

Approved: J. F. Bennett

(Professor Jonathan Bennett)

Date: July 20 1997

The Graduate School
Syracuse University

We, the members of the Oral Examination Committee,
hereby register our concurrence that

Stuart C. Rachels

satisfactorily defended his dissertation on

Monday, July 20, 1998

Examiners:

John Hawtthorne _____ (Please sign)

John Robertson _____ (Please sign)

Michael Stocker _____ (Please sign)

Stewart Thau _____ (Please sign)

_____ (Please sign)

Advisor:

Jonathan Bennett _____ (Please sign)

Oral Examination Chair:

Donald Siegel _____ (Please sign)

Contents

Chapter 1: Founding Ethics

How, if at all, can moral philosophers found ethics? I'll clarify the question. To found ethics would be to give evidence for normative doctrines without presupposing any—already a tall order, and such evidence must outweigh any contrary evidence. Moreover, those doctrines must be, or must help support doctrines that are, sufficiently robust; merely making a good case against nihilism isn't enough, for merely showing that normative reasons exist doesn't ground any ethical system or project. So, to found ethics, one must also show what some of the reasons are, or at least how to find them.

Does Ethics Need No Foundation?

Some grand ways of theorizing do not assume that ethics can be founded. Consider, for example, the idea of reflective equilibrium, as understood by Rawls. We enjoy such equilibrium when our considered beliefs about moral test-cases are consistent with our preferred principles.[1] This goal might be important, but my finding equilibrium provides no argument against nihilism; it merely entails my confidence in the other direction.

Some philosophers believe that ethics cannot be founded. Beardsley, for instance, says,

> When we are in the position of having to decide what is valuable, or more valuable, we are in Dewey's "problematic situation," and such a situation is one in which certain ends are in grave doubt and others are (on that occasion) taken as temporarily fixed. If the value of everything in the situation were in question at once, nothing could be decided at all . . .[2]

McDowell, commenting on Aristotle, says much the same thing.[3] Why do many philosophers believe that ethics can't be founded? "To found ethics," they might say, "would require one to have evidence for moral conclusions from a purely factual or scientific standpoint. But science concerns what is, not what ought to be. Hence, morality can be defended from within a moral perspective, but moral claims can't be positively supported."

Doing normative ethics without a foundation would seem, on its face, intellectually unrespectable. "If the foundations of an ideological position are knocked out from under it," Singer says, "new foundations will be found, or else the ideological position will just hang there, defying the logical equivalent of the laws of gravity."[4] However, philosophers who deny that ethics can be founded typically aren't nihilists; they believe that ethics needs no underpinnings. "Why should it?" they might say; or perhaps, "A good moral system supports itself."[5]

I believe, first, that ethics can be founded. Even if "science concerns what is, not what ought to be," what is can provide evidence for what ought to be. That evidence, of course, need not be conclusive; foundations need not be certain. I'll discuss my proposal in due course. Second, I believe that ethics needs founding. To help explain why, let me first offer evidence for nihilism.

Aristotelian teleology and medieval theology have been displaced by modern science, so normative beliefs can no longer be supported by appealing to purposes or why we are here.[6] A complete description of reality, it now seems, need not entail anything to be of value; the physical sciences can explain and predict all predictable phenomena without using evaluative concepts. And the physical world of science seems insignificant: we seem to have no reason to care about anything. As Sidgwick says,

> Let us examine . . . physical processes. . . . so long as we confine our attention to their corporeal aspect,—regarding them merely as complex movements of certain particles of organised matter—it seems impossible to attribute to these movements, considered in themselves, either goodness or badness. I cannot conceive it to be an ultimate end of rational action to secure that these complex movements should be of one kind rather than another, or that they should be continued for a longer rather than a shorter period.[7]

Sidgwick was a mind/body dualist, so he concluded that "if a certain quality of human Life is that which is ultimately desirable, it must belong to human Life regarded on its psychical side, or, briefly, Consciousness." But what should physicalists think? Dirt has no basic normative significance, and all other physical objects above a certain size are just rearranged dirt. Objectivity in ethics, it seems, can get no hold in the physical universe.

The naturalist's worldview also looks nihilistic from a different perspective. Values imply reasons—not merely motivating reasons, but, as they are variously called, normative, justifying, good or real reasons. What are reasons? We talk about them in three ways: as something we possess (psychological causes of action), as objective grounds (states of the world), and as entities in a Fregean third realm (abstract propositions the thinking

of which may motivate us). If reasons are Fregean entities, then one must explain how we can know anything about them—for example, when they apply, or even that they exist.[8] If reasons are objective grounds, then Sidgwick's worries arise: why should some complex physical movements or states be reasons, but not others? Or suppose in the spirit of naturalism that reasons are in the head. On this view, reasons are physically realized intentional states. But our brains are composed of neurons, glial cells, blood, chemicals, electricity, and so on. Hence, Sidgwick's worries arise again. Don't make the mistake Wittgenstein warned against; don't assume that the mysterious becomes banal if we put it inside the head or make it mental. Finally, it may be said that language deceives us: there aren't items called *reasons*; there are merely meaningful sentences that include words like "reason." However, if we don't quantify over reasons, we are burdened to explain reason-talk in a way that preserves its normativity.

Without a foundation for ethics, this evidence for nihilism carries the day. So, ethics needs founding. Could an unfounded ethical system be self-supporting? The notion of "self-support" is obscure, but anyway, why should a nihilist's worldview support itself less than a moralist's? Examining the foundations of ethics might also help resolve disputes at higher levels of normative inquiry.

An apologist for an unfounded morality might say, "In denying nihilism, I am no worse off than someone who denies another skeptical position—solipsism—without argument. Solipsism might then be said to 'carry the day' in virtue of its theoretical simplicity; but, for whatever reason, we needn't bother refuting it." However, the evidence I've presented supports nihilism more than the mere appeal to theoretical simplicity supports solipsism. And nihilism, like solipsism, has the

advantage of simplicity. Moreover, the idea that value is illusory has weighed heavily on many thoughtful souls; it is not, like solipsism, a wholly unbelievable fantasy. Finally, if moral doctrines can't be supported without assuming moral doctrines, which should we assume?

Two Unsuccessful Arguments Against Nihilism

Now I want to make nihilism stronger by criticizing two arguments against it. These arguments try to give evidence for the existence of normative reasons without entailing what any of them are.

Parfit says, "Suppose that, unless I move, I shall be killed by a falling rock, and that what I most want is to survive. Do I have a reason to move? It is undeniable that I do."[9] Parfit uses this premise to argue against the thesis that objective moral values cannot exist—that they are "too queer to be part of the fabric of the universe." "Since there are some reasons for acting," says Parfit, "it is an open question whether some of these are moral reasons." However, one might use the same premise to argue: "Some reasons for acting exist; this increases the likelihood that there are objective moral reasons for acting."

Let's see whether Parfit's premise supports our conclusion. In what sense do I (obviously) have a reason to get out of harm's way? (i) My own values—that can be denied—plainly entail that I have a reason to avoid the rock. (ii) If I see the rock and leap out of the way, obviously I was moved to do so; hence, I had a "motivating reason" to leap. (iii) Perhaps, if I had complete factual knowledge, then the values I would have would ensure

that I would move out of the way. But options (i)-(iii) are merely actual and counterfactual facts about my psychology; they don't tend to show the existence of objective reasons. Nor can "I have a reason to move" be interpreted as "I would be better off if I moved," for that would merely assume the existence of objective reasons.

Korsgaard asks, "if . . . we admit there are reasons for belief, then why not admit that there are reasons for action as well?"[10] We can interpret "there are reasons for belief" in several ways, but on no interpretation does it support the existence of objective moral reasons. (i) If "there are reasons for belief" just means "people have motives for some of their beliefs," then that would support the existence of motives to act, but not objective ones. (ii) If "there are reasons for belief" means "people are better off having certain beliefs" (as Pascal urged of theism), then Korsgaard's premise, being normative, needs support in this context. (iii) I imagine Korsgaard means something like, "People have nondeductive evidence that some non-normative beliefs are true." But this doesn't support the thesis that people have evidence for beliefs of the form, "S has reason to F." Moreover, we have evidence for certain factual beliefs because we are in causal contact with the natural world, but this doesn't support our being in contact with a normative world.

On its most challenging interpretation, the argument would state, "Believing both *P* and *P entails Q* provides a normative reason for believing Q. If such normative reasons for belief exist, then probably there are normative reasons for acting." But believing both *P* and *P entails Q* provides a normative reason for believing Q only if true belief has objective value—and that can't be assumed in this context. This argument shouldn't be confused with the following: "The conjunction of *P* and *P*

entails Q provides evidence for Q." Indeed, it suffices for Q's being true; but the mere existence of such evidence, as in the third interpretation, doesn't show that evidence bears on what one ought to do or what one ought to be.

Arguments against nihilism such as these may be too formal to succeed; successful arguments are likely to advocate specific objective values.

Objective Values

An item has *objective value* if it provides a normative reason for some agent. Almost every ethical theory posits such values. Foremost among them are *intrinsic values.*

"The intrinsic properties of something," David Lewis says, "depend only on that thing; whereas the extrinsic properties of something may depend, wholly or partly, on something else. If something has an intrinsic property, then so does any perfect duplicate of that thing; whereas duplicates situated in different surroundings will differ in their extrinsic properties."[11] An item's intrinsic value, therefore, depends solely on the item's monadic or nonrelational properties or, equivalently, on its internal nature.[12] Perfect duplicates cannot differ in intrinsic value. Intrinsically good things are good independent of all else, good *per se* or as such. Moreover, they are not merely good for me or good for human beings; all creatures have reason to create them and sustain them. In other words, intrinsic goods are not good *relative* to some person or group; they are just good. Such goods, however, might have intrinsically bad inner aspects or

parts.[13] For example, a life might be intrinsically good, even if a miserable part of it is intrinsically bad. Intrinsic goods are intrinsically valuable-on-the-whole.

Many nonintrinsic, objective values depend on intrinsic values. For example, some *instrumental goods* are valuable by virtue of causing intrinsically good items to exist. (Other instrumental goods cause goods of other types to exist.) Also, consider Moore's intrinsically good "organic wholes." The total value of an organic whole differs from the summed value of its parts.[14] A part has *contributive value*[15] if the organic whole has less intrinsic value without it, but that difference in value is more than the part's intrinsic worth. For example, a gustatory pleasure would have contributive value as part of Jimmy Carter's life if: Carter's having it is intrinsically good; Carter's life would have less intrinsic value without it; but such a pleasure lacks intrinsic value (Pol Pot's enjoying such a pleasure, perhaps, would not be good in any way). Korsgaard's *ends* are a species of contributive values. An end has no intrinsic value but the whole that comprises it does; that whole is the end conjoined with someone's rationally desiring it for its own sake.[16]

Dworkin's *sacred values* are items "whose destruction would dishonor what ought to be honored."[17] But we have no reason to create more of them, as we do with intrinsic values (which Dworkin calls "incremental values"). Persons and paintings, in Dworkin's opinion, have sacred value.

These are the most important categories of objective value: intrinsic goods (including organic wholes), contributive goods, ends, sacred goods, instrumental goods and relative goods.

A Taxonomy of Arguments on Intrinsic Value

Many philosophers agree with Hume that "morality is determined by sentiment"[18] rather than evidence; contemporary skeptics would say that ethical theorizing boils down to "mere intuition," where intuition is just shallow and unsupported belief. This outlook seems most accurate for claims about intrinsic value; I'll mention two reasons why evidence seems especially scarce for judgments of the form "X has intrinsic value."

First, "X has intrinsic value" is a strong claim: it asserts a thesis aboutX and all of X's perfect duplicates. So, to say that a token pleasure has a certain amount of intrinsic value implies the same for all of its perfect duplicates, including the worst tyrant's pleasure. Chisholm, however, misinterprets this point, saying:

> Where knowledge of the instrumental value of a state of affairs thus involves knowledge of its *causal* properties, knowledge of the *intrinsic* value of a state of affairs may be likened to knowledge of its logical properties. For attributions of intrinsic value are necessary. If pleasure is intrinsically good in this world, then it would be intrinsically good in any world in which it might be found. . . . Hence, as Brentano emphasized, the kind of knowledge we have of intrinsic value is properly said to be a priori.[19]

But our knowledge of intrinsic value is no more a priori than our knowledge of any intrinsic properties, for example, our empirical, a posteriori knowledge of how many nucleons an atom contains. If an atom

contains six nucleons, then a perfect duplicate of the atom will also contain six nucleons, in any world in which it might be found.

Second, one can't show that something is intrinsically valuable by showing that it leads to other goods. Hume took this to prove that ultimate ends cannot be determined by argument:

> It appears evident, that the ultimate ends of human actions can never, in any case, be accounted for by *reason*, but recommend themselves entirely to the sentiments and affections of mankind, without any dependence on the intellectual faculties. Ask a man, *why he uses exercise*; he will answer, *because he desires to keep his health*. If you then enquire, *why desires health*, he will readily reply, *because sickness is painful*. If you push your enquiries further, and desire a reason, *why he hates pain*, it is impossible he can ever get any. This is an ultimate end, and is never referred to any other object.[20]

Beardsley even says that "the concept of intrinsic value is inapplicable—that even if something has intrinsic value, we could not know it . . ." because we can give evidence for an item's value only by considering it "in the wider context of other things, in relation to a segment of a life or of many lives."[21] Thus Hume and Beardsley exemplify Korsgaard's generalization: "Modern philosophers have tended to hold that if you can say *why* something is valuable, that *ipso facto* shows that the thing is *extrinsically* valuable."[22]

Other philosophers have been just slightly more optimistic about argument in this area.[23] According to Moore and Mill, only one method can be used to show that something is intrinsically good[24] (though most philosophers reject Moore's intuitionism and Mill's proof of utility), while

in Smart's opinion, ethicists can only "remove confusions and discredit superstitions . . ." and then rely on the reader's benevolent sentiments.[25]

I am a bit more optimistic, however, given the great variety of arguments that philosophers have made about intrinsic value. I'll list twelve types of argument philosophers have brought to bear on whether an item has intrinsic value, eleven of which can be used to argue that an item is intrinsically good. Many of these strategies can also be used to argue that an item has other types of objective value. Most philosophers, I predict, will think that more than one of these argument-types can be fruitfully employed in some context. But the foundational context—in which no values are assumed—is especially demanding. Later I will discuss which of the argument-types might be used to found ethics.

1. One may argue that an item is intrinsically good by appealing to the merits of wider theories entailing it. Such theories may or may not entail plausible evaluative judgments; be internally consistent; be self-defeating;[26] make arbitrary or natural distinctions; have *ad hoc* principles; motivate their principles, goals and restrictions;[27] have desirable breadth; give plausible advice; cohere with established scientific theories (such as evolution by natural selection or the rejection of vitalism); and cohere with viable doctrines having a strong philosophical component (for example, human determinism or atheism). This strategy may also be used to support the idea that some action is right or reasonable. Also, if a theory entails that an item is intrinsically good, one might try to undermine that thesis by arguing against the theory.

2. One may argue that an item is intrinsically good, bad or neutral by appealing to intrinsic facts about it. This argument-type seems promising: to argue that a thing is intrinsically good, why not focus on its intrinsic nature? One can appeal, for example, to:

a) facts about its internal structure;

Korsgaard says: "there is an order within 'valuable wholes,' a conditioning of some elements by others . . . This order reflects the reason why the wholes are good."[28] Why, she asks, is happiness conjoining a good will intrinsically good, unlike happiness conjoining a bad will? "happiness in the one case is good because the conditions under which it is fully justified have been met (roughly, because its having been decently pursued makes it deserved). Those internal relations reveal the reasons for our views about what is valuable . . ." And Dworkin distinguishes "two processes through which something becomes sacred for a given culture or person. [Hence, Dworkin's topic is sacred, relative value.] The first is by association or designation. . . . Many Americans consider the flag sacred because of its conventional association with the life of the nation; the respect they believe they owe their country is transferred to the flag."[29] "The second way something may become sacred is through its history, how it came to be. In the case of art, for example, inviolability is not associational but genetic: it is not what a painting symbolizes or is associated with but how it came to be that makes it valuable."

b) facts about its parts;

Of course, if a whole's part is supposed to be intrinsically good, then we might also want an argument for that.

"The value of a whole," says Moore, "must not be assumed to be the same as the sum of the values of its parts."[30] To exemplify organic wholes,

Moore says, "we cannot attribute the great superiority of the consciousness of a beautiful thing to the mere addition of the value of consciousness to that of the beautiful thing." Here the consciousness and the beautiful thing exist simultaneously. But one might also think that the value of a whole doesn't equal the summed values of its temporal parts. Stocker says, "increases in the goodness and constituent goods of a life need not make the life a better life."[31] And Griffin says, following Seabright,

> suppose a person who has had seventy years of very good life has an accident, and imagine each of the alternative results. In one, he is killed outright. In the other, he lives on, diminished and in some pain, for another ten years, but years which on their own would be worth living. Now if maximizing means adding, then there is no question but that the second outcome is better: seventy years plus ten years of positive value. Yet if evaluation is sometimes holistic there is a question. Perhaps the life as a whole would be better if he is killed outright; perhaps the person would rationally prefer it.[32]

But even if the value of the whole doesn't *always* equal the summed value of its parts, the value of the parts might still provide a fairly reliable basis for drawing conclusions about the whole.

c) facts about the nature or essence of the item (where the item's essence or nature is internal to it);

Some people believe that post-adamite humans are inherently sinful; were this true, it would tend to show that we're intrinsically bad.

d) facts about the type of thing the item is;

If persons are nothing over and above their bodies and experiences, this may tend to show that persons aren't intrinsically good.

3a. One may argue that an item is (or isn't) intrinsically good by appealing to how it is viewed by God, an ideal observer, a fully rational person or a competent judge.[33]

3b. One may argue that an item is (or isn't) intrinsically good by appealing to actual human attitudes about it.

The strategies of 3a and 3b may also be used to support the idea that behavior is right or reasonable.

According to Aristotle, "Eudoxus thought that pleasure was the good because he saw all things, both rational and irrational, aiming at it . . ."[34] Aristotle evidently agreed, for he said, "the fact that all things, both brutes and men, pursue pleasure is an indication of its being somehow the chief good . . ." Hume says, "An action, or sentiment, or character is virtuous or vicious; why? because its view causes a pleasure or uneasiness of a particular kind."[35] And Mill holds that "the sole evidence it is possible to produce that anything is desirable is that people do actually desire it."[36] Narveson agrees.[37]

4. One may argue that an item is (or isn't) intrinsically good by appealing to perception or moral intuition. C. I. Lewis believes that we have "immediate experiences of good and bad."[38] One may also say that we immediately experience some behavior as right or wrong. More recently, Nagel says that:

> If the possibility of real values is admitted, specific values become susceptible to a kind of observational testing . . . In ethics, one infers from appearances of value to their most plausible explanation in a theory of what there is reason to do or want.[39]

For maximal reliability, we may be advised to intuit judgments about an item while imagining it in isolation, and to confront it or its duplicates in a variety of empirical circumstances. In each case, the aim is not to confuse the item's intrinsic nature with its extrinsic properties. Moore thinks that imagining an item in isolation is "the only method that can be safely used, when we wish to discover what degree of value a thing has in itself."[40]

5. One may argue that an item is (or isn't) intrinsically good by appealing to facts about the concept through which we refer to it. For example, an item picked out by a disjunctive concept (and not by a simple concept) might be thought unfit to be intrinsically good.

6. One may argue that an item is (or isn't) intrinsically good by appealing to "formal features of rationality or the logic of key moral terms."[41] This sort of strategy is typically used to show what one ought to do,[42] but it could also be used to show the some item is intrinsically good (for example, a satisfied preference).

Railton says that "According to a well-developed tradition in the theory of value, internalism, it is essential to the concept of intrinsic goodness that nothing can be of intrinsic value unless it has a necessary connection to the grounds of action"[43] (where one reading of "grounds of action" is "motives"). So one may argue, on internalist grounds, that some item isn't intrinsically good because it needn't motivate.

7. One may argue either that an item is real (or natural) and therefore, probably, intrinsically good, or that an item is less real or (unnatural) and therefore, probably, intrinsically bad.[44] This sort of strategy may be also

used to try to show what one ought to do (for example, by arguing that certain behavior is, or isn't, natural for the human animal).

8. One may argue that an item is (or isn't) intrinsically good by suggesting that it is (or isn't) virtuous, correct or fitting to promote or adore for its own sake.[45]

9. Sometimes philosophers try to show that a belief of the form, "X is intrinsically good" has nonrational causes (typically psychological, sociological, historical or evolutionary). For example, Marx thinks that economic factors determine moral beliefs; Nietzsche blames "slave morality" on the socioeconomic condition of Jews and early Christians; Smart says that belief in some basic values may be due to "conceptual confusion," "rule worship" or "tradition;"[46] utilitarians in particular hold that some moral beliefs are due to classical conditioning: we believe that justice or honesty, say, are intrinsically good merely because they have been conjoined with the intrinsic good of utility.[47]

What might such etiological accounts show?

First, such accounts might combat certain arguments for intrinsic goodness. If the belief that an item is intrinsically good is claimed to be self-evident, that claim is undermined by showing that such confidence has a nonrational cause. Or, if a belief is advocated on the strength of its wide or universal appeal, that argument is undermined by showing that such appeal has nonrational sources.[48]

Second, etiological accounts help us evaluate arguments in a subtler way. Sher says that, "psychological explanations of desert-claims are of interest only *after* justification has failed."[49] However, psychological

explanations can help us assess *whether* justifying arguments succeed, for understanding our biases may help us evaluate arguments. For example, we might assess an argument for an item's having intrinsic value less favorably if we knew that nonrational factors predispose us to accept its conclusion.

Third, etiological accounts might aid us in arguing that we have no good reason to believe that an item is intrinsically valuable. If that conclusion can be established, then one can say, "Most items are not intrinsically good; so, probably, this item isn't intrinsically good (or, most items are intrinsically neutral; so, probably, this item is intrinsically neutral)."

Fourth, etiological accounts may show that belief in an item's intrinsic goodness has a source that usually produces false beliefs; therefore, that belief is probably false.[50]

"In my opinion," Nagel says, "someone who abandons or qualifies his basic method of moral reasoning on historical or anthropological grounds alone is nearly as irrational as someone who abandons a mathematical belief on other than mathematical grounds."[51] Etiological accounts alone can never rationally compel us to *abandon* the belief that an item is intrinsically good; however, they can count against such beliefs in the four ways just noted.

One might also try to show that a belief of the form "X is intrinsically good" has rational causes and so is probably true. For example, one might argue that the belief is the result of reliable moral intuition. However, no psychological, sociological, historical or evolutionary accounts provides evidence for such beliefs. This territory belongs to skeptics. Of course, my

belief might have a rational source if I formed it by reflecting on a good argument, but then we would assess the belief by assessing the argument.

10. One may argue that an item is (or isn't) intrinsically good by drawing an analogy between that item and some other item which is assumed to be (or not to be[52]) intrinsically good. These two items may have similar features (and therefore, each is good, bad or indifferent) or they may have analogous features (so that one is intrinsically good and the other intrinsically bad). I have already used this argument-form once: dirt has no intrinsic worth; anything physical is just rearranged dirt-parts; this supports the idea that nothing physical has intrinsic value. Or, to take another example, if intense pain is intrinsically bad because it hurts, then intense pleasure is intrinsically good because it's delightful.

11a. One may argue that an item is intrinsically good (or bad) by arguing that it is the source of other values (or disvalues).

11b. One may argue that some item must exist which is intrinsically good (or bad) since something in the world has value.

Beardsley considers the argument that "'Instrumentally valuable' is a relational concept—X borrows its value from Y, or Y confers its value upon X. If the value Y confers is itself instrumental, so that it is merely passed along from Z, then where does Z get its value? In the last analysis, something must (according to this argument) possess its value in itself, or nothing can get any value."[53] This argument parallels the First Cause argument for the existence of God.

11c. One may assume that Z has instrumental value and then argue that "Z has instrumental value" means "Z is conducive to something that has

intrinsic value." And then one can conclude that some particular item has intrinsic value because that would best explain Z's instrumental value. Beardsley calls this "the argument from definition."[54]

12. One can always support a thesis by knocking down arguments against it; defending a view helps support it.

So, when philosophers disagree about whether an item has intrinsic worth, discussion need not end. And several of these methods, I think, can advance *rational* discussion in the right context.

Which of the twelve argument-types might be used to help found ethics? Strategies 9-12 are non-starters: 12 merely tells us to respond to opposing arguments; to use 11 and 10, one must presuppose that some items are valuable; strategy 9 (about the etiology of moral beliefs) is a merely skeptical weapon. What about the others? I'll address them in descending order:

8. "One may argue that an item is (or isn't) intrinsically good by suggesting that it is (or isn't) virtuous, correct or fitting to promote or adore for its own sake." Such arguments seem unpromising; I can't think of any way to identify what is virtuous to adore for its own sake apart from trying to judge which items are intrinsically good.

7. "One may argue either that an item is real (or natural) and therefore, probably, intrinsically good, or that an item is less real or (unnatural) and therefore, probably, intrinsically bad." This type of argument seems to assume a suspect theology or teleology. If the universe wasn't created by

God, and if human nature evolved by natural selection, why should the natural tend to be good?

6. "One may argue that an item is (or isn't) intrinsically good by appealing to 'formal features of rationality or the logic of key moral terms.'" Although first-rate contemporary ethicists advocate this strategy, I side with their critics: for even if the ordinary notions of "ought" and "right" entail substantive ethical doctrines, evaluative notions that are substantively neutral (or that entail different ethical doctrines) might be better. "[D]efinitions," as Rawls says, "have no special status and stand or fall with the theory itself."[55]

5. "One may argue that an item is (or isn't) intrinsically good by appealing to facts about the concept through which we refer to it." But one can usually respond: "We shouldn't use *that* concept to refer to the item." For example, one can argue that someone is intrinsically good by appealing to facts about the concept "hero," but why apply that concept to that person?

4. "One may argue that an item is (or isn't) intrinsically good by appealing to perception or moral intuition." But saying, "I intuit X to be good" doesn't obviously amount to more than asserting one's belief that X is good.[56] Moreover, even if we intuit values, our intuition is fallible (or, if "intuition" is a success term, then we can't always tell if we're intuiting), for people often make unveridical perceptual value judgments: for example, when some people see two men kiss, they think it's wrong, disgusting, or unnatural. So, we would still need to know when people accurately (or really) intuit.

3a. "One may argue that an item is (or isn't) intrinsically good by appealing to how it is viewed by God, an ideal observer, a fully rational person or a competent judge." How can we put this method into practice—how can we

find out what all these authorities think? Most philosophers, with good reason, don't think God exists, while we can determine the attitudes of an ideal observer or a fully rational human only by theorizing about its psychology—yet if we do, we must support why we select just those psychological properties.[57] Most promising, perhaps, is to defer to competent judges, insofar as these can be identified; if intelligent, thoughtful people think that X is intrinsically good, then this provides some evidence that it is. However, smart, reasonable people differ on all matters of fundamental value. To resolve their disagreements, we need to assess their reasons. (3b, which appeals to actual human attitudes, is plausible only insofar as it appeals to the actual attitudes of competent judges.)

2a-d. "One may argue that an item is intrinsically good, bad or neutral by appealing to intrinsic facts about it." Perhaps some argument of this type succeeds. But arguments that focus on the monadic properties of physical items are vulnerable to Sidgwick's kind of objection, that it seems implausible to attribute to particles of organised matter, considered in themselves, either value or disvalue.

1. "One may argue that an item is intrinsically good by appealing to the merits of wider theories entailing it." We may reject inconsistent or internally incoherent theories and those that conflict with independently established results in science or philosophy. This might pare down the viable candidates for intrinsic value. But many theories won't be eliminated by this process: we won't be able to deduce, at the end of the day, that anything has intrinsic value. We can narrow down the pool further by eliminating theories that strike us as implausible—but calling a theory "implausible" doesn't offer evidence against it.

So far, the prospects for founding ethics look dim: some evidence supports nihilism (while two arguments I discussed don't count against it); skeptics stand ready to use strategy 9 (about nonrational etiology) against believing that a particular item has intrinsic value; and none of the twelve ways to argue that an item has value seem especially promising for founding ethics.

Korsgaard's Attempt to Found Ethics on the Value of Humanity

The foundation of ethics is the topic of Korsgaard's recent and widely-discussed book, *The sources of normativity*, so I will discuss her proposal at some length. Korsgaard's Kantian approach is foreign to me; I find it hard to summarize and difficult to interpret. So, I will quote her at length to help the reader judge whether my interpretations and criticisms are fair.

Korsgaard says that the problem of justifying ethics arises because "our capacity to turn our attention onto our own mental activities is also a capacity to distance ourselves from them and to call them into question."[58] And she says:

> If the problem springs from reflection then the solution must do so as well. If the problem is that our perceptions and desires might not withstand reflective scrutiny, then the solution is that they might. We need reasons because our impulses must be able to withstand reflective scrutiny. We have reasons if they do. The normative word 'reason' refers to a kind of reflective success.[59]

Now Korsgaard must explain what this reflective process is, and why impulses that survive it deserve the title of reasons. Assuming human free will, she begins by invoking Kant:

> [Kant] defines a free will as a rational causality that is effective without being determined by any alien cause. Anything outside of the will counts as an alien cause, including the desires and inclinations of the person. The free will must be entirely self-determining. Yet, because the will is a causality, it must act according to some law or other.[60]

Nagel challenges this last bit: "If the will is self-determining, why can't it determine itself in individual, disconnected choices as well as according to some consistent law or system of reasons? A neo-Humean regularity theory of causation seems an inappropriate model for free self-determination."[61] This seems right; but Korsgaard offers another argument for the idea that wills must act according to laws:

> Alternatively, we may say that since the will is practical reason, it cannot be conceived as acting and choosing for no reason. Since reasons are derived from principles, the free will must have a principle. But because the will is free, no law or principle can be imposed on it from outside. Kant concludes that the will must be autonomous: that is, it must have its *own* law or principle. And here again we arrive at the problem. For where is this law to come from? If it is imposed from outside, then the will is not free. So the will must adopt the law for itself. But until the will has a law or principle, there is nothing from which it can derive a reason. So how can it have any reason for adopting one law rather than another?
>
> Well, here is Kant's answer. The categorical imperative tells us to act only on a maxim that we could will to be a law. And *this*,

> according to Kant, *is* the law of a free will. . . . *All that it has to be is a law.*[62]

Now Korsgaard must support Kant's answer. She must at least address Mill's worry that Kant "fails . . . to show that there would be any contradiction, any logical (not to say physical) impossibility, in the adoption by all rational beings of the most outrageously immoral rules of conduct."[63]

Korsgaard continues, now extending Kant's ideas:

> The reflective structure of the mind is a source of "self-consciousness" because it forces us to have a conception of ourselves. . . . When you deliberate, it is as if there were something over and above all of your desires, something that is *you*, and that *chooses* which desire to act on. This means that the principle or law by which you determine your actions is one that you regard as being expressive of *yourself*.[64]

So, she says, you identify with the principles that move you. The concept of identity at work here is:

> a description under which you value yourself, a description under which you find your life to be worth living and your actions to be worth undertaking. So I will call this conception your practical identity. Practical identity is a complex matter and for the average person there will be a jumble of such conceptions. You are a human being, a woman or a man, an adherent of a certain religion, a member of an ethnic group, someone's friend, and so on. And all of these identities give rise to reasons and obligations. Your reasons express your identity, your nature; your obligations spring from what that identity forbids.

Korsgaard hasn't yet addressed Mill's concern. Why are we ethically—as opposed to psychologically—bound by our practical identities? Joseph Mengele might value himself under the description, "Nazi doctor," but no obligations flow from that identity.

Korsgaard continues:

> It is the conceptions of ourselves that are most important to us that give rise to unconditional obligations. For to violate them is to lose your integrity and so your identity, and to no longer be who you are. That is, it is no longer to be able to think of yourself under the description under which you value yourself and find your life worth living and your actions worth undertaking. It is to be for all practical purposes dead or worse than dead. When an action cannot be performed without loss of some fundamental part of one's identity, and an agent would rather be dead, then the obligation not to do it is unconditional and complete.

But, as Griffin says, "To ask a person whose life is centered on resentment or revenge or vanity or one-upmanship to 'abandon' himself may be exactly what he needs. Some people most need, and all of us to no small degree would benefit from, some well chosen 'disintegration' and 'reintegration'."[65]

Korsgaard admits that so far she hasn't settled the question of how one should conceive her practical identity,[66] but she does claim to have established that:

> The reflective structure of human consciousness requires that you identify yourself with some law or principle that will govern your choices. It requires you to be a law to yourself. And that is the source

> of normativity. So the argument shows just what Kant said that it did: that our autonomy is the source of our obligation.[67]

To be under an obligation, one must be autonomous; this is why infants and lions lack obligations. In *that* sense autonomy may be regarded as a source of normativity, but Korsgaard hasn't shown that obligations exist; autonomy may not suffice for normativity. Moreover, Korsgaard has not shown that autonomy is "the" source of normativity. Suppose, for example, that pleasant experiences alone are intrinsically good, and unpleasant experiences alone are intrinsically bad, so I should conceive my identify in whatever way would best promote utility. On these assumptions, sentience is also a source of normativity, for a world without it would be a world without obligation.

Now Korsgaard begins arguing for what our self-concept should be. She says,

> you must be governed by *some* conception of your practical identity. For unless you are committed to some conception of your practical identity, you will lose your grip on yourself as having any reason to do one thing rather than another—and with it, your grip on yourself as having any reason to live and act at all. But *this* reason for conforming to your particular practical identities is not a reason that *springs from* one of those particular practical identities. It is a reason that springs from your humanity itself, from your identity simply as a human being, a reflective animal who needs reasons to act and to live. And it is a reason you have only if you value your humanity as a practical, normative form of identity, that is, if you value yourself as a human being.[68]

And "valuing ourselves as human beings involves valuing others that way as well, and carries with it moral obligations." What obligations? If I value humanity as a "practical, normative form of identity," must I have a reason to dissuade my terminally ill friend from committing suicide? Must I have a reason to help a stranger learn how to play chess, if she values that as an end? I will try to assess Korsgaard's argument without fully understanding its conclusion.

In the above passage, Korsgaard warns: "unless you are committed to some conception of your practical identity, you will lose your grip on yourself as having any reason to live and act at all." Why must I not lose my sense of having reasons to act and live? If the "must" is psychological—so it is psychologically incumbent on me to have practical identities—then normativity is not established, only compulsion (and anyway, it couldn't then be shown that I must respect the humanity of others, for no psychological compulsion makes me respect my neighbor's humanity). However, Korsgaard seems to believe that one could have no practical identity, that one could "live at random, without integrity or principle. . .,"[69] so this interpretation may be unfaithful to her intent.

Alternatively, perhaps Korsgaard means that I must not lose my sense of having reasons because that would be bad or bad for me. This is true but inappropriate to assume in this context: establishing normativity requires not assuming explicitly normative theses. Anyway, does the fact that it would be bad for me to lose my will to live show that I should value myself as a reflective animal? The premise doesn't seem to support the conclusion, unless they have very similar meanings.

Korsgaard says that she is just offering a fancy version of the following argument which she attributes to Kant:

> Kant saw that we take things to be important because they are important to us—and he concluded that we must therefore take ourselves to be important. In this way, the value of humanity itself is implicit in every choice. If complete normative scepticism is to be avoided—if there is such a thing as a reason for action—then humanity, as the source of all reasons and values, must be valued for its own sake.[70]

This tightly-packed argument is more easily grasped as follows:

1. We take things to be important because they are important to us.
2. So, whenever we take something to be important, our reason for doing so entails that we are valuable.
3. So, if we ever take something to be important (that is important) for the correct reason, then we are valuable.
4. So, if complete normative scepticism is to be avoided, then humanity is valuable for its own sake.

This argument is too weak to found ethics, for humanity's value would only follow from the further premise that normative skepticism is false—a premise nihilists deny. Nihilists believe that nothing is valuable, so they won't allow that we ever take something to be important that is important. But the argument is worth pursuing further; perhaps the best we can do is a conditional proof: "if nihilism is false, then . . ."

Korsgaard's use of "we" makes the inferences from 2 to 3 and 3 to 4 seem less controversial than they are. In the context of her discussion, "We take things to be important because they are important to us" means "I take

things to be important because they are important to me" for each value of "I." But if so, all that follows at 3 is that if I ever take something to be important (that is important) for the correct reason, then I am important *to myself*—I have merely relative value. And then it does not follow that humanity is valuable if anything is—just that I am valuable to myself if anything is (for each value of "I"). But this thesis, if established, would be significant; so let's examine the argument's first two steps.

Premise 1, "We take things to be important because they are important to us," admits of three interpretations. (For the reader's convenience, I'll include the second premise, suitably revised, with each interpretation.)

1a. I take things to be important because I believe them to be important—so every time I value something, I express trust in myself and value my opinion.
2r. So, whenever I take something to be important, my reason for doing so entails that I am valuable to myself.

2r doesn't follow from 1a, for you can trust your opinion on a particular occasion without valuing yourself—you need only think you're right once.

1b. I take things to be important because I believe them to be good for me.
2r. So, whenever I take something to be important, my reason for doing so entails that I am valuable to myself.

But the reason why I take things to be important isn't always that I believe them to be good for me—that assumes some form of egoism. We might

accept a weakened form of 2r: "whenever I take something to be important because I believe it's good for me, that pattern of reasoning assumes that I am valuable." But nothing about the value of persons follows from this. Even if I take my pleasure, for instance, to be important because it's good for me, it may be good merely because of what it's like to be in that state.

Premise 1 admits of a third understanding that is not Korsgaard's but is worth exploring:[71]

1c. We take things to be important because they are important to people whom we regard as objectively valuable. Sometimes I take things to be important because they are important to me, but on other occasions I defer to the judgment of others.
2r. So, whenever I take something to be important, my reason for doing so entails that I or someone else is valuable.

But I may not always take something to be important because someone whom I value does: I may think, for example, that ending poverty is important just because I think that poverty is bad. 2r, anyway, won't entail, given 3, that "if complete normative scepticism is to be avoided, then humanity is valuable for its own sake." It would only entail that "if complete normative scepticism is to be avoided, then some people must be valuable for their own sake." Hence, I don't think Korsgaard's argument, on any interpretation, supports the thesis that humanity is valuable.

A few pages later, Korsgaard seems to suggest that one can intuit one's own value:

> Value, like freedom, is only directly accessible from within the standpoint of reflective consciousness. And I am now talking about it externally, for I am describing the nature of the consciousness that gives rise to the perception of value. From this external, third-person perspective, all we can say is that when we are in the first-person perspective we find ourselves to be valuable, rather than simply that we are valuable. There is nothing surprising in this. Trying to actually see the value of humanity from the third-person perspective is like trying to see the colours someone sees by cracking open his skull.[72]

Can I simply see, from the first-person perspective, that I have value? Such a claim, by Korsgaard's own lights, is "just an expression of confidence and nothing more."[73]

Prior to *The sources of normativity*, Korsgaard argues that:

> we regard some of our ends as good, even though they are obviously conditional; there must be a condition of their goodness, a source of their value; we regard them as good whenever they are chosen with full rational autonomy; so full rational autonomy itself is the source of their value.[74]

Acting with full rational autonomy is equivalent to acting with a good will; so, Korsgaard's conclusion is that the good will is intrinsically valuable. And she says that the good will alone has intrinsic value.[75] This contradicts her conclusion that humanity has value, for not all human being have good wills.

The above argument is of type 11a: "One may argue that an item is intrinsically good by arguing that it is the source of other values." But the argument, as it stands, is too quick and intuitive to succeed. People are

often mistaken in thinking their ends are good; and when those ends are good, perhaps they are intrinsically good or good as a means to something of intrinsic worth.

And so, Korsgaard's Kantian arguments fail to found ethics. She admits early on in *The sources of normativity* that she will have little to say about the content of our obligations;[76] I am inclined to think that her strategy is too formal to succeed.

Now I want to inquire whether the thesis that persons are intrinsically valuable, if established, *would* found an approach to normative ethics. I will examine an argument by Sher that motivates a position similar to Korsgaard's. According to Sher, persons are intrinsically valuable and confer value on whatever they desire. According to Korsgaard, the good will is intrinsically valuable and persons confer value on just those objects that they rationally desire.[77]

Sher argues as follows:

1. Persons are ends in themselves or have absolute value.
2. Persons taking "other things to have value is central to what makes them valuable."[78]

C. "Thus, it does seem natural to hold that a portion of their value devolves upon what they value—that some of the absolute value of persons is transferred to, or inherited by, the things they care about."

If Sher's argument succeeds, then ethical theory is off and running, for many things would be valuable.

Korsgaard rejects the thesis, entailed by the conclusion, that "value is conferred by desire," observing that

> there are desires that conflict with one's health or happiness or that are self-destructive or pathological or simply burdensome out of all proportion to any gratification their fulfillment can provide. This already shows that the existence of a desire is not by itself a sufficient reason for the realization of its object; further conditions exist.[79]

Korsgaard's argument is invalid; even if some desires are *insufficient* reason to realize their objects, all desires may provide *some* reason for doing so. On Korsgaard's view, value is conferred, but only by fully rational choice; "it is the reasoning that goes into the choice itself—the procedures of full justification—that determines the rationality of the choice and so certifies the goodness of the object."[80] I don't know what Korsgaard's "procedures of full justification" are, so I can't assess her proposal.

Sher's premises can be challenged, but I'll grant them. I'll just suggest that they don't support the conclusion.

What does the conclusion mean, according to which a portion of someone's value "devolves" or "transfers" to what she cares about? This mimics the theory according to which causes always transfer something to their effects. But to suppose that goodness could *transfer* in any non-metaphorical sense falsely reifies value. Anyway, to suppose that value literally transfers from person to object entails, absurdly, that the person's value decreases each time she desires something. (Similarly, if I inherit my father's fortune, the family wealth doesn't increase, although, on Sher's view, when I care about something, the world's value increases.) Saying that persons "confer" value, as Sher sometimes does, is better. Still, "conferring value" is metaphorical; what is literally meant? Perhaps only that an item's value is greater after being desired than before. But now how

do the premises support the conclusion? Even if persons acquire value, in part, by desiring things, why should those things—passively being desired—acquire value?

Sher, as we've seen, holds that a portion of our value devolves on what we value. And Korsgaard says that "goodness . . . flows into the world from the good will, and there would be none without it."[81] Each picture has the following drawback. Just as what we (rationally) care about seems to have value, what we (rationally) despise seems to have disvalue. These two intuitions are of a piece. But the source of a despised object's badness wouldn't be an intrinsically bad person or will—rather, an intrinsically good person or will would confer badness. And if "goodness conferring goodness" seems plausible, "goodness conferring badness" seems implausible.

So, I've found no reason to think that an approach to normative ethics could be founded on the intrinsic value of persons.

Founding Ethics on Sentient Experience

Suppose I believe on the basis of a newspaper article that Andy Kaufman has died. By showing you the article, I share my evidence for this belief. In this instance I can share my evidence because we have similar access to its bearer, the newspaper. Sometimes, however, I can't share my evidence with others. Suppose, for example, that I see an elderly woman driving a red sports car at twice the speed limit. Later I can tell you about it, but I can't share my evidence with you that it happened, for it's too late

for you to see it happening. You may, of course, regard my testimony itself as evidence, but then your evidence would differ from mine.

Ethics is founded on evidence that can't be shared. My experience of severe pain gives me reason to believe that nihilism is false. In others words, when I am in severe pain, that pain, as it's presented to me, gives me evidence that it's bad in some way. I can't share this evidence with you; you can't feel my pain. Even if you could peer inside my head and see it, you wouldn't be presented with it in a way that gave you evidence of its badness. But you, of course, are in the same position regarding your pain: when you are in severe pain, that pain, as it's presented to you, provides you with evidence that it's bad in some way. So, each of us has evidence for his or her severe pain being bad in some way. In the case of infants and nonhuman animals, the evidence is there, but the creature is too unsophisticated to recognize it as such.

Other philosophers have made similar points. According to James Rachels, "Suffering is so obviously an evil, just on account of what it is like, that argument would be superfluous; and the same goes for enjoyment as a good."[82] Our knowledge of what suffering is like, I emphasize, comes from experiencing it. And Nagel says,

> If I have a severe headache, the headache seems to me to be not merely unpleasant, but a bad thing. Not only do I dislike it, but I think I have a reason to try to get rid of it. It is barely conceivable that this might be an illusion, but if the idea of a bad thing makes sense at all, it need not be an illusion, and the true explanation of my impression may be the simplest one, namely that headaches are bad, and not just unwelcome to the people who have them.[83]

Passages like these merely express confidence, Korsgaard thinks.[84] Of course, having severe pain imbues me with confidence that the pain is bad, but this confidence, I believe, is appropriate. Whether this is whistling in the dark I'll let you decide, based on your experience of intense pain.

Earlier I discussed twelve types of argument bearing on whether an item has intrinsic value (or, with slight modifications, other kinds of objective value). Now I am suggesting that a severely painful experience gives its subject evidence for its being bad in some way. Which argument-type is this? It isn't an instance of "2. One may argue that an item is bad by appealing to intrinsic facts about it." I don't appeal to facts about severe pain; I appeal to pain itself. Do I "appeal to perception or moral intuition," as in 4? To say "yes" would be misleading. I *appeal* to severe pain and merely *assume* that, in particular instances, a subject can apprehend that such pain is bad. Moreover, nothing I say entails that a special faculty of intuition exists. Perhaps my argument is closest to instancing

> 3. One may argue that an item is bad by appealing to how it is viewed by a competent judge.

But 3 fails to capture the fact that my argument specifically addresses the subject of the pain; I want you to realize that your experience constitutes evidence for founding ethics. 3 can be modified to capture the argument: "3m: One may argue to person Z that Z's severe pain is bad by appealing to how Z views it," where Z is the only competent judge of *that* instance of pain.

Earlier I suggested, in so many words, that the modern scientific worldview supports nihilism. How might our larger scheme accommodate

the badness of suffering? I'll discuss two possibilities. On dualism, experiences are neither physical nor physically realized. Dualists have no problem acknowledging the disvalue of pain; anyway, my arguments for nihilism merely assume that dualism is false. But dualists are burdened to explain why experiences exist only in connection with certain types of complex physical states. Alternatively, if experiences are physical or physically realized, then my earlier arguments support nihilism. However, my experience lends *greater* support to my believing that my severe pain is bad.[85] Hence, nondualists should believe that a complete account of the physical world—that includes (realized) suffering—would entail that some experiences are bad, even if scientists need not characterize them as bad for their purposes. As Nagel says, "To assume that only what has to be included in the best causal theory of the world is real is to assume that there are no irreducibly normative truths."[86] Nondualists will have trouble explaining why the relevant physical processes are bad; but on my view that amounts to a problem they have anyway, of explaining why certain physical processes are or realize experiences having the (awful) qualitative character they do.

In sum, the evidence for severe pain being, in some way, objectively bad outweighs the evidence for nihilism. Mackie says, "if we were aware of [objective values], it would have to be by some special faculty of moral perception or intuition, utterly different from our ordinary ways of knowing everything else."[87] Were pains bad by virtue of participating in a Platonic form or having the non-natural property *bad*, perhaps we would need a special faculty to detect them as such. But we seem to need no such faculty to know that pains have disvalue because of how they feel. As Churchland says, "we do have an organ for understanding and recognizing moral facts.

It is called the brain."[88] Moral facts, as opposed to objectively valuable items, might be in a third realm, but if they are, then so are empirical facts.

My argument posits something like an instance of "immediate justification." Someone is immediately justified in a belief, says Alston, just in case she is justified by something other than its relation to her other justified beliefs.[89] But in a later paper, Alston argues persuasively that there is no unique item properly called *epistemic justification*. Rather, we should abandon justification-talk in favor of talk about epistemic desiderata, or "different ways in which beliefs can be better or worse from an epistemic point of view."[90] One such desideratum is "based on adequate grounds (reasons, evidence)." So, I assert an instance of immediate justification, substituting that desideratum for "justified." On my view, I have adequate evidence for believing that an instance of severe pain is bad by virtue of having that pain.

My argument has overcome nihilism. At least one value exists: severe pain. But this conclusion is too weak to found a systematic ethical project. So, now I will argue that all sentient experiences have normative significance. These arguments needn't outweigh the evidence for nihilism; for if our worldview must accommodate the badness of some experiences, then it can accommodate the badness or goodness of others without further cost.

All sentient experiences are either pleasures or unpleasures. Pleasures are pleasant; unpleasures are unpleasant. Severe pains are a subset of intense unpleasures. I'll argue that all unpleasures are bad, and all pleasures are good, using two types of argument.

First, I'll extend the argument about severe pain to all sentient experience. All of my pleasurable experiences provide me with evidence

that they're good; all of my unpleasurable experiences provide me with evidence that they're bad. This evidence is greater, or at least more obvious, for the more intense pleasures and the more intense unpleasures. But even having a mild pleasure gives one reason to think it's good in some way.

Second, I'll appeal to analogical arguments (strategy 10). All intense unpleasures, I've argued, are bad; each gives its subject evidence of its badness. If so, then (i) less unpleasant experiences should also be bad, though less bad; and (ii) intense pleasures should be good. Furthermore, (i) and (ii) each support the thesis that less intense pleasures are good; mild pleasures are similar to intense pleasures and mild unpleasures.

These arguments found the project of developing a theory of hedonic value.

Conclusion

To found ethics, one must show that the balance of evidence favors the existence of normative reasons. Also, or in the same breath, one must adequately support a project for developing an ethical theory. In this chapter, I tried to found ethics and to raise problems for other approaches.

Two arguments against nihilism proved unsuccessful. These arguments tried to show that normative reasons exist without entailing what any of them are. A successful argument against nihilism, it seems, must support particular values.

I discussed twelve types of argument bearing on whether an item has intrinsic value. Several of these strategies might be used in founding

ethics, but each seemed to have problems. Korsgaard's Kantian approach was inadequate in the final analysis.

On my view, a severely painful experience gives its subject evidence that it's bad in some way. This evidence outweighs our evidence for nihilism. Moreover, similar considerations, as well as analogies, support thinking that all unpleasures are bad in some way and all pleasures are good in some way.

But in what way? Let's use "a pleasure" to refer just to an experience. My arguments thus far leave open the following options:

(1) Pleasures—those experiences—might be extrinsically pleasant, just as fathers—those persons—are extrinsically fathers. If so, then pleasures might have contributive value; a pleasure conjoined with whatever its pleasantness consists in might be an intrinsically good organic whole.
(2) Pleasures might have merely relative value; only I would have basic reason to promote my pleasure. Or, more weakly, perhaps you have some basic reason to promote my pleasure, though I have more.
(3) Existing pleasures, like sacred values, might be basically good even though we don't always have reason to ensure that someone will feel pleasure. In particular, the fact that someone will feel pleasure is no reason to bring him or her into existence.

In chapters 2, 3 and 4, I argue against (1), (2) and (3) respectively. Pleasures, I will ultimately conclude, are intrinsically good, while unpleasures are intrinsically bad.

Chapter 2: Is Unpleasantness Intrinsic to Experience?

Unpleasant experiences include itches, backaches, phantoms pains and moments of embarrassment. What does their unpleasantness consist in? Philosophers have offered the following answers:

1. The unpleasantness of an experience consists in its representing bodily damage. (Damage)
2. The unpleasantness of an experience consists in its inclining the subject to fight its continuation. (Motivation)
3. The unpleasantness of an experience consists in the subject's disliking it. (Dislike)
4. The unpleasantness of an experience consists in features intrinsic to it. (Intrinsic Nature)

Each of these theories stands or falls with its corresponding view of pleasure. So, I will assess Motivation, for instance, alongside the idea that the pleasantness of an experience consists in its inclining the subject to fight for its continuation.

Why does this issue matter? First, if Intrinsic Nature is true, then unpleasantness doesn't consist in unpleasures playing any sort of functional role. Second, a correct account of unpleasantness should cast light on *why* unpleasures are bad. Third, if we knew why unpleasures are bad, perhaps we could conclude that other things are bad for the same reason. Consider the following argument:

(i) Unpleasures are non-instrumentally bad.

(ii) So, on Motivation or Dislike, some experiences are non-instrumentally bad because one is moved to end them or dislikes them.

(iii) To dislike or to be moved to end an experience is to desire that it end (even if one's overall desire is for the experience to continue).

(iv) So, on Motivation or Dislike, desires sometimes confer non-instrumental disvalue (or non-instrumental value, in the case of pleasures); or, on Motivation or Dislike, sometimes satisfying a desire (by ending unpleasure) is good just to have it satisfied. This supports:

(v) Desires confer disvalue or value in other instances; there is basic reason to satisfy desires in other instances.

I will challenge this argument by criticizing Motivation and Dislike.

Fourth, if Intrinsic Nature is false—if unpleasantness is extrinsic to unpleasures—then such experiences are not intrinsically bad. Several philosophers have made this point about painful experiences. Nelkin, who supports Damage, says, "Pains are *bad*, but no phenomenal state *in and of itself* wears that evaluation."[91] Korsgaard, championing Motivation, says that "someone who says he is in pain is not describing a condition which gives him a reason to change his condition. He is announcing that he has a *very* strong impulse to change his condition."[92] Parfit, advocating Dislike, says, "Some have claimed that pain is intrinsically bad, and that this is why we dislike it. As I have suggested, I doubt this claim."[93] And Hall says, "Why don't we like pain sensations? . . . Because they accompany nociceptual reports of bodily damage, and bodily damage is something we don't like to hear about. It is like the ruler who slew the messenger who

brought the bad news; pain sensations are no more inherently bad than the messenger."[94] These remarks flow from false theories of unpleasantness. The correct account—Intrinsic Nature—supports unpleasures being intrinsically bad.

Damage

According to Tye, "pains are sensory representations of bodily damage or disorder."[95] Pains are merely one species of unpleasure, but authors rarely distinguish painfulness from unpleasantness—perhaps some would say that unpleasures are sensory *or nonsensory* representations of bodily damage. But in what sense do unpleasures *represent* damage? There are several views in this area to consider.

(A) The unpleasantness of an experience consists in its being caused by bodily damage.

Pitcher holds that "to be aware of a pain is to perceive—in particular, to *feel*, by means of the stimulation of one's pain receptors and nerves—a part of one's body that is in a damaged, bruised, irritated, or pathological state, or that is in a state that is dangerously close to being one or more of these kinds of states."[96] This view fails even for painful unpleasures. Brains in vats can have painful experiences without being diseased, injured or in danger; cortical stimulation suffices for unpleasure. Moreover, "Gentle touch, vibration and other non-noxious stimuli can

trigger excruciating pain" in some patients.[97] Such patients are unhealthy, but their pain is caused by *non-noxious* stimuli. So, (A) fails: an unpleasant experience needn't be caused by bodily damage.

Variants of (A) also appear unpromising. For example, "the unpleasantness of an experience consists in its being of a phenomenological type whose instances are typically caused by bodily damage." But bodily damage never causes unpleasure in a world of vatted brains. Moreover, many unpleasant psychological states in the actual world—such as loneliness, anxiety, boredom and sorrow—are not typically caused by bodily damage.

(B) The unpleasantness of an experience consists in the subject's believing it to signal bodily damage.

But if I interpret my unpleasure as part of a physical or psychological process of healing, then I won't believe it to signal bodily damage. Similarly, if I know that my diet contains too much fat, I won't believe that the delicious taste of chocolate signals my good health.

Nelkin says, "Pains consist entirely of a phenomenal state and the simultaneous, spontaneous appraisal of that state as representing a harm to the body."[98] So, I may *spontaneously* take my experience to signal bodily harm, even if I believe it to signal healing, all things considered. How can infants, rats, cats and bats feel pain despite lacking the concept of bodily harm? Nelkin says that the generalized form of all pain evaluation is, "The state here represented is harmful!"[99] Why should we believe that cats make such assessments? Nelkin points to instances in which evaluations seem to affect whether the subject feels pain.[100] But evaluations may affect whether

an experience is painful or unpleasant by subtly altering the experience itself (Intrinsic Nature), by moving the agent to fight its continuation (Motivation), or by causing one to dislike it (Dislike). Also, Nelkin emphasizes that on his view, pains share no characteristic qualitative feature. But Motivation, Dislike and Intrinsic Nature (as I develop it) don't imply otherwise. Hence, Nelkin's evaluative theory is undermotivated.

Extending Nelkin's view to cover all unpleasures weakens it further. Consider: "the unpleasantness of an experience consists in the subject's spontaneously evaluating it as signalling bodily harm." Many nonpainful unpleasures don't seem to signal bodily harm in any way. Mouthwash marketers, for instance, make their product taste bad because consumers associate that taste with bodily benefit (though with "harm" to bacteria); sorrows and anxieties often accompany worries about others only; and, with loneliness, what seems wrong is that one lacks company, not that one is in harm's way. These intuitive remarks are inconclusive; perhaps loneliness, sorrow, anxiety and the yucky taste of mouthwash are spontaneously evaluated as representing harm, even if we're not conscious of the evaluation as such. But such a view needs support.

(C) The unpleasantness of an experience consists in its intrinsically representing bodily damage. (This is also a variant of Intrinsic Nature.)

How might an experience intrinsically represent something? (i) The mental image of a square might intrinsically represent square things by robustly resembling them. But unpleasures, as far as I can tell, do not robustly resemble bodily damage. (ii) The mental image of a square might intrinsically represent square things by being itself an instance of

squareness. However, no one is tempted to think that each unpleasant experience is or includes an instance of bodily damage. So, (C) seems untenable.

Motivation

Korsgaard says, "The painfulness of pain consists in the fact that these are sensations which we are inclined to fight."[101] But that is not exactly right; a priest may be strongly inclined to fight his licentious *pleasures* because he views them as sinful. On the best form of Motivation, unpleasantness consists in only that inclination which the experience causes. As Brandt says, "for an experience to be pleasant is for it to make the person want its continuation," where *wanting* is understood in terms of action-tendencies.[102]

Unpleasures typically incline us to fight against them; moreover, Motivation elegantly accounts for unpleasure intensity. An unpleasure's intensity, on this view, is the degree to which it inclines the subject to end it. So, an intense unpleasure highly inclines one to end it, a mild unpleasure slightly inclines, and so on.

Even if unpleasure always inclines one to fight its continuation in proportion to its unpleasantness, one might not so fight, for any of four reasons. First, other motives might partly, or wholly, mask the expression of that inclination. Nagel says that someone may pursue pain "as a means to some end or . . . backed up by dark reasons like guilt or sexual masochism."[103] If such motives do not wholly mask the tendency to fight,

then the experience would limit the vigor of the pursuit. Second, one might not avoid unpleasant stimuli because one can't do so for nonpsychological reasons: one might itch but not scratch because one is pinned down in the wrestling ring. Similarly, one's capacity to fight might be diminished by such causes: one might be too exhausted from swimming to scratch hard. Third, other motives might cause one to fight unpleasure in excess of its intensity: one might work harder to end a headache because the wedding is in an hour. Fourth, one might fight disproportionately hard because of nonpsychological causes external to the experience: one might fight against unpleasure more vigorously if one just drank a pot of coffee. In all these kinds of case, Motivation entails counterfactuals: if one didn't feel guilty (or weren't pinned to the mat), then one would fight the unpleasure more vigorously; if one weren't concerned about the wedding (or weren't on a caffeine rush), then one would fight less vigorously.

Objection: "People want their unpleasure to end because it's unpleasant. However, on Motivation, this amounts to saying that people want their unpleasure to end because they want it to end, which is nonsense." Motivationists can reply: "Initially, Agent doesn't want her unpleasure to end because it's unpleasant. She tastes quinine, it seems, no sooner than she wants to spit it out; so the unpleasantness of the taste doesn't seem to cause her urge to spit. However, Agent's urge, once established, may cause further urges; and so, the unpleasantness of the experience may incline her to end it."[104]

Sidgwick criticizes Motivation in a forgotten passage. He considers Bain's view that "pleasure and pain, in the actual or real experience, are to be held as identical with motive power."[105] And Sidgwick objects that "some feelings which stimulate strongly to their own removal are either not

painful at all or only slightly painful:—e.g. ordinarily the sensation of being tickled."[106] Call this objection (i). Moreover:

(ii) The soothing pleasure of a massage can be highly intense yet relaxing, so the subject is not strongly moved to prolong it. If the masseuse pauses to rest, her client might say, "Please don't stop" but won't protest in any other way. On such occasions, the pleasure seems to motivate the subject less than its intensity would require.

(iii) Some depressives have no impulse or only a slight impulse to change their condition, perhaps because they cannot imagine feeling happy.[107] "Depression," says comedian Steven Wright, "is merely anger *without enthusiasm*."

(iv) Severe embarrassment can be motivationally crippling—it can cause the agent to "freeze up." When this happens, the unpleasure seems greater than any tendency the agent has to fight its continuation. Similar remarks hold for anxiety.

(v) Intensely unpleasant physical pains are usually motivationally crippling.

In (i), degree of motivation exceeds degree of unpleasantness; in (ii)-(v), it falls short. Motivationists can say that in (i) the extra motivation is caused by something other than the experience, and that in (ii)-(v) the motive power is masked. Empirical findings might vindicate Motivation, but lacking strong reasons to accept that view, this seems unlikely.

Dislike

Hall says that, "The unpleasantness of pain sensations consists in their being disliked."[108] Broad entertains Dislike when he wonders whether "This experience of mine is pleasant" means "I like this experience for its non-hedonic qualities."[109] On this view, unpleasures are conscious episodes that are disliked when experienced, and the intensity of an unpleasure is how much the subject dislikes it.

Parfit seems to vacillate between Dislike and Motivation. He says:

> On the use of 'pain' which has rational and moral significance, all pains are when experienced unwanted, and a pain is worse or greater the more it is unwanted. Similarly, all pleasures are when experienced wanted, and they are better or greater the more they are wanted.[110]

Wanting is more akin to motivation than to liking. But Parfit also says, ". . . the badness of a pain consists in its being disliked . . .,"[111] which implies that, on the use of 'pain' which has rational and moral significance, all pains are when experienced disliked.

Dislike and Motivation are easily conflated because one almost always wants to avoid experiences one dislikes and dislikes experiences one wants to avoid. To dislike an experience to some degree, however, does not entail being moved to fight its continuation to that degree; one can dislike being depressed but be resigned to it.

In what sense must I dislike my unpleasure? Unpleasures can be liked in some ways—for example, I can approve of my pain as part of a

healing process or as a means to benefiting others. Perhaps, on Dislike, one must have an unfavorable emotional attitude towards the experience "considered merely as feeling" (to use Sidgwick's phrase). The priest disapproves of his sexual pleasure, not merely as feeling (in that light it's delicious) but as a sin against God. Such an attitude needn't be cognitively sophisticated, given that kittens feel unpleasure.

Yesterday I felt pleasure, to different degrees, all day long. In what sense did I have a "favorable emotional attitude" to my experiences when I was not focusing on them? The notions of liking and disliking need further explication. One might say: "You felt pleasure all day long in the sense that, had you reflected on your experience, at any time during the day, you would have liked it. And how intense your pleasure is equals how much you would like it on reflection." But doesn't focusing on one's experiences change them? And sometimes one seems to like an experience more when one focuses on it—typically, the taste of wine—and sometimes less—typically, sexual pleasures. Why should pleasure intensity be the intensity of liking were one to focus, rather than actual liking?

Normally, one distinguishes unpleasant from other experiences based on whether one dislikes them merely as feeling, and one gauges unpleasure intensity based on the strength of such disliking.[112] However, this doesn't show that one must dislike unpleasure, for we often categorize phenomena based on features accidental to the category. For example, we normally categorize water as *water* based on features that pick out XYZ on Twin Earth, which isn't water. Perhaps we identify unpleasures based on our disliking them as feeling because such disliking typically includes the judgment that the experience is intrinsically unpleasant.

I'll offer counterexamples to Dislike in which the subject's emotional attitudes are skewed.

First, how much one likes an experience might be influenced unduly by how it contrasts with a prior state. For example, suppose you are in ecstasy, but then your pleasure plummets to a level only mildly pleasant. Of course, you will dislike no longer being in ecstasy, but mightn't you also dislike the mildly pleasurable state—resent it, as it were, for being so mild? And Trigg says, "The way in which people apparently enjoy searching with their tongue for a tooth which aches slightly, and deliberately trying to manipulate it, might confirm that it is quite intelligible to like pain."[113] People like such unpleasures because they contrast with the normal experience of touching a tooth with a tongue. People like them less as their novelty wears off.

Second, how much one likes an experience might be influenced unduly by one's prior expectations. Suppose you are blindfolded and told to expect the touch of a hot poker on the small of your back. As you wait, your anxiety rises. But instead of a hot poker, your back is touched with ice. You cry out, until you feel the object start to melt. Didn't you greatly dislike the experience, even though it was not intensely unpleasant?[114]

Third, someone's desire for attention can skew her attitudes. For example, a child wanting pity might throw a tearful fit over a minor injury. In some cases, it seems, she is not just acting—she dislikes her conscious state out of proportion to its unpleasantness.

Dislikers might respond as follows: "In these examples, the subject's disliking doesn't correspond to intensity because of causes foreign to the experience: a prior contrasting state, heightened anxiety, a strong desire for attention. But the unpleasantness of an experience consists in *its*

causing the subject's dislike." This would revise Dislike to mirror Motivation. Paul Churchland endorses such a view in passing: "[pain] qualia are similar in causing a reaction of dislike in the victim . . ."[115] But causes external to an experience almost always influence how much one dislikes it. Such causes might include noncontrasting prior states, moderate levels of anxiety and modest desires for attention. So, on this view, pleasure intensity will almost never be supposed to correspond to the subject's *actual* degree of liking. If so, then Dislike is hard to test. Such a view might be true, but why believe it?

Intrinsic Nature

On this view, certain experiences are intrinsically or nonrelationally unpleasant.[116] Unpleasantness, on this view, supervenes on qualia: there cannot be a change in unpleasure intensity without a change in qualia. Also, unpleasantness does not reduce to motivation or disliking or bodily damage relating in the right way to experience. But Intrinsic Nature leaves open whether unpleasantness is irreducible. For example, unpleasantness might reduce to (or supervene on) a physical property intrinsic to all unpleasures.

Hall assumes that if pains are intrinsically unpleasant, then they're necessarily unpleasant,[117] but this should not be assumed. If a pain is intrinsically unpleasant, then its perfect duplicates are unpleasant, but it might not be identical with only its perfect duplicates. In another world, the pain (or its counterpart) might be similar to it in this world without

being unpleasant. If so, the actual pain would be intrinsically, but not necessarily, unpleasant.

Two Reasons For Intrinsic Nature

1. When you twist your ankle or jam your finger, the experience itself seems to hurt; the unpleasantness seems to be right there in it. Don't just think about this abstractly; examine your unpleasures. Introspection, though fallible, provides evidence for Intrinsic Nature.

2. Why do some physical states, but not others, cause, constitute or realize qualitative experiences? This problem arises on any theory of unpleasantness. On Intrinsic Nature, one aspect of the problem is to explain why certain physical states cause, constitute or realize qualitatively unpleasant experiences.

However, the other views we've considered face an additional problem: why does conjoining an intrinsically neutral experience with that extra element produce an awful, normatively significant state of affairs? Consider Dislike. Why should having an unfavorable emotional attitude towards an otherwise neutral experience be awful? Why does disliking an otherwise neutral experience have normative significance? Dislikers might say: "I don't know, but this is not just my problem. Everyone believes that liking and disliking confer significance. For example, baseball cards matter because people like having them." But not everyone believes that liking and disliking confer significance. My having an

intrinsically worthless baseball card is good, I think, because it gives me intrinsically pleasant experiences—not because I like having it.

Objections to Intrinsic Nature

1. According to the higher-order perception theory (HOP), a mental state is conscious just in case it is introspected. According to the higher-order thought theory (HOT), a mental state is conscious just in case it is accompanied by the thought that one is in that state. Proponents of HOP or HOT might object to Intrinsic Nature as follows:

(i) HOP or HOT is true.
(ii) So, mental states are not intrinsically conscious.
(iii) So, some perfect duplicates of unpleasures are not conscious.
(iv) Such duplicates are not unpleasant.
(v) So, unpleasant mental states are not intrinsically unpleasant.
(vi) So, Intrinsic Nature is false.

I won't assess HOP and HOT[118] or (iv); I'll bypass these big issues by amplifying Intrinsic Nature so that it's compatible with them. Intrinsic Nature holds unpleasantness to be intrinsic to experience—that is, intrinsic to the minimum unit required for conscious experience. Hence, an experience is either (a) a mental state; (b) a mental state conjoined with one's introspecting it; or (c) a mental state conjoined with one's thinking that one is in it. Intrinsic Nature is therefore compatible with (v): even if

unpleasant *experiences* are intrinsically unpleasant, unpleasant *mental states* may not be. If both of these are true, then functionalism about unpleasantness would be true in one sense and false in another. True, if functionalism is the view that token mental states are type-identified by their relational properties; false, if functionalism holds that token conscious experience are type-identified by their relational properties.

2. What integrates the category *unpleasant experience*? If unpleasures are *intrinsically* unpleasant, then an intrinsic property, it seems, should unite that category. Do unpleasures share a distinctive, intrinsic feature? I have rejected the idea that unpleasures intrinsically represent bodily harm. Broad thought that, ". . . there is a quality, which we cannot define but are perfectly acquainted with, which may be called 'Hedonic Tone.' It has two determinate forms of Pleasantness and Unpleasantness."[119] Some writers agree with Broad,[120] though most disagree.[121] Broad's claim that we are perfectly acquainted with, but cannot define, "Hedonic Tone" reminds me of Hume's insistence that beliefs share a common quality of vivacity, even though "'tis impossible to explain perfectly this feeling or manner of conception."[122] Unpleasures, I think, are too varied for any common feel to distinguish them from all other experiences. And so Korsgaard says,

> If the painfulness of pain rested in the character of the sensations . . . our belief that physical pain has something in common with grief, rage and disappointment would be inexplicable. For that matter, what physical pains have in common with each other would be inexplicable, for the sensations are of many different kinds. What

> do nausea, migraine, menstrual cramps, pinpricks and pinches have in common, that makes us call them all pains?[123]

Nausea, migraines and menstrual cramps have no common qualitative feel distinctive of unpleasure, but if unpleasures are physical (or physically realized), then they (or their realizations) might share an intrinsic physical property. That property might be experienced differently depending on what other physical properties contribute to the conscious state; thus unpleasures could have no characteristic *qualitative* aspects, even though they had a characteristic physical aspect. Such a physical property could explain why we call unpleasures "unpleasant"—we do so when we discern that property's presence.

Positing such a property, however, might be considered empirical wishful thinking. Intrinsic Nature lovers would do better to explain unpleasure's unity in any of the following ways:

(A) Unpleasures are just those experiences that are intrinsically bad because of how they feel.
(B) Unpleasures are just those experiences that are bad for the people who have them because of how they feel to those people.
(C) Unpleasures are just those experiences that one ought to dislike merely as feeling; disliking is an appropriate response to unpleasures alone considered merely as feeling.

On these views, "unpleasure" is evaluative. Variants of this idea dot the literature. Nelkin says, "When one ascribes 'pain' to one's self, one is not merely describing a condition of oneself. One is also evaluating that

condition."[124] And utilitarians have traditionally understood "pleasure" and "pain" normatively. Sidgwick, for example, defines "Pleasure—when we are considering its 'strict value' for purposes of quantitative comparison—as a feeling which, when experienced by intelligent beings, is at least implicitly apprehended as desirable or—in cases of comparison—preferable."[125]

3. Two experiences may differ in pleasantness but seem intrinsically indistinguishable (or only unimportantly different). Some of these cases involve medical intervention. Parfit says that:

> After taking certain kinds of drug, people claim that the quality of their sensations has not altered, but they no longer dislike these sensations. We would regard such drugs as effective analgesics. This suggests that the badness of a pain consists in its being disliked, and that it is not disliked because it is bad.[126]

Brandt and Hall dispute Intrinsic Nature on similar grounds.[127] Let's distinguish two varieties of this kind of case. When drugs relieve pain (as in Parfit's example), the subject says that the experience has not altered, but she no longer minds it. When drugs prevent unpleasure from occurring, the subject says that the resulting experience is intrinsically just like pains normally are. Both varieties are potential counterexamples to Intrinsic Nature. In each case, the subject says that a current experience that doesn't hurt is intrinsically indistinguishable from a past experience that did. The subject's report is more credible when pain is relieved, for then the intrinsically indistinguishable experience is one the subject just had.

Brandt and Parfit do not cite any studies, interviews or personal experiences. Hall quotes a number of sources,[128] but in them subjects are not probed for the right information. All parties agree that analgesics have psychoactive properties. Demerol, for example, is psychoactively identical with heroin. So the question is not whether the subject's conscious state has qualitatively altered—obviously it has. Rather, the question is whether those changes account for why the state is no longer unpleasant. One cannot answer such a question based on a simple inspection of the introspective data. Under *favorable* circumstances, ordinary subjects might get these matters wrong. The circumstances, moreover, are unfavorable because the subjects are high on drugs. Hall believes that the altered state of the subjects does not undermine the credibility of their reports. "Is it really likely that they could actually be in an unpleasant or even awful mental state and not notice it or be able to report it correctly?"[129] But the danger is that the subjects' experience is *not* unpleasant, but they do not discern why.

Furthermore, subjects may be asked, "Though your pain has been relieved, is it still there?" with "yes" being taken as evidence against Intrinsic Nature. But the ordinary concept of "pain" might naturally extend to pleasant or neutral experiences that have much—but not everything—qualitatively in common with paradigmatic experiences of pain, and the qualitative differences might account for the hedonic difference. Similarly, the ordinary concept of "same sensation" might naturally apply to the sensation I have now (that is not unpleasant), and the unpleasant sensation I had two minutes ago, if they have much, but not everything, in common qualitatively.

After taking Demerol for pain relief, my father thought that the drug changed the sharpness of his pain to something duller that didn't hurt.[130] "Before anaesthetics proper came into use," says Hare, "surgeons used to give their patients whisky before operations; as anybody may verify, this does not diminish substantially the intensity of the pain-sensation, but may make it a great deal easier to bear."[131] My experience is that alcohol makes pains *intrinsically* duller, which might make them less unpleasant.

Trigg says, "We talk of 'acquiring a taste for something' and this just means coming to like a taste which we previously did not care for. It is a reasonable assumption that the taste has not changed in such instances."[132] If so, then whether a taste experience is pleasant or unpleasant depends on whether it is liked, not its internal nature. But it is also reasonable to suppose that the taste has changed. As I child, I despised crunchy peanut butter; now I like it, and I think my peanut butter experience itself has changed.

And in support of Motivation, Korsgaard says:

> Pain really is less horrible if you can curb your inclination to fight it. This is why it helps, in dealing with pain, to take a tranquilizer or to lie down. Ask yourself how, if the painfulness of pain rested just in the character of the sensations, it could help to lie down? The sensations do not change.[133]

Do they not? When relaxing soothes pain, doesn't pain's quality change? Might not a Stoic attitude affect the intrinsic properties of experience?

4. Assume the Humean thesis that distinct existences needn't be related—anything can be conjoined, or not conjoined, with anything else.

On Intrinsic Nature, unpleasures are conceptually and ontologically distinct from emotional or behavioral responses to them. I assume that, if this is true of unpleasure, then it is true of pain. And so, on Intrinsic Nature, severe pain can be conjoined with liking or pursuit. It is a further question whether human beings, given the actual laws, can like and pursue severe pains. But seemingly we can; we can leap tall buildings in a single bound, given the occurrence of multiple improbable quantum events. The fourth objection to Intrinsic Nature holds that no creature—and certainly no human being—can like severe pain, and severe pain can never cause a creature to pursue its continuation. (Motivation is vulnerable to the first half of this objection, and Dislike to the second.)

I won't appeal to masochism as a counterexample to this objection, for the familiar claim that masochists like and pursue pain might be true but not in the sense that masochists have a favorable emotional attitude towards severe pain itself, considered merely as feeling, and seek its continuation for that reason. Masochists might like pain but not severe pain; masochists might dislike pain but like the pleasure that accompanies it; masochists might seek stimuli that cause most people pain but cause them pleasure; masochists might dislike how pain feels but like its gratification of their self-loathing.

Should Intrinsic Nature lovers reject the Humean thesis? Sprigge suggests that we "turn away from that denial of genuinely intelligible necessities stemming from David Hume . . ." and "not be afraid of the idea that pleasures and pains are of their very nature liable to affect behavior in certain directions."[134] But, first, Sprigge's thesis is vulnerable to some of my objections to Motivation: if pleasures and pains are liable to affect behavior in certain directions, then why don't massages and severe

physical pains move us more than they do? Second, one can visually imagine and propositionally describe (in some detail) creatures pursuing all types of pain; this tends to show that creatures can pursue all types of pain.[135] Third, what evidence tells against creatures pursuing pain? Sprigge says that Intrinsic Nature lovers siding with Hume must "pretend that there would be nothing intrinsically odd to counter-hedonistically guided behavior."[136] But pursuing pain, on such a view, may be intrinsically odd by being so obviously imprudent.

Can creatures like severe pain merely as feeling? Consider this argument: "to like a token severe pain merely as feeling, one must not fully understand it; one must fully understand one's current experience; therefore, one cannot like severe pain." This argument has intuitive force, but its second premise is false. Consider a more humble argument: "to like a token severe pain merely as feeling entails robustly misunderstanding it, which is impossible. So, one cannot like severe pain." But why can't there be robust self-misunderstanding? And why couldn't some strange creature know that its pain was awful without disliking it?

Intrinsic Nature lovers should not be faulted for believing that some creatures can like and pursue severe pain. "It is a contingent fact that [pleasures] do cause and sustain desire and do shape behavior."[137] By the same token, Motivationists should not be faulted for thinking that creatures can like severe pain; nor should Dislikers be faulted for believing that creatures can seek to prolong severe pain.

5. The fifth objection goes as follows. "On the view that combines the Humean thesis with Intrinsic Nature, any psychology that includes one element from each of 1-4 is possible:

1a. Creatures generally avoid unpleasure.

1b. Creatures generally pursue unpleasure.

1c. Creatures exhibit no general motivational pattern towards unpleasure.

2a. Creatures generally pursue pleasure.

2b. Creatures generally avoid pleasure.

2c. Creatures exhibit no general motivational pattern towards pleasure.

3a. Creatures generally dislike unpleasure.

3b. Creatures generally like unpleasure.

3c. Creatures exhibit no general emotional pattern towards unpleasure.

4a. Creatures generally like pleasure.

4b. Creatures generally dislike pleasure.

4c. Creatures exhibit no general emotional pattern towards pleasure.

So, 81 distinct psychologies are possible (assuming that liking and disliking are distinct from behavioral tendencies). If so, the human psychology (1a, 2a, 3a, 4a) is improbable—a 'gross, empirical accident,' to use Findlay's phrase.[138] But the human psychology is uniquely rational among the 81 possibilities; surely we can explain its existence rather than regard it as a fluke; we can do better than to say 'perhaps it is the greatest luck of all that for some creatures that which is intrinsically motivating is also intrinsically valuable.'[139] So, the Humean version of Intrinsic Nature is false."

Similar objections apply to combining the Humean thesis with Motivation or Dislike. For example: "On Dislike, any psychology that includes one element from each of 1 and 2 is possible. If so, the human psychology (1a, 2a) is improbable. But that psychology is uniquely rational

among the nine possibilities; surely we can explain its existence rather than regard it as a fluke. Therefore, the Humean version of Dislike is false."

These objections are similar to the following argument from design: "According to atheist metaphysics, vastly most possible worlds do not include intelligent life. But surely we can explain the existence of intelligent life rather than regard it as an improbable fluke. Therefore, atheist metaphysics is false."

The fifth objection, like the argument from design, is hard to assess. How might Intrinsic Nature lovers respond to it? Appeals to natural selection are worthless, for at least the following reason: fitness-producing activities such as eating and mating might cause unpleasure in some worlds, and so evolution would favor unpleasure-seekers. Instead, Intrinsic Nature lovers might say that sentient species must have the human psychology, for any of three reasons. First, they might hold that pleasure, by its nature, motivates and causes liking no matter what other causal regularities a world exhibits (and similarly for unpleasure). Second, they might adopt the more general view that the laws and initial conditions of this world are metaphysically necessary, and so the human psychology is no fluke. Third, they might say that God must exist, and human psychology is rational because God is wise and good. These responses entails that the Humean thesis is false: unpleasure can't be conjoined with liking and pursuit. None of them, however, seem well-motivated.

But two better responses are compatible with the Humean thesis:

(1) One might respond to the fifth objection and to the argument from design by rejecting *a priori* probabilities: "Although most possible worlds do not include intelligent life, it is not *improbable* that the actual world does. And although most possible psychologies aren't rational, it is no *fluke* that ours is. Probabilistic notions presuppose regularities and so operate within worlds; they do not apply to worlds as a whole."[140] I won't discuss the difficult issues that this response raises.

(2) Suppose the universe includes a great many sentient life-forms, and that each of the 81 psychologies occurs with roughly equal frequency. Then the fact that the *human* psychology is rational would need no explanation; we could say: "It was likely that *some* life-forms would have the rational psychology; we're just lucky that ours does." Even if all terrestrial psychologies are of the (1a, 2a, 3a, 4a) variety, there are three non-exclusive ways in which the actual universe might have a roughly equal distribution of the 81 psychologies: first, a great many sentient life forms of varying psychologies might exist on distant planets; second, our cosmos might be one of many actual cosmoi (causally isolated worlds with different laws) that contain sentient beings of various psychologies; third, our cosmos might have cyclically expanded and contracted in the past, with various creatures existing during different periods of expansion and contraction. So, Intrinsic Nature lovers can say, "We have no reason to think that the universe doesn't have a roughly equal distribution of the 81 psychologies; so, we have no reason to think that the rationality of the human psychology needs explaining."[141]

Intrinsic Nature lovers can respond to the fifth objection by affirming that either (1) or (2) is true.

Conclusion

Damage, Motivation and Dislike face worrisome difficulties, while Intrinsic Nature rebutted five objections. Moreover, two additional considerations support Intrinsic Nature. So, our evidence now seems to favor Intrinsic Nature. As empirical psychology develops, this might change.

I haven't assessed all possible competitors to Intrinsic Nature. The ones I've discussed offer a single condition as necessary and sufficient for an experience to be unpleasant. A more complex functionalist view would hold that a token of unpleasantness can consist in multiple relational elements, and perhaps that unpleasantness can be realized by different, single elements in different instances. No such view, to my knowledge, has been explored,[142] but here are two brief challenges. First, can multi-factor functionalisms preserve the integrity of "unpleasure," or must that category seem as artificial as "itchy and anxious experience" or "itchy experience or anxious experience?" Second, can such views offer a natural way of determining when one experience is more unpleasant than another—of weighting contributors to unpleasure intensity?

If Intrinsic Nature is true, then "unpleasant" admits of no functional analysis. Moreover, on Intrinsic Nature, the best way to explain the unity of unpleasures is that unpleasures are just those experiences that are bad

in some way because of how they feel. This suggests the following transcendental argument:

1. Unpleasant experiences exist.
C. Some experiences are bad.

Some philosophers, however, may reject the concept of "unpleasant experience" as normatively loaded. On their view, "unpleasant experiences" don't exist, although experiences exist that are mislabeled as unpleasant.

Goldstein says, "It is by feeling the way it does, i.e. awful and bad, that pain justifies our aversion to it. Similarly, our justification for desiring pleasure and calling pleasure 'desirable' and 'good' lies in the intrinsic quality of the experience."[143] If unpleasures are bad because of how they feel, as I believe, then nothing else is bad for the same reason.

Chapter 3: Is Hedonic Value Agent-Neutral?

The Spectrum

Do I have basic reason to promote my pleasure, but not yours? In other words, does pleasure have agent-relative value? If so, then pleasures are not intrinsically good but are good merely from one person's perspective. Utilitarians have often disputed the more general thesis that I always have a reason to promote only my *happiness*. But have utilitarian arguments been successful? Mill, it is widely thought, failed to prove that the general happiness is a good to all persons, given that each person's happiness is good for him, and Gauthier says that "a hundred years of ever more sophisticated efforts to avoid Mill's fallacy have not advanced the cause of utilitarianism a single centimetre."[144] Sidgwick thinks that the conflict between egoism and utilitarianism, "the profoundest problem of Ethics,"[145] cannot be resolved by argument, and contemporary utilitarians such as Singer and Parfit seem to agree.[146]

Philosophers such as Bennett and Nagel believe that I have *some* basic reason to promote the pleasure or welfare of others, although I have more basic reason to promote mine. Bennett, for example, says that "each person is morally entitled to give some special weight to his own wants and needs and interests, just *qua* his."[147] If this special weight is light, call the view *Modified Benevolence*; if it is heavy, call it *Modified Egoism*. Call the range of views from light to heavy *the Spectrum of Weighted Self-Interest* or

the Spectrum for short. Egoism is the spectral point where the weight is heaviest. If one is not entitled to give one's interests any special weight, then well-being has agent-neutral value.

Rational Behavior

Promoting one's own interests over others' may seem rational, if not always moral. But calling behavior "rational" seems to praise it; hence, if it's rational to give one's welfare special weight, then one's pleasure, it seems, has more value to one than to others. *Is* self-interested behavior especially or uniquely rational? To answer this question, we need to know what "rational" means in this context. To this end, I'll identify the three ways in which philosophers commonly apply the term to behavior.

Bayesianism

According to Pettit and Smith, Bayesian decision theory has been the orthodox account of rationality for the last two hundred years.[148] Bayesians believe, roughly, that the rational thing for me to do is to maximize expected utility, given what outcomes I think are possible as well as how likely and desirable I find them. Rational behavior is thus what an agent "ought" to do, given her beliefs and preferences. If an agent has wicked desires, on this account, then monstrously immoral behavior that satisfies them can be

rational. Russell, Rawls, Foot, Harsanyi, Simon and Nagel espouse or assume something like this view of rational behavior.[149]

Rational Behavior as Ethical Behavior

On this view, acting rationally is acting ethically; theories of rational behavior are, at bottom, theories of right action, since both concern what one has most reason to do. This use of "rational" is uncommon in economics, but Sidgwick, Moore, Smart, Brandt, Hare and Gibbard use "rational" in this way.[150]

Rational Behavior as Prudent Behavior

The word "rational," as Hare says, "is sometimes used more or less synonymously with 'prudent'. . ."[151] Rawls, for instance, says that, "I have assumed throughout that the persons in the original position are rational. In choosing between principles each tries as best he can to advance his interests."[152] And now David Lewis:

> And some—I, for one—who discuss Prisoners' Dilemma think it is rational to rat no matter how much alike the two partners may be, and no matter how certain they may be that they will decide alike. Our reason is that one is better off if he rats than he would be if he didn't, since he would be ratted on or not regardless of whether he ratted.[153]

Using "rational" to refer to self-interest has deep roots. "It has been assumed," says Parfit, "for more than two millennia, that it is irrational for anyone to do what he knows will be worse for himself,"[154] while according to Sen, "the self-interest interpretation of rationality . . . has been one of the central features of mainline economic theorizing for several centuries."[155] "The concept of rationality familiar in social theory," says Gauthier, "identifies rationality with the maximization of individual utility."[156] Finally, in his sympathy for the view, Sidgwick finds company in the ancients, Spinoza, Butler, Clarke and Bentham.[157]

Why "Rational" Does More Harm Than Good in Ethics

The three accounts of rational behavior, in a nutshell, are:

(1) Rational behavior is Bayesian behavior.
(2) Rational behavior is ethical behavior.
(3) Rational behavior is prudent behavior.

Are these three competing theories of rationality? If they are, then I haven't the foggiest notion what they're competing over. 2 and 3 (but not 1) might compete over how one ought to live, if 3 expresses egoism and 2 means one ought to live ethically, where not all ethical behavior is self-interested. But most philosophers who use "rational" interchangeably with "prudence" aren't egoists, so that wouldn't account for most uses of "rational" in accordance with 3.

These ways of using "rational" might be linked by a weak family resemblance, if in each case the term is used to praise behavior: calling Bayesian behavior "rational" would praise it for being well-connected to the appropriate ends (like calling something "a good knife"); calling ethical behavior "rational" would praise it for being best; calling prudent behavior "rational" would praise it for being best in terms of self-interest. Such a weak common thread, however, would fail to support a unified sense for the term. For example, characterizing "rational behavior" as "action that is praised for any reason" would let too much behavior count as rational. Moreover, to add to the semantic clutter, it is not clear that economists typically use "rational" as a term of approval; often they may use it descriptively to refer to the self-interested behavior of real or hypothetical consumers. Furthermore, philosophers might use "rational" to mean "self-interested" without connoting praise.

I conclude that "rational," as applied to action, has at least three meanings at work in the philosophical literature: prudential, ethical and Bayesian. Moreover, "rational behavior" sometimes means something like "behavior that depends on adequately supported beliefs." Suppose, for instance, I believe that touching doorknobs ordinarily endangers one's health. One might call my behavior—scrubbing away in rubber gloves—"irrational" for having such a basis. This usage is less common in philosophy than the others. But all of these meanings for "rational" are sufficiently similar that one can easily be unsure what an author or speaker intends in using the term. (Or worse, different readers or listeners may confidently assume different interpretations.) "Rational" is not like "bank" or "bishop;" one would never think that Chase Manhattan Bank is part of the Hudson River or that Roman Catholic bishops are more effective

than knights in open positions. Moreover, when an author uses "rational" in its prudential or Bayesian senses, I am often unsure how much praise she intends to bestow. Using "rational" to modify behavior, I think, breeds misunderstanding.

Does "rational" have a *proper* meaning in philosophy? The Bayesian account at least gives the term a distinctive place in the philosophical lexicon; on other uses, "rational" may be replaced with greater clarity by "ethical" or "prudent." However, I think that rational behavior is properly equated with ethical behavior (or moral behavior, where these are used interchangeably), for "rational" derives from "ratio," the Latin word for "reason," while ethical behavior is behavior supported by the best reasons.

Parfit rejects this view after attributing it to Sidgwick:

> Sidgwick thought that [What is rational? and What is right?] were, in the end, the same, since they were both about what we had most reason to do. This is why he called Egoism one of the 'Methods of Ethics.' A century later, these two questions seem further apart. We have expelled Egoism from ethics, and we now doubt that acting morally is 'required by Reason.'[158]

Parfit's objections are cryptic. First, he says that in the last century we have expelled egoism from ethics. But the term "ethical egoism" is still in use.[159] More commonly, philosophers define "morality" such that moral theories must sanction impartiality (which expels egoism from morality); but, if so, then many ethicists endorse nonmoral ethical theories. Second, Parfit says that we now doubt that acting morally is required by reason. What do we now doubt? (i) We may doubt that acting immorally entails an inconsistent will or inconsistent values or beliefs. (ii) "Morality is required

by reason" is a Kantian slogan. Denying it might merely express unsympathy for Kant's ethics. Neither interpretation counts against "ethical" being the proper meaning for "rational" (although, as I've indicated, using "rational" in this way breeds misunderstanding).

Parfit agrees with Sidgwick up to a point. Moral theories and theories of rationality, he thinks, *are* about what we have most reason to do.[160] However, Parfit believes they differ fundamentally, and I can't discern how. He doesn't define "rationality,"[161] nor can I glean how he understands it from his absorbing discussions. He doesn't understand rational behavior in terms of self-interest, for he says that it can be rational to act in the interests of other people, even when one knows that one's act is against one's self-interest.[162] Nor does he bar theories of rationality from criticizing goals; "It is irrational to desire something," says Parfit, "that is in no respect worth desiring, or is worth avoiding." I am at a loss as to how theories of rationality and ethical theories, on Parfit's view, differ.

Rationality has turned out to be a red herring. The idea that self-interested behavior is rational may mean nothing more than "self-interested behavior is prudent." Or, the intuition that prudence is rational may just be the belief that I am entitled to give my interests special weight. To find reasons for that belief, we must look elsewhere.

Sidgwick on Egoism

According to Sidgwick, "the Egoistic first principle" may be "formulated" in two manners. One is easily refuted:

> When . . . the Egoist puts forward, implicitly or explicitly, the proposition that his happiness or pleasure is Good, not only *for him* but from the point of view of the Universe,—as (e.g.) by saying that 'nature designed him to seek his own happiness,'—it then becomes relevant to point out to him that *his* happiness cannot be a more important part of Good, taken universally, than the equal happiness of any other person.[163]

But, Sidgwick thinks, egoists can hold their ground if they don't imply that one's good is good from an impersonal perspective:

> If the Egoist strictly confines himself to stating his conviction that he ought to take his own happiness or pleasure as his ultimate end, there seems no opening for any line of reasoning to lead him to Universalistic Hedonism as a first principle; it cannot be proved that the difference between his own happiness and another's happiness is not *for him* all-important.[164]

Earlier, however, Sidgwick himself seemed to endorse an argument against all forms of egoism:

> So far we have only been considering the 'Good on the Whole' of a single individual: but just as this notion is constructed by comparison and integration of the different 'goods' that succeed one another in the series of our conscious states, so we have formed the notion of Universal Good by comparison and integration of the goods of all individual human—or sentient—existences. And here again, just as in the former case, by considering the relation of the integrant parts to the whole and to each other, I obtain the self-evident principle that the good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other; unless, that is, there are special grounds for believing that more good is likely to be realised in the one case

> than the other. And it is evident to me as a rational being that I am bound to aim at good generally,—so far as it is attainable by my efforts,—not merely at a particular part of it.[165]

If, on Sidgwick's view, it is evident to me that I should aim at good generally, why does he later despair of refuting egoism? First, he uses "evident" to mean "evidence," not "overwhelming evidence." Second, Sidgwick thinks that the egoist also offers evidence. (I'll discuss that argument in the next section.) Third, he must realize that appealing to the "Universal Good" and to "the point of view of the universe" begs the question against the egoist, who favors Individual Good and the personal point of view.[166]

Sidgwick's distinction helps rebut the objection that egoism involves a "failure to recognize oneself as just one person among others."[167] If the egoist sees her interests as important from the point of view of the universe, then the existence of other people is embarrassing—why shouldn't their interests be important too? But the more resilient egoist may recognize others by recognizing that their interests are good-for-them, just as hers are good-for-her.

Also, consider Korsgaard's argument against egoism (following Nagel in *The Possibility of Altruism*). If you are tormenting me, she says, I can say, "How would you like it if someone did that to you?" And then, you will realize that if you were in my position,

> you would not merely dislike it, you would resent it. You would think that the other has a reason to stop, more, that he has an obligation to stop. And that obligation would spring from your own objection to what he does to you. You make yourself an end for others; you make yourself a law to them. But if you are a law to others in so far as you

> are just human, just *someone*, then the humanity of others is also a law to you.[168]

So, according to Korsgaard, I can oblige you to stop tormenting me by forcing you to reason in this way:

"(a) If I were in your position, I would (rightly) think, 'He should stop tormenting me simply because of my humanity.'
(b) Therefore, in actuality, I am obliged to stop tormenting you."

But instead of "He should stop tormenting me simply because of my humanity," one might think, "It would be good for him to stop tormenting me, not because of my humanity, but because it's me." Again, the second form of egoism proves to be more resilient.

Arguments For the Spectrum

Most human beings are selfish—profoundly so—by any decent standards, and even the best of our lot feel pulled in that direction. Self-bias seems deeply engrained in human beings prior to ethical theorizing, even prior to speech. Philosophers, I think, rarely *argue* for the Spectrum because their main motive for that ethic—caring more about themselves than others—is transparently weak as a rationale. Is there one better?

The Separateness of Persons

I have my pain, and you have yours. Does the separateness of persons support giving one's own interests special weight? According to Sidgwick,

> It would be contrary to Common Sense to deny that the distinction between any one individual and any other is real and fundamental, and that consequently "I" am concerned with the quality of my existence as an individual in a sense, fundamentally important, in which I am not concerned with the quality of the existence of other individuals: and this being so, I do not see how it can be proved that this distinction is not to be taken as fundamental in determining the ultimate end of rational action for an individual.[169]

Parfit responds by challenging Common Sense. In Part Three of *Reasons and Persons*, he argues persuasively that persons are not simple soul-like entities; rather, they consist in subpersonal items such as personality traits, memory-links and physical continuity. Most metaphysicians now agree. Such "reductionist" views of personhood, many philosophers believe, count *against* the Spectrum.[170] Here I assert more humbly that nothing about personhood counts for the Spectrum; in other words, the nature of persons provides no reason for me to think that a pleasure's being mine rather yours should be important to me.

Brink disagrees, saying that the separateness of persons "suggests" and "explains the appropriateness of" the principle that it is unreasonable to make uncompensated sacrifices.[171] But if so, why wouldn't the separateness of the sexes provide a reason to think that it's unreasonable

for a man to make sacrifices that don't benefit men? Brink says, "part of what it is for me to be a separate person is for me to be unwilling to sacrifice my interests without appropriate compensation." If so, then making sacrifices is unlikely or impossible, not unreasonable. And he says that:

> Even if we do not suppose that special concern is itself constitutive of persons, it is commonly believed that we do have special reason to be concerned about our own lives that we don't have with regard to the lives of others. The rationality of an agent's action seems not in general to be proportional to the good that she does.[172]

But this is a mere intuitive appeal; to argue against agent-neutrality, one must explain *why* the separateness of persons, but not the separateness of the sexes, provides adequate grounds for bias. And no one, I think, has done that.

Other Arguments For Favoring One's Self Over Others

Partiality towards oneself may be natural, but is it good or right? As I remarked in the first chapter, arguing from naturalness to value most likely assumes a suspect theology or teleology. Moreover, counterexamples to "what's natural is good" abound: cystic fibrosis, male aggression, heavy snowfall in Syracuse, jealousy, depression in winter and competitive infanticide each seem natural in many instances.[173] Of course, if ought implies can, and if I cannot be perfectly impartial, then it is not the case that I ought to be. But that result doesn't support the Spectrum. On the

Spectrum, I'm entitled to give my interests special weight because they're mine, not because I must.

Not only are human beings naturally partial to their own interests, but our *strongest* desires are typically for our own well-being. Nothing of interest to us, however, follows from this; Bishop Butler warned against confusing power and authority,[174] and here we shouldn't confuse the power of our desires with their authority. But one might hold that a person's most fundamental, integrated and unshakable desires or tendencies-to-desire are of special moral significance. Those desires constitute one's *grain*—like the grain of wood—and, on this view, one has special reason not to go deeply and steadily against them.[175] This could provide a reason for giving one's own welfare special weight. But what about people who are rotten to the core—do *they* have a normative reason not to go against their most basic desires? Such a normative reason, of course, could be outweighed by others; the view in question doesn't entail that one should never go against one's grain (although, as before, it might be false to say that one "ought" to go against one's grain, if one can't). Nevertheless, the thesis that basic desires have special moral significance seems hard to motivate given that not all human beings are basically good.

Many philosophers worry, "If I have no moral reason to favor myself over others, then morality will demand too much of me." But a savvy moralist wouldn't demand that these philosophers act like saints, for their having this worry suggests that such demands would be counterproductive. The following *is* true: if I have no moral reason to favor myself over others, then I would need to make vast sacrifices to be the best I could be (assuming that I could make vast sacrifices). But being one's best should hardly be easy, "for the world's more full of weeping than you can understand."

Arguments Against the Spectrum

Before arguing against the Spectrum, I want to criticize some arguments I haven't yet discussed for the same conclusion. The first is conceptual: Narveson, following Moore, says that "The word 'good,' and other evaluative words, are logically incapable of denoting any sort of 'private' object . . ."[176] This semantic claim is doubtful; why couldn't I commend only those benefits accruing to me? Anyway, egoists might want to revise the logic of evaluative words.

The next argument—inspired by Sidgwick, Nagel and Parfit[177]—is analogical:

(1) It doesn't matter whether one receives a benefit *now* or at some other time.
(C) It doesn't matter whether *I* or someone else receives a benefit.

Why should (1) support (C)? Parfit presses the analogy between "oneself and the present, or what is referred to by the words 'I' and 'now.'" He says:

> This analogy holds only at a formal level. Particular times do not resemble particular people. But the word 'I' refers to a particular person *in the same way in which* the word 'now' refers to a particular time. And when each of us is deciding what to do, he is asking, 'What should I do *now*?' Given the analogy between 'I' and 'now', a theory ought to give both the same treatment.[178]

Parfit mentions two similarities between "I" and "now": (i) both are indexicals; (ii) "when each of us is deciding what to do, he is asking, 'What should I do *now*?'" I'll relate these ideas to the argument as follows. One might want to assign the present special importance because practical deliberation paradigmatically involves now-thoughts: what should I do *now*? Similarly, one might want to favor oneself because practical deliberation paradigmatically involves I-thoughts: what should *I* do now? But "now," being indexical, refers to different times on different occasions; hence, no particular time is privileged by the form of practical deliberation. And, similarly, "I," being indexical, refers to different persons on different occasions, so practical deliberation puts no halo around any particular person. Reply: all this may be true, but it cuts ice only against the egoist who is moved by the form of paradigmatic practical deliberation.

If analogical arguments are hard to assess, then formal analogical arguments are especially problematic, so I'll restrict my conclusion to this: the analogy between the present and oneself needs more work to be persuasive.[179]

Egoism entails the existence of private (or agent-relative) reasons. The egoist, for example, believes that I, but not you, have a reason to prevent my being harmed. Korsgaard argues that reasons are essentially public (or agent-neutral). She says,

> The public character of reasons is indeed created by the reciprocal exchange, the sharing, of the reasons of individuals. . . . If these reasons were essentially private, it would be impossible to exchange or to share them. So their privacy must be incidental or ephemeral; they must be inherently shareable. . . . what both enables and forces us to share our reasons is, in a deep sense, our social nature.[180]

"Individuals share their reasons with one another"—does this show that reasons are agent-neutral? If I alone have reason to promote my interests, I can still share my reasons with you in the sense of explaining my motives to you. And, in response, you may be moved to help me, but your motive needn't constitute a normative reason.

I agree with Nagel in *The View From Nowhere* (echoing Sidgwick): "no completely general argument about reasons can show that we must move from the admission that pleasure and pain have relative value to the conclusion that they have neutral value as well."[181] So, one must look elsewhere for evidence against egoism.

Nagel's Arguments for the Agent-Neutrality of Pain

In *The View From Nowhere*, Nagel argues for severe pain having agent-neutral disvalue. First, he says,

> There's a reason for me to be given morphine which is independent of the fact that the pain is mine—namely that it's awful.[182]

If so, then presumably my pain has agent-neutral disvalue. But *is* the awfulness of my pain independent of its being mine? Here "awful" means "highly unpleasant." I argued in chapter 2 that the category *unpleasant* is unified by the fact that, among experiences, all and only the unpleasant ones are bad due to how they feel. But unpleasant experiences could be unified by having either agent-neutral disvalue (because of how they feel) or

agent-relative disvalue (because of how they feel). And if pain has agent-relative disvalue, then an awful experience just is an experience that is very bad for the person who has it (in this case, me). By assuming that my pain's awfulness is independent of my having it, Nagel merely assumes that pains don't have agent-relative disvalue.

Also, Nagel says,

> [My] pain can be detached in thought from the fact that it is mine without losing any of its dreadfulness. It has, so to speak, a life of its own. That is why it is natural to ascribe to it a value of its own.[183]

So, according to Nagel, my pain can be detached in thought from its being mine without losing any of its dreadfulness; therefore, my pain has agent-neutral disvalue.

On the next page Nagel elaborates on the idea, expressed by his premise, that the badness of a pain doesn't seem tied to the subject of the pain:

> [the sufferer's] awareness of how bad [his pain] is doesn't essentially involve the thought of it as his. The desire to be rid of pain has only the pain as its object. This is shown by the fact that it doesn't even require the idea of *oneself* in order to make sense: if I lacked or lost the conception of myself as distinct from other possible or actual persons, I could still apprehend the badness of pain, immediately. . . . To regard pain as impersonally bad . . . does not involve the illegitimate suppression of an essential reference to the identity of its victim. In its most primitive form, the fact that it is mine—the concept of myself—doesn't come into my perception of the badness of my pain.[184]

I'll use Nagel's premise in my argument below. That premise, however, doesn't show by itself that one's pain has agent-neutral value. For someone who holds that pain has agent-relative disvalue may reply: "I agree that if severe pain is detached in thought from its being mine—or if I lost the concept of myself—that pain would still seem dreadful. But all this means is that, under such circumstances, I wouldn't know who the pain is dreadful for; I wouldn't know the identity of the person who alone has a reason to end it. Pain has a disvalue of its own in the sense that a pain is bad for whoever has it." (Of course, if a severe pain could be detached, not just in thought, but in reality from any person or subject, without losing its dreadfulness, then *that* pain—and presumably all others—would have agent-neutral disvalue. But unowned pains seem metaphysically or absolutely impossible, perhaps because conscious states are merely ways beings are.)

Although Nagel's arguments, as formulated, strike me as unsuccessful, I'll now offer an argument in the spirit of his approach.

A Related Argument

One's most basic source of evidence for the badness of pain is one's experience of it. That much is clear. This is why, in chapter 1, I began my project by observing that, when I am in severe pain, that pain, as it's presented to me, gives me evidence that it's bad in some way. Let's take up the argument at this point. In doing so we do not assume the point of view of the universe; our discussion proceeds from the individual's perspective.

If a pain of mine is bad, who has reason to end it? Now Nagel's point comes into play: from my own perspective, while in pain, what seems bad is just the experience itself; the fact that it's mine, whether recognized or not, is incidental to my judgment. Moreover, I could think "this pain is bad" if I lost the concept of myself or if I followed Buddha[185] in believing that subjects of experience don't exist.

Egoists rightly believe that I have reason to end my pain. But the grounds for this belief should derive from the most basic source of evidence; so, the ground is my judgment "this pain is bad." Hence, I have reason to end my pain because the pain itself would end; that bad thing would stop. But if *this* grounds my having reason to end my pain—as I think it does—then it also supports the thesis that other people have reason to end my pain. For the pain would be "just as stopped" by someone else than by me. Hence, my pain has agent-neutral disvalue. A similar argument shows that any severe pain of yours has agent-neutral disvalue.

Does this argument succeed? The egoist would like to say, "The grounds for my having a reason to end my pain is that *my* pain is bad, not that *this* pain is bad." However, even an egoist should concede that, in assessing the disvalue of one's pain, one should begin with one's experience of it. Thus, I won't let the egoist—now quoting Sidgwick—to strictly confine himself to stating his conviction that he ought to take his own happiness as his ultimate end. To insist on this doesn't "beg the question" against the egoist, for I am not foisting an impersonal perspective on him. But once the egoist concedes that we should begin by examining the first-person experience of pain, he can't return to his safe perch because what seems bad to me most basically is not my pain but *that pain*. If so, then what grounds my self-interested concern grounds the concern of

others. And, of course, it's a two-way street, for what grounds the self-interested concern of others also grounds my concern for them.

This argument, I hope, beats the egoist on his home turf; the evidence against egoism is drawn from *within* the first-person perspective. I know from within that my experience of my severe pain provide a reason for anyone to end it; anyone should end it because of how it feels. *How a pain feels* is clearly independent of its being mine, even if *a pain's being awful* isn't clearly independent of its being mine.

Echoing the first chapter, two types of argument now support the thesis that all of my pleasures and unpleasures have agent-neutral significance (and the same arguments apply to your pleasures and unpleasures). First, extending the argument just given, all of my pleasant experiences seem good to me just as experiences; and all of my unpleasant experiences seem bad to me just as experiences. Hence, I have evidence that all of my pleasures and unpleasures provide agent-neutral practical reasons. Second, I'll repeat my appeal to analogies. All intense unpleasures, I've just argued, are impersonally bad. If so, then (i) less unpleasant experiences should also be impersonally bad, though less bad; and (ii) intense pleasures should be impersonally good. Furthermore, (i) and (ii) each support less intense pleasures having agent-neutral or impersonal value.

Conclusion

I found no evidence for the thesis that one has a basic normative reason to favor oneself over others. Moreover, I endorsed an argument for pleasures and unpleasures having agent-neutral significance. So, I conclude, they have such significance. If so, this brings me one step closer to concluding that pleasures are intrinsically good and unpleasures are intrinsically bad.

Chapter 4: Is it Good to Make Happy People?

If pleasures have intrinsic value, then there would always be reason to bring about additional pleasure. Usually one would do so by bringing pleasure into the life of an existing person, but one could also bring to life someone who would feel pleasure. Is the fact that someone would feel pleasure a reason to bring her into existence? This question raises peculiar issues, not about hedonic value, but about potential persons. I'll argue that it would be good for additional people to exist whose lives would be worth living. (I'll refer to people whose lives are worth living as "happy people.") If so, then part of the reason why such people should exist is that they would feel pleasure.

Arguments Against Additional Happy People Being Good

The No-Obligation Argument

1. If it would be good for additional happy people to exist, then we would be morally obliged to have children.
2. We are not morally obliged to have children.

C. Therefore, it is not the case that it would be good for additional happy people to exist.

The first premise of this familiar argument is questionable. For one thing, creating people who would flourish might be good *in one respect* yet not be obligatory because countervailing considerations ensure that it would not be good *all things considered*. Although the new person might be happy, the consequences of bringing him or her into existence might be bad enough for other people that having the child would not be overall good. Second, creating people might be overall good but not good enough to warrant calling it an "obligation," just as buying Girl Scout cookies is good even though we are not obliged to do so. Third, having children might be good but involve such sacrifices or be so psychologically demanding that it would be supererogatory rather than obligatory. And finally, having children might not be obligatory because I can better use my time and money. The thousands of dollars required to raise one child can save the lives of many starving children who already exist.[186]

These replies can be circumvented. A variant of the argument goes like this:

1. If it would be good for there to be additional happy people, then, if God exists, God would be obliged to create infinitely many people.
2. If God exists, God would be under no such obligation.

C. Therefore, it is not the case that it would be good for additional happy people to exist.

Now I deny the second premise, as would anyone who rejects the conclusion.

Failing to Create is Bad for No One

Consider another argument:

1. If there are reasons against behaving in a certain way, then that behavior is bad for someone.
2. If we do not create additional happy people, our behavior is bad for no one.

C1. Therefore, there are no reasons against not creating additional happy people.

4. But if it would be good for additional happy people to exist, then there would be reasons against not creating them.

C2. Therefore, it is not the case that it would be good for additional happy people to exist.[187]

Parfit's Non-Identity cases show that premise 1 is false.[188] One such case is "Depletion." If we adopt a social policy of depleting resources, we both alter the identity of future generations and lower the future quality of life, although life would still be worth living. Depletion is bad for no one because the people who would exist would be happy, and without depletion they wouldn't exist. Nevertheless, there is reason not to deplete.[189]

Following Narveson,[190] Bennett endorses a principle similar to premise 1:

> The question of whether action A is morally obligatory depends only upon the utilities of people who would exist if A were not performed.[191]

Suppose we must perform either action D or action ~D. D results in depletion and ~D does not. Is ~D morally obligatory? If it is, then Bennett's principle fails because ~D would be obligatory because the utilities of those who would exist were it performed would be greater than the utilities of those who would exist if it were not. Alternatively, perhaps there is reason to perform ~D but ~D is not "morally obligatory"—for example, ~D might be supererogatory, so D would be permissible. This position is consistent with Bennett's principle but inconsistent with using it to deduce that it isn't good for additional happy people to exist. For the position affirms that the utilities of people we could create are relevant to whether we have reason to create them; it merely denies that such reasons suffice to create obligations.[192]

Tooley's Appeal to Rights and Obligations

Tooley advances the general thesis:

> S: An action is prima facie wrong if and only if it involves a failure to fulfill an obligation regarding some individual, when it was possible to do so, or it makes it the case that there is some individual with respect to whom there will be an obligation that cannot be fulfilled.[193]

S revises the idea that, if there are reasons against behaving in a certain way, that behavior is bad for someone. S entails that, if there are reasons against behaving in a certain way, then that behavior is either bad for someone or creates an obligation that cannot be fulfilled. Thus, S entails

that "there is no prima facie obligation to produce additional persons" because "refraining from producing additional persons does not in itself either violate an obligation with respect to any individual, or make it the case that there is an individual with respect to whom there are obligations that cannot be met."[194] But S doesn't entail that it is prima facie wrong to fail to create someone with respect to whom obligations can be satisfied. However, when we create a person whose life must be miserable, our obligation to respect that person's right to life worth living cannot be fulfilled.

Tooley says that S is "free of unacceptable consequences." By itself, S has few consequences; it needs to be supplemented with principles specifying obligations. However, S is vulnerable to Non-Identity style counterexamples. Consider the case of Carlo and Jane. A woman trying to get pregnant is in danger of passing along a heart condition to her child—but only to a male child. This condition would be unpleasant at times but not fatal. The woman can take a pill to ensure a female child. Without this pill, she might have a boy named Carlo, and with it, she would have a girl called Jane. To ensure that Carlo and Jane are different people, assume that the pill would alter the timing of her pregnancy. Suppose that, if the woman does not take the pill and (unluckily) gives birth to Carlo, Carlo would be provided for, but he would not be nearly as happy as Jane would have been.[195] The woman would be prima facie wrong not to take the pill, even though, if she has Carlo, it seems that she would have no unfulfilled obligations with respect to him.

Tooley agrees that the woman should take the pill.[196] However, he believes that there would be an unfulfilled obligation to Carlo. In defending this claim, he appeals to a principle of equal opportunity:

> Every person has a right to an equal chance of enjoying those natural resources, both environmental and genetic, that a person living in his society might enjoy, and that make it possible for one to lead a satisfying life.[197]

Hence, Carlo, with his genetic heart ailment, is denied an equal chance of enjoying at least some important genetic resources.

But there are two problems with this view:

1. Suppose a couple is poor but happy. Would it be prima facie wrong for them to have a child? If we interpret S and the principle of equal opportunity such that they entail that it would be wrong for the woman not to take the pill, then those principles imply that it would be prima facie wrong for poor people to have children. After all, their children would have a less than equal chance of enjoying environmental resources.

One might think this is acceptable, since the poor couple's obligation is only a prima facie obligation that could be overridden. But the prima facie obligation can be easily turned into an absolute obligation. First, bear in mind that, according to S, the fact that a person would be happy is not a reason to create her. And suppose that this child's existence, on the whole, would neither benefit nor harm other people. It then follows that it would be wrong, all things considered, for the poor couple to bring this happy child into the world. This conclusion seems mistaken, especially since the child would be happy, despite enjoying fewer resources than most others.

2. Suppose half the population enjoys all the relevant natural resources. A woman is deciding whether to take a fertility pill with specific side-effects. If she takes the pill, she will have twins. One of those twins (we can't know

which) will be healthy and enjoy all the resources, but the other will become sick and won't. Taking this pill is permissible, on Tooley's view, since each twin would have the same chance of enjoying the natural resources as the population at large (one chance in two). (It might count against Tooley's view that taking the same pill would be unacceptable if 52% of the people in the population enjoy the resources, for then each twin would have a less than equal chance of enjoyment.)

Change the example slightly. Call the twins the woman would have "Lefty" and "Righty." Suppose the woman knows that Lefty would be the sick one. Now Tooley's view entails that the woman mustn't take the pill, since Lefty wouldn't have a fair chance of enjoying health. But suppose that Lefty is the same person who would have become sick in the original example. Then, on Tooley's view, taking the first pill is permissible, while taking the second pill is not, even though (i) they would lead to the same result, and (ii) if the second pill led to a different result, that result would be as bad as what would actually happen (Righty's being sick would be as bad as Lefty's being sick). This implication counts against Tooley's theory.

The No-Benefit Argument

Some philosophers say that *we do not benefit people by creating them.* This suggests the following argument:

1. A person who is created does not benefit from having been created.
2. If a person does not benefit from having been created, then adding that person to the world (even if she would be happy) does not make the world better.

C. Therefore, it is not the case that it would be good for additional happy people to exist.

The first premise is not obviously true; Parfit deftly defends denying it.[198] Perhaps 1 is supposed to be true for the following reason: a *benefit* , by definition, promotes one's interests; but to *promote one's interests*, one must, again by definition, have interests. If we define words like this, then the second premise begs the question, for the opposing camp is defined by the belief that we have reason to create happy people even if they don't benefit in that sense. Similarly, we have reason not to create miserable people even if doing so harms no one.

The No-Preference Argument

Some people might be persuaded by this argument:

1. If it would be good for additional happy people to exist, then creating those people must satisfy some of their preferences.
2. Creating happy people does not satisfy any of their preferences.

C. Therefore, it is not the case that it would be good for additional happy people to exist.

We cannot fully understand this argument without knowing what theory of preference lies behind it. However, we understand it well enough to reject it. Consider the following, parallel argument:

1. If it would be bad for unhappy people to exist, then creating unhappy people must frustrate some of their preferences.
2. Creating unhappy people does not frustrate any of their preferences.

C. Therefore, it is not the case that it would be bad for additional unhappy people to exist.

The conclusion of this parallel argument is false because it is bad for a person to exist who suffers intensely without compensation. Where does it go wrong? Since the argument is valid, it must have a false premise. But neither of its premises can be denied without undermining the corresponding premise of the No-Preference Argument. Its second premise may be challenged in two ways: (i) creating unhappy people frustrates desires because those people will prefer not to have been born; (ii) creating an unhappy person frustrates those desires that constitute the person's unhappiness. If either of these is true, then we should be able to say, against the second premise of the No-Preference Argument, that (i) creating happy people satisfies desires because those people will prefer to have been born; or (ii) creating a happy person satisfies those desires that constitute the person's happiness. A similar strategy applies to the first premise of the parallel argument. That premise utilizes the principle that *what is bad for people must frustrate their preferences*. It will be hard to defend this while denying that *what is good for people must satisfy their preferences*.

In short, the No-Preference Argument stands or falls with its parallel. Since its parallel falls, it falls.

Contractualism

Egoistic contract theories base ethics entirely on agreements of mutual benefit. Thus, Gauthier calls contractualism "the morality of mutual advantage"[199] and says that his theory "denies any place to rational constraint, and so to morality, outside the context of mutual benefit."[200] This implies that I have no reason to create additional happy people, unless doing so benefits me.

Contractual egoism is unsatisfactory because moral reasons exist outside the context of such agreements. If I see an animal suffering, and I can easily help it, I should do so even though the animal cannot help me. And humans who suffer bear the same relation to God, if God exists.

Turn now to a Rawlsian "rightness as fairness" framework.[201] The contractors are still self-interested, but they do not know what place they will occupy in society. Why would they prefer fewer happy people to exist in society? Consider two contractual scenarios:

(a) Suppose that, in the original position, the contractors know they will exist in society. In this case they wouldn't want additional happy people, provided that the smaller population would be better off, on average, than the larger population. And, given limited resources, this might be the case.

But now rightness-as-fairness has ludicrous implications. For if the contractors know that they will exist in society, then they will prefer society X over society Y if each person in X is better off than each person in Y. And this is often unacceptable when all the people in X and Y have lives not worth living. For example, the contractors would prefer a society in which billions of people live in ghastly, repulsive conditions to one in which five people live in conditions very slightly worse.[202]

(b) Suppose that, in the original position, the contractors do not know whether they will exist in society. Would they be willing to risk not existing by affirming principles that don't grant the mere potential for happy existence much weight?[203]

If the contractors don't know whether they will exist in society, it isn't clear what population principles they would act on. Maximin alone would be a poor guide. For in a population containing only lives worth living, the least advantaged class would be the non-existent, if some contractors do not exist in society. The theory would then imply that ten billion and one lives barely worth living are preferable to ten billion lives of very high quality (plus one life not lived)—an unacceptable result. Of course, the contractors could apply principles characteristic of other theories; questions about population do not show the incompleteness of contractualism or Kantian constructivism. However, contractualism has no distinctive method for answering these questions.

The No-Sympathy and the Less-Sympathy Arguments

Some philosophers might emphasize that we are *unmoved* by thoughts concerning potential people. The problem, it might be said, is that, prior to life's beginning, there is no one with whom to sympathize. Therefore, potential people don't matter. Alternatively, someone might say, we sympathize with potential people less than living people because we can't at this time see, touch or befriend them; because as yet they have no projects or desires; because they can't help us, harm us or hold us responsible for not creating them.

The arguments are these:

1. We feel no sympathy for potential people.

C. Therefore, it is not the case that it would be good for additional happy people to exist.

1. We have less sympathy for potential people than for living people.

C. Therefore, it would be slightly good for additional happy people to exist.

These arguments are weak because our sympathies are unreliable. Our capacity for sympathy is notoriously underdeveloped, for example, in connection with animal suffering and human starvation. We find it difficult to sympathize with pigs or people we don't know, but we should oppose factory farming and support human rights. Why shouldn't something similar be true of our lack of sympathy for potential people? Why shouldn't the thought that *if I were to have a child, that child would be*

happy move a benevolent adult toward becoming a parent? This thought does not assume that there now exists someone with whom to sympathize.

The Less Sympathy Argument compromises: it concludes that potential people matter but not very much. There is another compromise to consider:

The Average Utility Compromise

Some people believe that insofar as utility matters, what matters is average utility or the average level of well-being. Average utilitarianism entails that it is good for there to be additional happy people, but only when their existence augments average utility. However, this is unacceptable because it yields pairs of judgments such as the following:

—It is good to create someone who is mildly happy in a world in which happiness and unhappiness are balanced.
—It is bad to create someone very happy in a world of even higher average happiness.

Other arguments also show that the principle of average utility is mistaken. If a population lives in writhing agony, it would be better with respect to average utility if someone were added to the population whose agony was just a tiny bit less severe because this would increase the average level of well-being. But one more suffering soul does not improve the universe with respect to well-being.[204]

I have now rejected the most important arguments and principles entailing that it is not good for additional happy people to exist.

Arguments Supporting Additional Happy People As Good

Appeals to Intuition

Some plausible beliefs suggest that it's good for additional happy people to exist.

1. It would have been bad if the happiest million people in human history had never been born.
2. It would have been bad if human beings had never evolved (assuming here and in 3 that the good in human life outweighs the bad).
3. It would be wrong for everyone to take a drug that causes both infertility and an indifference to being infertile, thus ensuring that the youngest generation alive will be the last.[205]
4. It would be good for God to add ten billion flourishing people to a distant part of the cosmos.

If 1-4 are true, then the best explanation for their truth is that it would be good for additional happy people to exist. 1-4 are true; so it would be good for additional happy people to exist. Against this, one may offer principles and arguments implying that some of 1-4 are false and the rest are true but not

because it is good for happy people to exist.[206] I won't try to adjudicate this dispute, but I submit that 1-4 provide slight evidence for the conclusion.

An Implausible Consequence of Asymmetry

It is bad for people to exist whose lives are not worth living or worse than nothing. This is not to say that such people should be killed; but we do believe it would be better, other things being equal, if such people never came into existence. Consider Seabrook's remarks:

> When one thinks of a truly awful genetic disease, like Lesch-Nyhan syndrome, a rare mutation in the single X chromosome in males which causes mental retardation, extreme physical pain, and compulsive self-mutilation—children savagely gnaw their fingers and lips unless they are constantly constrained—one wonders whether the child who would have to suffer from this condition has a right *not* to be born.[207]

Now suppose you believe that, while it would be bad for additional unhappy people to exist, it would not be good for additional happy people to exist. This combination of beliefs entails that it is wrong, other things being equal, to act with the predictable consequence that (a) 100,000 happy people come to be, if one's behavior also creates a slight chance of (b) one person existing whose life is not worth living or worse than nothing.[208] Such behavior would be considered wrong because (b) counts against it while (a) does not count in its favor. Imagine that the one unhappy life would be *barely* not worth living and that everyone else is blissful. Does it not seem

that the bliss of the one hundred thousand outweighs the chance of mild misfortune for one? If so, then those additional happy people existing is good.

The Symmetry Argument

It is bad for people to exist whose lives are not worth living (at least in part) because their conscious states are unpleasant. Analogously, it is good for happy people to exist (at least in part) because their conscious states are pleasant. We can state the Symmetry Argument as follows:

1. The fact that someone's conscious experiences would be unpleasant is a reason against bringing that person into existence.
2. If (1), then the fact that someone's conscious experiences would be pleasant is a reason for bringing that person into existence.

C1. Therefore, the fact that someone's conscious experiences would be pleasant is a reason for bringing that person into existence.

C2. Therefore, it would be good for additional happy people to exist.

The Symmetry Argument is analogical. Just as unpleasant experiences should be avoided, pleasant experiences should be sought. How strong is the analogy? Pleasant and unpleasant experiences are ontologically similar: they are conscious items (whatever consciousness may be). Also, normally we have the same type of epistemic access to each: we know them by introspection. Furthermore, both have normative significance because

of how they feel. If a sensation feels good, then to some extent it is good; and if a sensation feels bad, then to some extent it is bad.

Several objections to this argument rely on positions that I have already criticized. Appealing to Tooley's principle, one might say that we shouldn't create a miserable person because doing so creates obligations that cannot be fulfilled; but there is no analogous reason to create a blissful person. Or one might say that creating a miserable person is bad for someone, but not creating a happy person is bad for no one. I have already addressed these and other challenges.

The Analogy Between Parts of Lives and Whole Lives

Suppose you expect to live many more years of high quality. If tomorrow you were to die tragically in a bizarre gardening accident,[209] this would not merely be worse for you, it would be bad impersonally. Thus, it is worse that one should die rather than live many golden years before dying. Similarly, it is worse that one should not exist rather than live a golden life before dying. This argument is Parfit's:

> Consider someone dying painfully, who has already made his farewells. . . . He might decide that, at some point in the past, if he had known what lay before him, he would or would not have wanted to live the rest of his life. He might thus conclude that these parts of his life were better or worse than nothing. If such claims can apply to parts of lives, they can apply, I believe, to whole lives.[210]

The argument proceeds by analogy: parts of lives can be good (or bad), so whole lives can be good (or bad). Whole lives and parts of lives are the same sort of entity. It is hard to show a relevant disanalogy between them that does not rely on a principle or argument criticized above.

I have argued in four ways that it would be good for additional happy people to exist. These arguments provide better evidence than the opposing arguments. Therefore, it is good for additional happy people to exist.

How Valuable is the Happiness of Potential Persons?

I'll approach this issue by asking whether the well-being of potential people matters as much as the well-being of actual adults. Suppose that Cayce, an adult, has a life of neutral value. Further suppose that we must choose between (i) raising Cayce's life to a high level of prosperity and (ii) creating a new person at that high level. Most of us believe that, other things being equal, (i) is better than (ii), and not merely because (i) is better with respect to average utility. Most of us believe that adult potential matters much more than the potential of the unconceived. I believe, however, that (ii) is as good as (i). I will argue for the *Strong Thesis*, which holds that potential people matter as much as adults.[211] The Strong Thesis suggests that potential life should weigh heavily in our reasoning. It entails that *this life would be worth living* provides the same sort of reason for creating life as it does for augmenting and prolonging adult life.

The Strong Thesis should be understood to include the idea that *the longer duration of happy life an individual has in prospect, the stronger the reason for creating or saving her*. So, other things being equal, the reason for prolonging life will be stronger if the individual is younger—say, an infant rather than a college student. This does not entail that infant interests normally outweigh adult interests because our resources often stretch further among adults than among the young. Infants, after all, need considerable care before they can provide for themselves.

A. A version of the Symmetry Argument supports the Strong Thesis. Consider:

The Strong Thesis As Regards Lives Not Worth Living: Suppose that Cayce has a life of neutral value. Further suppose we must choose between (i) making Cayce terribly miserable and (ii) creating a new person who will be terribly miserable. (i) is not worse than (ii), assuming away any influence the average principle might have on our assessment.[212]

This principle is (plainly) true: potential people becoming actual and suffering is as bad as adults suffering. But if so, then potential people becoming actual and flourishing is as good as adults flourishing. Together these two assertions imply the Strong Thesis. It may seem that disanalogies spoil the argument. However, I tried to answer such objections in assessing arguments against the idea that it is good for additional happy people to exist.

B. The analogy between parts of lives and whole lives can be used to support the Strong Thesis:

1. Parts of my life are better or worse than nothing.
2. If a part of my life is better or worse than nothing, then a whole life of similar quality and duration is exactly as good or bad as that part.

C1. Therefore, whole lives matter as much as parts of lives.

C2. Therefore, potential people matter as much as adults.

As before, the argument rests on a comparison between entities of the same sort: whole lives and parts of lives. I have not found satisfactory arguments to block the analogy.

Conclusion

Our evidence supports the Strong Thesis, according to which the well-being of potential persons matters as much as that of adults. And, in particular, we have just as much reason to promote pleasure, other things being equal, by creating people who would be pleased than by benefiting existing adults.

Chapter 5: Counterexamples to the Transitivity of *Being Better Than*

Why the Thesis is Not Too Ridiculous to Take Seriously

Ethicists and economists commonly assume that if A is intrinsically better than B, and B is intrinsically better than C, then A is intrinsically better than C. Call this principle *Transitivity*.[213] Transitivity provably stands or falls with the corresponding principle for *intrinsically worse than*, so I will treat them together.

I will offer counterexamples to the transitivity of *being intrinsically better than*. These examples employ more than three states of affairs, but a set of possibilities of any size provably violates Transitivity if it can be ordered so that each succeeding possibility is worse than its predecessor but the last one isn't worse than the first. My examples, if successful, also show the nontransitivity of "being all things considered better than," "being hedonically better than" and "being better for a person than."

In general terms, Transitivity might fail because factors determining how A&C compare differ (or differ in significance) from those determining how A&B and B&C compare.[214] In my examples, differences in duration and hedonic intensity are the only relevant factors. Intensity differences always matter, but their significance increases dramatically once they become sufficiently great. Temkin endorses this type of counterexample, using arguments based on earlier drafts of this chapter.[215]

Most philosophers strongly believe in Transitivity. This belief may derive force from the idea that value is like a line. If all states of affairs can be represented as points along a line, with better possibilities represented to the right of worse ones, then *being intrinsically better than* seems transitive because *being to the right of* seems transitive. However, this linear view of value is not sacrosanct. Many philosophers already reject it, for example, those who believe that some items can't be compared in terms of value. Einstein, moreover, discredited an analogous picture of time. On that picture, all events are represented as points along a line, with later events represented to the right of earlier ones. This view entails absolute simultaneity, which Einstein rejected, for two events either are or aren't represented at the same point on the line.

Let me press this analogy further. Conceptual arguments for Transitivity, I think, are no more effective than conceptual arguments for absolute simultaneity. Perhaps the linear view of value has become part of the meaning of value-terms; this might account for Transitivity's conceptual appeal. If so, we should revise those concepts. Similarly, the idea of a universal time might be part of what temporal notions mean (at least for ordinary speakers); this would account for relative simultaneity and the twin paradox seeming somehow incoherent. But physicists reject those definitions.[216]

Although Transitivity isn't apodictically certain, inductive evidence supports it: philosophers have often noted that an A is better than a B both of which are better than a C. Any successful argument against Transitivity must outweigh this evidence.

The First Counterexample: Nine Bad Headaches

The first counterexample consists in nine possibilities:

J: 5 minutes: an agonizing, excruciating migraine headache.

K: 10 minutes: a pounding migraine headache somewhat less bad than the headache in J.

L: 20 minutes: a hideous headache somewhat less bad than the headache in K.

M: 40 minutes: a terrible headache somewhat less bad than the headache in L.

N: 90 minutes: a dreadful headache somewhat less bad than the headache in M.

O: 3 hours: a headache somewhat less bad than the headache in N.

P: 6 hours: a headache somewhat less bad than the headache in O.

Q: 12 hours: a headache somewhat less bad than the headache in P.

R: 1 day: a headache somewhat less bad than the headache in Q. Its pains are slightly worse than temporary unconsciousness.

As we move down the alphabet, the possibilities get intrinsically worse because having a painful headache is worse than having a headache somewhat more painful but for only half as long. The pains in R are still bad: each moment is worse than nothing. Were those pains less bad, then R might not be worse than Q: R's pains would be less intense than Q's, and since the conscious states in R would not be worse than neutral, its extra duration would not count against it. Nonetheless, R is not worse than J.

Transitivity is violated because the path from J to R consists in nothing but steps for the worse.

One can construct a similar counterexample involving pleasure. J, five minutes of the best sexual pleasures, is not worse than R, one day of lousy sex, each moment being barely better than unconsciousness. But R could be reached from J with 8 merely moderate reductions in pleasure intensity, thus ensuring that J is worse than K (with its double duration), K is worse than L (with its double duration), and so on.

Can migraine pain be so agonizing that five minutes of it would be at least as bad as a day of R's milder pain? Can an agonizing headache be transformed into a R-level headache with 8 reductions of intensity? Answering "yes" to both questions entails the denial of Transitivity. Let's consider them in turn.

First, can J be at least as bad as R? Think about that question in terms of this one: would you prefer to have the day-long headache or five minutes of the worst headache pain?

You might prefer the shorter headache so you could return to normal life after five minutes rather than a day. Let's stipulate, however, that if one has a headache for less than a day, then one will spend the rest of the day unconscious, or at least having a day neither better nor worse than if one were.

What Parfit calls the *bias towards the future* suggests another motive for preferring the shorter headache. "Looking forward to a pleasure," Parfit says, "is, in general, more pleasant than looking back upon it. And in the case of pains the difference is even greater."[217] For this reason, Parfit says, we may bring pains into the nearer future and postpone pleasures. If so, then we may want a shorter duration of pain so we can

stop dreading the continuation of the pain at an earlier time. And so, for that reason, we may prefer J to R. But let's assume that dread is either absent in J-R or counterbalanced by some pleasure. Hence, we may ignore the bias towards the future.

People who have had severe migraines can best judge whether a day of R's pain is worse than five minutes of J. Most of the migraine sufferers I've talked to prefer R. So, I answer "yes" to the first question: migraine pain *can* be so awful that five minutes of it would be at least as bad as a day of the milder pain.

To avoid where this argument is headed, some might object, "R is worse than J even though one should choose R when J is the only alternative. R is worse because it compares less favorably than J to some unavailable possibilities." This strategy could be deployed against any comparative judgment. But which of the endless unavailable options are relevant to comparing J and R, if comparing them directly doesn't establish their relative value? And what comparisons could reverse our initial judgment that R is not worse than J? These questions appear unanswerable.[218]

Second, can an agonizing J-level headache be transformed into a milder R-level headache with 8 reductions of intensity? Since duration doubles with each move down the alphabet, each change should be for the worse even if the pain intensity is considerably reduced at each step. 8 reductions, I think, is more than we need. Again we should answer "yes."

Hence, J-R entail that *being intrinsically better than* is not transitive. "More cautiously," says Temkin, "one may decide that the concept of "better than" is limited in scope, and that for [apparent counterexamples to Transitivity] one needs another concept for comparing alternatives that is

similar in meaning, but intransitive."[219] But one can use an intransitive concept in all of one's comparative, normative judgments—why multiply concepts?

The Second Counterexample: Long Periods of Pain

The second counterexample consists in the possibilities A-Z, each of which involve a single person's experience:

A: 1 year of excruciating agony.

B: 100 years of pain slightly less intense than that in A.

C: 10,000 years of pain slightly less intense than that in B.

D: 1 million years of pain slightly less intense than that in C.

. . .

Y: 10^{48} years of pain slightly less intense than that in X.

Z: 10^{50} years of pain slightly less intense than the pain in Y. The pains in Z are slightly worse at each moment than unconsciousness.[220]

If you doubt that 25 slight reductions in intensity could turn A's unpleasures into Z's, replace "slightly less intense" with "somewhat less intense" or "A-Z" with "1-100."

This counterexample retains its force even if one amplifies A-Z in various ways. To keep matters simple, one should amplify A-Z only with details that have little or no effect on intrinsic disvalue.

Z is not worse than A; a tremendously long period of Z's milder pains is not worse than horrible agony for one year. To make this rhetorically compelling, I would now describe a method of torture that would conjure up horrible agony. But I'll leave that task to your imagination, if you are tempted to think that Z is worse than A.

Although Z is not worse than A, the example creates a path from A to Z involving only changes for the worse. These changes are for the worse because increasing a pain's duration 100-fold outweighs slightly reducing its intensity. So, the possibilities get worse and worse. but Z is not worse than A. This contradicts Transitivity.

Or does it? According to Transitivity, if A is intrinsically better than B, and B is intrinsically better than C, then A is intrinsically better than C. I interpret "A," "B" and "C" to stand for *any* consistent possibilities, but some philosophers might say, "Evaluative concepts such as *better than* are essentially practical, so we should restrict Transitivity to possibilities that might bear on action." Such a restriction runs counter to the spirit of theoretical philosophy; but anyway, showing that A-Z violates the unrestricted version of Transitivity would, in various ways, support the first counterexample which *does* apply to the restricted version.

Two worthwhile objections don't say how the example goes wrong—just that it must. According to the first, "Sorites arguments—and this is one—are known to be unsound." Sorites arguments appeal to a series of changes that individually make no difference. For example, having one hair fewer makes no difference to whether someone is bald. But in this example, each change matters: each succeeding possibility is worse; each change in intensity makes the pain worse. So, the first counterexample is not of the Sorites type.[221]

According to the second undiagnostic objection, "This example can't be assessed; our judgments about such bizarre lives can't be trusted." But each comparison involves just two factors: duration and pain intensity. And most of us have strong beliefs about each comparative judgment; the questions don't strike us as too bizarre or difficult to answer correctly. So, this objection is too quick. But the next objection tries to explain more precisely why our intuitions fail.

Two Diagnostic Objections to the Second Counterexample

(i) "Z *is* worse than A, although A seems worse because we have difficulty conceiving how much disvalue can accumulate, bit by bit, over 10^{50} years." But even after reflecting on Z's length—also bearing in mind that Z's pain doesn't worsen over time—I still don't prefer A to Z. In fact, I strongly prefer Z to A. Should I? The following argument emphasizes the ratio of time spent suffering in A to time spent in the mild pain of Z (where "mild pain" means "pain at each moment slightly worse than temporary unconsciousness"):

1. One million years of mild pain are worse than three seconds of agony; (premise)
2. A period of pain is exactly as bad as the summed badness of its sub-periods of three seconds or more; (premise)
3. One year of agony is exactly as bad as the summed badness of each of its 10 billion three-second sub-periods; (from 2)

4. 10^{16} years of mild pain is exactly as bad as the summed badness of each of its 10 billion million-year sub-periods; (from 2)
5. 10^{16} years of mild pain are worse than one year of agony; (from 1, 3 and 4)
C. 10^{50} years of mild pain are worse than one year of agony (Z is worse than A). (from 5)

All of the derivations in this argument are valid. So, if the argument is unsound, then either 1 or 2 is false. Therefore, the argument proves that the following triad is inconsistent:

1. One million years of mild pain are worse than three seconds of agony.
2. A period of pain is exactly as bad as the summed badness of its sub-periods of three seconds or more.
~C. 10^{50} years of mild pain are not worse than one year of agony (Z is not worse than A).

At least one of these is false. But which? Any member of the triad can be denied by appealing to the other two, so let's look for independent evidence.

Thoughtful people who know suffering overwhelmingly prefer an indefinitely long period of Z's pain to a year of agony. So, ~C is well-supported independently of the trilemma. (In fact, a slightly stronger thesis is also well-supported: that Z is preferable to A.)

According to Moore, "two bad things . . . may form a whole much worse than the sum of badness of its parts."[222] Is a period of pain worse than the summed disvalue of its temporal parts? A period of pain might be worse than the summed disvalue of its parts that last only a millionth of a

second; such parts may be too brief to have disvalue. But 2 says, "A period of pain is exactly as bad as the summed badness of its sub-periods *of three seconds or more*," and one can suffer in three seconds. Is 2 true? 2 seems true for the following reason: a period of pain is bad because of how it feels or what it's like; and how it feels or what it's like consists in how its subperiods feel or what they're like. 2 does presupposes the controversial thesis that badness can be summed, but we can revise 2 as follows:

> 2R: The reasons why a period of pain is bad are exhausted by the reasons why each of its subperiods of at least three seconds are bad.

2R can replace of 2 in our discussion. And the same reasoning that supports 2 supports 2R. Hence, 2R is well-supported independently of the trilemma.

Now consider 1: "Three seconds of agony are preferable to a million years of mild pain." Is this true of the worst agony? I find the issue too hard to decide intuitively. Many other people I know agree.[223]

According to Gurney, "torture" is "incommensurable with moderate pain"—so, any duration of torture is worse than any duration of moderate pain.[224] Perhaps "moderate pain" is roughly at the level of Z's pains. If so, Gurney rejects 1. Sidgwick responds as follows:

> [Gurney's] doctrine . . . does not correspond to my own experience; nor does it appear to me to be supported by the common sense of mankind:—at least I do not find, in the practical forethought of persons noted for caution, any recognition of the danger of agony such that, in order to avoid the smallest extra risk of it, the greatest

> conceivable amount of moderate pain should reasonably be incurred.[225]

Sidgwick's reply assumes that, if an indefinitely long period of Z's pain is preferable to a short duration of agony, then one should choose that long period of pain over a *slight* chance of the agony. Is this true? Let "F" represent behavior that has a 100% chance of causing an indefinitely long period of Z's pain, while "G" represents behavior that has a .1% chance of causing 3 seconds of horrible pain followed by normal life and a 99.9% chance of being followed by normal life. Is F wiser than G, on Gurney's view? Perhaps not: someone who has a "normal life" might occasionally enjoy ecstasy, which might counterbalance the possibility of three seconds of agony. But even if F were wiser than G, even cautious persons would probably opt for G. None of us want certain pain, and we're all prone to disregard small chances.

Hence, no solid, independent evidence counts for or against 1. And 1 conflicts with the conjunction of ~C and 2R, which independent evidence supports. So, 1 is false; Z is not worse than A.

(ii) "At least one of our other comparative judgments is false; some period of slightly more intense pain isn't worse than a period of slightly less intense pain that lasts 100 times longer." This thesis is tempting only when the lesser pain is mild: perhaps a period of pain isn't worse than a period of slightly lesser, mild pain that lasts 100 times longer. However, no one believes that a severe pain's becoming slightly less intense outweighs its duration increasing 100-fold. So, perhaps W is worse than V, but A is certainly not worse than B.

This objection, though tempting, is utterly implausible given that even Z's pains are worse than temporary unconsciousness. If pain is that bad or worse, then reducing its duration by 99% would obviously be worth a slight increase in intensity. So, Transitivity fails.

Again, note that one can easily construct a similar example involving pleasure. A is a year of ecstasy, B is 100 years of pleasure slightly less intense, C is 10,000 years of pleasure still slightly less intense, and so on. Z is 10^{50} years of experience slightly better than temporary unconsciousness at each moment—the pleasures of muzak and potatoes.[226] B is better than A, C is better than B, and so on, but Z is not better than A. This contradicts Transitivity.

Rational Choice

Transitivity's failure entails that, for some possibilities, X is better than Y, Y is better than Z, but X is not better than Z. To simplify the coming discussion, I'll adopt the stronger thesis (which I believe) that for some possibilities, X is better than Y, Y is better than Z, and X is *worse* than Z.[227] This thesis may seem to run afoul of the following "money-pump" objection[228] (I am not quoting):

> On your view, an informed, rational person may prefer X to Y, Y to Z, and Z to X. But such preferences lead to irrational behavior. For example, in some circumstances, she would pay a small amount to trade X for Z, then pay a small amount to trade Z for Y, then pay a small amount to trade Y for X—the same X she started with. Then

> she would lose more money in the same way, again trading X for Z, Z for Y and Y for X.[229]

But, on my view, an informed, rational person won't prefer X to Y, Y to Z, and Z to X in a sense that commits her to such insanity.[230] She would prefer X to Y were they alone relevant to which of them she should prefer—X is, after all, better than Y. And, similarly, she would prefer Y to Z and Z to X. But when all three possibilities are available, she won't prefer X to Y, Y to Z, and Z to X, to avoid being money-pumped. Hence, a rational person will not always prefer what is better. Given a choice among X, Y and Z, either Z's presence bears on whether she should prefer X to Y, or X's presence bears on whether she should prefer Y to Z, or Y's presence bears on whether she should prefer Z to X. For example, starting off with X, she may consistently (and wisely) refuse to trade it for Z, even knowing that Z is better. Unusual cases often need unusual treatment.

But what should she do: stay with X; trade X for Z and then stay put; or trade X for Z, then trade Z for Y and then stay put?

For cases that violate Transitivity, one lacks a powerful reason to justify any particular course of action: one cannot say *this choice results in a state of affairs that is best.* But there may be reasons, regarding the possibilities themselves, to prefer some choices over others. For example, perhaps we have special reason to avoid a possibility with *much* greater pains than occur in some other option. (So, in the second counterexample, we have special reason to avoid A, given that Z is available.) If such principles entail that one of X, Y and Z is most rationally targeted, then an informed, rational person could be money-pumped only until she lands on that choice. If, alternatively, there are no reasons, regarding the

possibilities themselves, to prefer some choices over others, then all available choices are equally wise (which, of course, doesn't entail that all possibilities are equally good). If so, then she could not be pumped of a cent: she would keep what she originally has (whether X, Y or Z), rather than giving up a quarter to trade for something better.

Transitivity may persist in our reasoning as a rule of thumb, as exceptions to it are rare. Avoiding the principle altogether would require making fewer inferential judgments. Instead of inferring that A is better than C (given that A is better than B and B is better than C), we would need to compare A and C directly.

So, I see no reason to accept Parfit's view, reported by Temkin, that Transitivity's failure would entail skepticism about practical reasoning.[231]

Conclusion

According to Transitivity, if A is intrinsically better than B, and B is intrinsically better than C, then A is intrinsically better than C. I offered general reasons, in the first section, for thinking that Transitivity is not analytically true. Transitivity's support, instead, is merely inductive. But if so, then a sufficiently compelling counterexample can overturn it.

I offered two counterexamples to Transitivity. The first consists in nine headaches, ranging from five minutes to a day. This example, I think, outweighs Transitivity's inductive support all by itself. And it may constitute my best case against Transitivity as a conceptual truth; false propositions aren't true conceptually.

The second counterexample consists in 26 lives, ranging from a year to 10^{50} years. On one objection to it, Z is worse than A: 10^{50} years of mild pain are worse than a year of agony. I discussed this objection in terms of the following inconsistent triad:

1. One million years of mild pain are worse than three seconds of agony.
2. A period of pain is exactly as bad as the summed badness of its sub-periods of three seconds or more.

~C. 10^{50} years of mild pain are not worse than one year of agony.

On this objection, 1 and 2 are true. I accept 2 in revised form. But I reject 1 and accept ~C, that one year of agony is at least as bad as 10^{50} years of pain slightly worse than unconsciousness. ~C seems obviously true to many of us when we reflect on our worst pains. The best response to the second counterexample, in my opinion, is the following:

> The fact that ~C seems to be true counts only minimally in its favor, since we have trouble imagining how much badness accumulates over Z's 10^{50} years. Such modest evidence is less than the inductive evidence for Transitivity. The second example, therefore, fails on its own to refute Transitivity.

I think that reflection supports ~C more than minimally, but I have no argument to offer. Because this issue is so difficult, the first counterexample is better than the second. But they work together against Transitivity. Together they outweigh the inductive evidence that supports that principle. "Being intrinsically better than," I conclude, is nontransitive, as is "being hedonically better than."

Since Transitivity fails, so does the linear picture of value: possibilities cannot all be represented as points on a line, with the intrinsically better ones represented to the right of those intrinsically worse; no complete ordinal ranking of states of affairs exists. If so, then, of course, no complete cardinal ranking exists; possibilities can't all be quantified so that the better ones have higher numbers than those worse (never mind other possible constraints on cardinality, for example, that equal numerical differences always correspond to equal differences in value).[232] Moreover, my counterexamples show that the hedonic value of possibilities can't be similarly conceived, even though philosophers often think of hedonic value as a paradigmatic (or as the sole) type of value that can be precisely quantified.[233] If hedonic value escapes neat and comprehensively quantification, this increases the chances that other spheres of value do as well (provided there are others).

Philosophers often discuss the ontology of value rather dismissively, bringing it up only to deny that "value" has a referent.[234] But most of us, prior to argument, are firmly wedded to Transitivity; moreover, quantitative expressions pervade ethical theory, for example, "average utility" and "the principle's weight." Even "more value" and "less value" connote quantity. This suggests that, at some level, philosophers take the linear view of value seriously—not just as a convenient heuristic, but as representative of value itself. For that reason, it may be worthwhile to note that Transitivity's failure entails that value itself (if such an item exists) is neither precisely quantifiable nor robustly isomorphic with items that are.

Chapter 6: How to Assess Comparative Hedonic Value

How can we accurately compare the hedonic value of states of affairs? Two general methods seem promising.

Ideal Observer Imitation

Some states of affairs include many people. This complicates assessing hedonic value, but many philosophers, including myself, accept the following, simplifying principle:

> The Prudence Principle: Alternative B is hedonically better than alternative C if it would be better for one to have all the experiences in B rather than all those in C (having, within each alternative, all the experiences of one sentient being, then all the experiences of another, and so on, in random order of creature).

On this principle, two pleasurable lives are hedonically better than one life of slightly greater pleasure (just as, within a single life, two minutes of pleasure are hedonically better than one minute of slightly greater pleasure). Average utilitarians might disagree, but average utilitarianism is untenable, as I explained in chapter 4.

The Prudence Principle sanctions translating multiperson comparisons into single person comparisons; instead of asking, "Is it hedonically better for Cindy and Bill, or for Jim and Carol, to be at the

Braves game?" we may ask, "Would it be better for one to have all of Cindy-and-Bill's relevant experiences or all of Jim-and-Carol's?" Translating like this conflates lives, but permissibly: the distinction between persons, though relevant to rights, desert, fairness and equality, is irrelevant to hedonic value.

The Prudence Principle supposes that having the experiences of several people can be better than some other set of experiences. But, according to Nagel, several lives being amalgamated into one is "unimaginable."[235] Yet we can imagine one person having experiences qualitatively identical to those had by many others; or, if my having Cindy's experiences (as though I were Cindy) followed by Bill's experiences (as though I were Bill) seems to violate the idea that *I* have these experiences, then modify the Prudence Principle so that I would merely need to have (or imagine having) pleasures and unpleasures just like Cindy's and Bill's in length and intensity.

In one state of affairs, let's suppose, Persephone and Hades live forever. Can we compare this alternative to any other, using the Prudence Principle? I can't have all the Persephone-and-Hades experiences by first having all the experience of one, then having all the experiences of the other (as the Prudence Principle would require), for I'd never get to the second set of experiences. Could I have all their experiences by a different route? If my consciousness split into a Persephone branch and a Hades branch, presumably I couldn't survive as both. However, I could have Persephone's experiences on even days, odd days being devoted to Hades. So, the Prudence Principle, if properly reformulated, could apply to such cases.

Given the Prudence Principle, many philosophers would accept something like the following ideal observer principle: "B is hedonically better than C if an ideally knowledgeable and sensitive person prefers to have B's experiences to C's." This principle makes two controversial assumptions. First, why believe that an ideally knowledgeable and sensitive person would always have the right preferences? Perhaps no decision procedure or perfect neural program for forming such preferences exists, either because some features of hedonic value are closed to all minds or because knowing some hedonic facts are incompossible with knowing others. A more modest principle avoids this worry: "The best possible evidence for B's being hedonically better than C is that an ideally knowledgeable and sensitive person prefers B's experiences to C's." Second, why should all ideally knowledgeable and sensitive persons agree? Hare believes they would but offers no argument.[236] Mill says, "On a question which is the best worth having of two pleasures . . . the judgment of those who are qualified by knowledge of both, or, if they differ, that of the majority of them, must be admitted as final."[237] So, let's revise the principle further: "The best possible evidence for B's being hedonically better than C is that *most* ideally knowledgeable and sensitive persons prefer B's experiences to C's." A problem remains, however, crippling to my aim. I asked, how can we accurately compare the hedonic value of states of affairs? Ideal observer principles can't help us compare alternatives: we don't know any ideal observers to consult, and wondering which alternative an ideal observer would prefer is like wondering which quantum theory an ideal physicist would believe: it just postpones the hard work of determining which is preferable.

A more serviceable principle is as follows:

> **Ideal Observer Imitation (IOI)**: Think hard about the pleasures and unpleasures in alternatives B and C qua experience. If you would prefer to have all the experiences in B (having all the experiences of one sentient being, then all the experiences of another, and so on, in random order of creature), then conclude that B is hedonically better than C (although, if you feel uncertain about this preference, then be uncertain of your conclusion); if you can't decide which you prefer or if you are indifferent between B and C, then conclude that they are the same in hedonic value (or, if the question seems difficult, roughly the same).[238]

This principle begins, "Think hard about the pleasures and unpleasures in alternatives B and C qua experience." But just saying "think hard about X" gives little guidance about how to think. IOI, as it stands, counsels meditating on pleasures and unpleasures without applying rules, but Sidgwick's confession shows how unhelpful such advice is:

> Now, for my own part, when I reflect on my pleasures and pains, and endeavor to compare them in respect of intensity, it is only to a very limited extent that I can obtain clear and definite results from such comparisons, even taking each separately in its simplest form:—whether the comparison is made at the moment of experiencing one of the pleasures, or between two states of consciousness recalled in imagination. This is true even when I compare feelings of the same kind: and the vagueness and uncertainty increases, in proportion as the feelings differ in kind.[239]

Introspection reveals nothing about how our minds make even the simplest comparative hedonic judgments. Which feels worse, a slap on the back or a blow to a kneecap? I know I prefer the slap, but how I form that preference is hidden in a neural "black box," opaque to consciousness. Knowing how

brains function at such moments, as well as knowing the neural correlates of sentient experience, could aid this inquiry. But, for now, we face the same limitations as Sidgwick.

IOI yields even more indefinite results as the comparisons get more complex. How well, using IOI, could we answer the following questions posed by Rawls?

> Are we to choose a brief but intense pleasant experience of one kind of feeling over a less intense but longer pleasant experience of another? Aristotle says that the good man if necessary lays down his life for his friends, since he prefers a short period of intense pleasure to a long one of mild enjoyment, a twelvemonth of noble life to many years of humdrum existence. But how does he decide this? Further, as Santayana observes, we must settle the relative worth of pleasure and pain.[240]

IOI provides few resources for answering such questions.

Hedonic Figuring

Ethicists, especially utilitarians, have wanted to quantify hedonic value. "Sum up all the values of all the *pleasures* on the one side," says Bentham, "and those of all the pains on the other."[241] According to Mill, "the truths of arithmetic are applicable to the valuation of happiness, as of all other measurable quantities."[242] And similarly, Sidgwick says, "all pleasures . . . are capable of being compared quantitatively with one another and with all pains . . .[243] If so, then B is hedonically better than C

just in case B contains the greater surplus of pleasure over pain. Given a thorough knowledge of neurophysiology, on this view, we might someday measure pleasure like sugar or to count pleasures like sugar granules; but for now, the best strategy for quantifying hedonic value might be something like the following:

> **Hedonic Figuring (HF)**: Imagine that someone in horrible agony feels gradually better, improving at a constant rate, until she enjoys intense pleasure. Identify a moment in this process when her conscious state seems neither better nor worse than unconsciousness. Assign that state the number 0. Now work forwards and backwards in time, assigning her state five seconds after 0 the number 1 and five seconds before -1. Then assign her state ten seconds after 0 "2" and ten seconds before 0 "-2," and so on. These numbers represent base experience intensity. To confirm that the subject felt better at a constant rate, verify that you are indifferent among the following alternatives: having 0 for time t; having 1 and -1 for .5t apiece; having 2 and -2 for .5t apiece, and so on.
>
> To determine the hedonic value of any state of affairs, sequentially order its experiences. Assign a number to represent the intensity of the first identifiable moment of experience by comparing it with the base. For example, if the first identifiable moment seems closest in intensity to the 8 pleasure, call it 8. Continue down the line, assigning a new number each time hedonic intensity changes. Once you have finished, multiply each number by the duration of that experience, then add the products. The sum represents the alternative's overall hedonic value. One alternative is better than another just in case its number is higher.[244]

Implementing this principle looks frightfully difficult. How does one originally grasp what unpleasure -21 feels like? Does the person implementing the procedure need to be the base subject? (How much would

that help?) And how does one remember, when assessing new experiences, what the 9 pleasure felt like? Moreover, HF requires IOI both to verify that the base intensities change gradually and to assign numbers to new experiences. And IOI yields "clear and definite results," as Sidgwick wrote, "only to a very limited extent." It would be glib to call these "merely technical" problems for HF, since we don't know to what extent we can solve them.

HF assumes that an alternative's hedonic value can be precisely represented with a number. Must moral reality be determinate? Often I am unsure whether two pleasures are equal in value. My subjective uncertainty might reflect objective indeterminacy. Perhaps, for some comparisons, none of the following statements is true:

(i) B is hedonically better than C.
(ii) C is hedonically better than B.
(iii) B and C are of equal hedonic value.

Or, for some alternatives, perhaps the following is true:

(1) M and N are of equal hedonic value.
(2) P is hedonically better than N.
(3) P is not hedonically better than M.

For example, suppose I am indifferent between tasting french toast (= M) and tasting grits and eggs that I cook (= N). And suppose that, in fact, these experiences are equally valuable, so 1 is satisfied. Of course, tasting my mom's grits and eggs (= P) beats tasting mine, so 2 is satisfied. Despite

this, my mom's grits and eggs might not taste better than french toast, so 3 might be satisfied. (1)-(3) might all be true because M and N differ phenomenologically. And if they are, then numbers cannot be coherently assigned to these alternatives.[245]

HF entails, absurdly, that having a pleasure at level 1 forever is exactly as good as having a pleasure at level 20 forever, since 1(infinite duration) = 20(the same infinite duration), whatever the order of infinity. Some would say, "Transfinite value theory may produce strange results, but so does transfinite mathematics." Transfinite mathematics does produce strange results—for example, that the integers are no more numerous than the even integers—but many people *believe* such results, based on mathematical argument. But what arguments support the crazy thesis that all eternally pleasurable lives are equal in value? Such arguments are likely to assume that value is quantifiable; but that assumption should be rejected before we accept the conclusion.[246]

(Traditional theists, incidentally, might want to accept the crazy thesis to explain away suffering. "Sufferers," they might say, "will go to heaven and enjoy blissful pleasure. After a while, this pleasure will counterbalance the pain they once felt. At that moment, their lives will have been, on the whole, hedonically neutral. And from that moment on, they will enjoy the infinite value of eternal bliss. Hence, they will have maximally good lives from a hedonic point of view.")

HF also entails that pleasures and unpleasures can't differ *lexically*. If unpleasure I is sufficiently more intense than unpleasure S, then S is *lexically superior* to I, meaning that no finite duration of S is as bad as some relatively short duration of I. Let's say, somewhat arbitrarily, that S is lexically superior to I just in case a year or more of I is worse than any

finite duration of S. On HF, a sufficient duration of a -1 unpleasure would yield a lower value than a year of agony (a -200 unpleasure, say), even though the year of agony is worse. My argument for lexical differences[247] relied on something like IOI: the shorter duration seems worse to those who think vividly about agony and mild unpleasure qua experience. I prefer IOI on this question to HF because HF merely asserts a single-scaled system of calculation, while IOI offers evidence without prejudging the issue.

HF might also miscompare pleasures and unpleasures less far apart in intensity. For example, even if some duration of -1 is worse than -20 for a year (so, they don't differ lexically), that duration might need to be longer than 20 years and a day. Perhaps, as intensity difference increases, so does the additional time needed for the lesser unpleasure to be worse than the greater. I don't know whether this is so.

IOI and HF, though imperfect, offer good starting points. Now I will offer principles and advice to supplement them.

Value Principles

(1) More intense pleasures are always better qua experience than less intense pleasures.

HF presupposes (1) by enjoining us to assign higher numbers to pleasures of greater intensity. IOI, on the other hand, tells us merely to

think hard, leaving open whether less intense pleasures can be better qua experience than more intense pleasures. So, (1) supplements IOI.

I'll define *qualitative hedonism* as the view that pleasures differ not only in intensity and duration, but also in a third normatively significant way—call it "dignity." So-called "higher" pleasures are better in terms of dignity than "lower" pleasures. On this view, less intense pleasure can be better qua experience than more intense pleasure. Hutcheson exemplifies this tradition:

> In comparing pleasures of different kinds, the value is as the duration and dignity of the kind jointly. We have an immediate sense of a dignity, a perfection, or beatifick quality in some kinds, which no intenseness of the lower kinds can equal, were they also as lasting as we could wish. No intenseness or duration of any external sensation gives it a dignity or worth equal to that of the improvement of the soul by knowledge, or the ingenious arts; and much less is it equal to that of virtuous affections and actions.[248]

On Hutcheson's lexical view, no "intenseness or duration" of a lower pleasure can equal the "worth" of a higher pleasure. But dignity doesn't trump duration and intensity on all qualitative hedonist views; even if higher pleasures are better, in a way, than lower pleasures, a sufficient duration of an intense lower pleasure might be preferable to a single appreciative reading of Ezra Pound's "Hugh Selwyn Mauberley."

Which pleasures are considered higher, and which lower? Here are typical suggestions:[249]

Higher	**Lower**
human pleasures	animal pleasures
mental or intellectual pleasures	bodily or sensory pleasures
moral pleasures	amoral pleasures
generalized pleasures	localized pleasures

So, human pleasures are better, in a way, than animal pleasures, according to some philosophers; mental pleasures are higher than bodily pleasures; and so on.

These pleasure categories, as listed, are vague. We can identify some clear cases: Mikhail Tal's chess combinations afford higher, mental pleasure; "throwing the crockery around when drunk," to use Moore's example,[250] affords lower pleasure, although it's not obvious what type of lower pleasure accompanies such activity. Many cases, however, are not clear. Is the pleasure most people get from drinking wine mental, bodily or both? Do I always get human pleasure from my friends' company, given that nonhuman primates have friends? Two questions now arise: Which pair of categories should we focus on? How can we make them more precise? Finding an adequate ground for qualitative hedonism would be the first step toward answers. Let's examine Mill's influential argument.

First, as we have seen, Mill says,

> On a question which is the best worth having of two pleasures, or which of two modes of existence is the most grateful to the feelings, apart from its moral attributes and from its consequences, the judgment of those who are qualified by knowledge of both, or, if they differ, that of the majority of them, must be admitted as final.[251]

Mill should now say, "Most competent judges prefer higher pleasures to more intense lower pleasures; so, qualitative hedonism is true." However, what he says concerns much more than pleasure:

> Now it is an unquestionable fact that those who are equally acquainted with and equally capable of appreciating and enjoying both do give a most marked preference to the manner of existence which employs their higher faculties. Few human creatures would consent to be changed into any of the lower animals for a promise of the fullest allowance of a beast's pleasures; no intelligent human being would consent to be a fool, no instructed person would be an ignoramus, no person of feeling and conscience would be selfish and base, even though they should be persuaded that the fool, the dunce, or the rascal is better satisfied with his lot than they are with theirs.[252]

In these pages Mill uses the language of pleasure, but his real topic is happiness or well-being ("manner of existence," as he puts it). Few of us, he says, would become beasts, fools, ignoramuses, dunces or rascals in exchange for the pleasures of their ilk. And, indeed, few would. But many non-hedonic values influence this judgment. We prize many things that beasts and fools are short on: knowledge, virtue, grace, success, intelligence, compassion, wisdom, liberty, personal independence and power.[253] So, the intuitions Mill elicits concern greater well-being, not greater hedonic well-being.

Keeping Mill's first premise, and substituting my second premise for his, here is the familiar 'Millean' argument:

(1) If most people who have had pleasures J and K prefer J qua experience, this is decisive evidence that J is better qua experience.
(2) Most people who have had both higher and lower pleasures prefer higher pleasures qua experience, even when they are less intense.
(C) Qualitative hedonism is true.

Were this argument sound, we could try to identify the higher and lower pleasures by surveying clearly competent judges. Is it sound? Let's examine its premises.

The Competent Judge Criterion (Premise 1)

This premise is too strong: most people's preferring J to K qua experience does not provide "decisive" evidence that J is better qua experience; people might mistakenly prefer higher to lower pleasures, having felt both, for at least two reasons.

First, learning to appreciate higher pleasures can lessen one's enjoyment of lower pleasures. Acquiring a taste for fine wine, for example, can diminish one's enjoyment of Beaujolais. Someone who no longer appreciates lower pleasures may underrate them.[254]

Also, some people, at various levels of awareness, may dismiss lower pleasures, snobbishly priding themselves on having superior taste.

I can revise the first premise to assuage these worries:

(1*) If the majority of *unpretentious* people who *fully appreciate* pleasures J and K prefer J qua experience, this is decisive evidence that J is better qua experience.

But the first premise, in both forms, seems plausible only assuming that J would be better than K qua experience *because J is more intense than K as a pleasure*. On that reading, the first premise is plausible: we may have no better recourse when comparing pleasures in terms of intensity than to poll those who have had the relevant experiences. However, in the Millean argument, the issue is whether J is better than K qua experience because J is more dignified. When Hutcheson says, "We have an immediate sense of dignity, a perfection, or beatifick quality in some kinds," he implies that one can easily introspect the intrinsic superiority of higher pleasures, just as one introspects intensity. However, no properties of pleasures *obviously* bear on their intrinsic value except those affecting intensity. So, dignity is not obviously good; on this issue, we need reasons, not tallied opinions. To advance discussion, qualitative hedonists should describe dignity and explain why it's desirable.

What Do Competent Judges Prefer? (Premise 2)

Premise 2 can be revised in accordance with 1*:

(2*) If the majority of unpretentious people who fully appreciate both higher and lower pleasures prefer higher pleasures qua experience, even when they are less intense.

We could try to assess 2* by surveying unpretentious people who have recently enjoyed higher and lower pleasures. But those who voice a

preference for higher pleasures might still provide poor data. First, despite being told to consider the pleasures qua experience, they might overrate higher pleasures qua experience because of their extrinsic properties: higher pleasures, it's hard to forget, often signal understanding, friendship and virtue, in contrast to lower pleasures; mental pleasures, as Mill says, have greater permanency, safety and uncostliness than bodily pleasures;[255] finally, even an unpretentious person might be influenced by the effort or skill it took to enjoy the higher pleasure. Second, higher pleasures tend to be less intense as sensations than lower pleasures (for example, the pleasures of philosophy are less intense than the pleasures of skydiving), so respondents might compare lower pleasures to higher pleasures which they believe to be less intense as pleasures but are really less intense only as sensations. And then, again, the data would be tainted.

For these reasons, surveying competent judges would support 2* only if a significant majority of respondents say they prefer the higher pleasures. And we have no reason to think they would.

So, both premises are weak; the Millean argument fails. We should reject qualitative hedonism because we lack evidence for a higher/lower distinction among pleasures.

(2) Some pleasures and unpleasures differ lexically.

This principle helps to implement IOI but corrects HF.

Pleasures sufficiently apart in intensity *differ lexically* in the sense that a year of the greater is better than any finite duration of the lesser.[256] In this section, my remarks about pleasure apply, with simple changes of

phrase, to unpleasure. How far apart in intensity must such pleasures be? Since I can think of nothing to say except "very," I'll discuss only the clearest case, of ecstasy and mild pleasure.

How long must ecstasy last to be better than any finite duration of mild pleasure? In chapter 5, I committed myself to answering "only three seconds," based on the following argument: fifty years of ecstasy is better than any finite duration of mild pleasure;[257] however, this wouldn't be so if some finite duration of mild pleasure were as good as three seconds of ecstasy, for then 500 billion times that duration of mild pleasure would be as good as fifty years of ecstasy; hence, even three seconds of ecstasy are better than any finite duration of mild pleasure.

What is the practical upshot of lexicality? According to Sidgwick,

> It is not absolutely necessary to exclude the supposition that there are some kinds of pleasure so much more pleasant than others, that the smallest conceivable amount of the former would outweigh the greatest conceivable of the latter; since, if this were ascertained to be the case, the only result would be that any hedonic calculation involving pleasures of the former class might be simplified by treating those of the latter class as practically non-existent.[258]

But mild pleasure cannot always be ignored when ecstasy is present. For example, imagine an hour in which you experience equally intense agony and ecstasy for five minutes apiece. That hour could still be better than neutral if you also enjoy mild pleasure. Or, if B and C each include the same duration of ecstasy, then how they compare may turn on which has better mild pleasures.

Practical Strategies

These strategies supplement IOI but are superfluous to HF (if perfectly implemented).

(3) Simplify comparisons by "canceling out" experiences.

Judges of hedonic value can simplify comparisons by "canceling out" experiences of equal value. For example, suppose that B includes the experiences of Simka and Latka, while C includes Barry's and Trudy's experiences. If you are indifferent between having Simka's and Barry's experiences, then compare B and C by comparing the experiences of Latka and Trudy.[259] Also, an alternative can be simplified for comparative purposes by "canceling out" pleasure-unpleasure pairs of equally great value and disvalue. For example, if you are indifferent between having some neutral experience and having Simka's unpleasure at noon followed by her pleasure at midnight, then suppose, for purposes of comparison, that Simka had neither experience.

If experiential variety as such were hedonically desirable, then canceling out pleasure-unpleasure pairs would lessen an alternative's hedonic value[260] (unless it included other experiences similar to the canceled ones); however, I find experiential variety desirable merely because it makes life more intriguing or less boring: the pleasures of memory and anticipation are greater if one's life is varied. Also, pleasures become duller upon repetition (who wants to eat the same meal twice in one day?). In each case, variety is valuable for making pleasures more intense.

Simplifying in these ways might distort some comparisons for the following reason. Just as "being better than" is not transitive, so is "being exactly as good as," for similar reasons. Let me pause to argue this. A is fifty years of ecstasy; B is a longer duration of lesser pleasure such that B is exactly as good as A; C is a still longer duration of still lesser pleasure such that C is exactly as good as B; and so on, until Z, which is a *very* long duration of mild pleasure, where Z is exactly as good as Y. Z, however, is worse than A, so "being exactly as good as" is not transitive.

(Earlier I speculated that some pleasures might be neither better, nor worse, nor equal in value to some others. If the above argument fails because no duration of B is *exactly* as good as A, then perhaps the only alternatives equal in hedonic value are qualitatively identical. If so, simplifying techniques would be problematic, as would the project of making many precise hedonic comparisons.)

If "exactly as good as" is not transitive, then two stretches of pleasure, in different alternatives, might be exactly equal in value yet compare differently to a third stretch; also, a pleasure-unpleasure pair of no overall value or disvalue might compare significantly to some other experience. So, canceling out either pair might alter the alternatives such that their overall comparative status changes; but I don't know whether this is so.

(4) Compare representative portions of alternatives.

Some alternatives may be assessed by comparing representative portions of them. In the simplest case, to compare uniform stretches of experience, compare any substretches proportional in duration. For

example, to compare two years of uniform pleasure to one year, compare two minutes of the former to one minute of the latter.

With more complicated alternatives, deciding which portion is representative can be as difficult as assessing the value of the whole. Moreover, lexical differences may complicate using the representative method. For example, suppose that B = 24 hours of mild, moderate and intense pleasures, while C = a much longer duration of fairly mild pleasure. B is hedonically better than C because B includes pleasures lexically superior to those in C; however, an average pleasure in B would be moderate, and a few minutes of moderate pleasure are worse, on my view, than a much longer period of fairly mild pleasure.

Mistakes to Avoid

(5) Don't be influenced by features extrinsic to experience.

This restates IOI's admonition to consider pleasures and unpleasures qua experience. Should some pleasures be assessed, not qua experience, but in light of a larger context? The following pleasures, some philosophers think, are bad:

(A) pleasures taken in bad intentional objects (such as the belief that someone is suffering);[261]

"Suppose that I perceive or think of the undeserved misfortunes of another man with pleasure. Is it not perfectly plain that this is an intrinsically bad state of mind, not merely in spite of, but because of, its pleasantness?"[262] (Broad) Such pleasures, Zimmerman says, are "inappropriate" or "incorrect" under the circumstances.[263]

(B) pleasures accompanying bad behavior;

"Now since activities differ in respect of goodness and badness, and some are worthy to be chosen, others to be avoided, and others neutral, so, too, are the pleasures; for to each activity there is a proper pleasure. The pleasure proper to a worthy activity is good and that proper to an unworthy activity bad; just as the appetites for noble objects are laudable, those for base objects culpable."[264] (Aristotle)

(C) pleasures that depend on false beliefs or cognitive error;

"Suppose I believe, incorrectly, that my father has performed certain magnificent deeds and suppose I take pleasure in what I believe that he has done. . . . Brentano is clear that such a false . . . pleasure is one that it would be better not to have—this because of the intrinsic evil that error involves."[265] (Chisholm)

(D) undeserved pleasures.

"The sight of a being who is not graced by any touch of a pure and good will but who yet enjoys an uninterrupted prosperity can never delight a rational

and impartial spectator. Thus a good will seems to constitute the indispensable condition of being even worthy of happiness."[266] (Kant)

Similarly, one might think that some unpleasures are good.[267] I'll focus on (A), to illustrate how I want to treat all of these cases.

According to (A), pleasures taken in bad objects are bad. What does this mean? According to Lemos, a pleasure's intrinsic value may turn on the value of its intentional object.[268] If so, perhaps (A) means

(A*) Pleasures are intrinsically bad that are taken in bad intentional objects (such as the belief that someone is suffering).

A* assumes that the pleasure's object is intrinsic to the pleasure; for, were it not, it wouldn't affect the pleasure's intrinsic worth. Consider this reductio: were a pleasure intrinsically bad, extrinsically taken in the belief that someone suffered, then a subjectively similar experience would be intrinsically bad, extrinsically taken in the belief that someone prospered.

Are intentional objects intrinsic to pleasurable experiences? Lemos thinks so because he characterizes pleasure too broadly, "as a state of affairs that implies pleasure and no pain."[269] On this view, *any* state of affairs that doesn't entail pain is intrinsic to some pleasure. For example "Mars exists, and Groucho Marx felt pleased" would name a pleasure, so Mars' existing would be intrinsic to that pleasure. (A proper use of "pleasure" occurs *within* Lemos's characterization of pleasure.)

"Pleasure" has different meanings. If someone says, "Swinging a golf club is one of my pleasures," then a pleasure, in that sense, is an activity that causes pleasurable experience. (But the activity needn't be

obviously public; I might also say, "Thinking about my children is one of my pleasures.") However, I use "a pleasure" to refer to an experience. So, I'll ask, "When I take pleasure in someone's suffering, is my belief that someone suffers part of my phenomenology?" Hume, of course, thought that all empirical beliefs have a qualitative feel to them,[270] but this got him into trouble. It seems to me that I could have qualitatively identical experiences taking pleasure in someone's suffering and taking pleasure in someone's getting fired. So, I am inclined to think that intentional objects are extrinsic to phenomenology. Hence, I am inclined to think that A* is false on metaphysical grounds alone. However, I have no theory of belief to offer, so my remarks in this area are superficial.

Someone might object: "But when I am pleased by someone's beauty, isn't my belief that she's beautiful woven into my pleasure, contributing to its being one kind of pleasure rather than another?" I can admit both points (or something like them): just as the fibers in a weave remain distinct, so do my pleasure and the belief it's taken in; just as paintings may be sorted by their cause (who painted them), so pleasures may be sorted by what they're taken in.

I suggest we interpret (A) instead as:

(A**): Pleasure's being taken in a bad intentional object is intrinsically bad. (Here "pleasure's being taken in a bad intentional object" names a state of affairs that entails the existence of both a pleasure and a distinct intentional object.[271])

Why might A** be true? Not because "pleasure's being taken in a bad intentional object" entails that a bad object exists, for it doesn't; if Lex

Luthor believes that Superman suffered, he may take pleasure in Superman's suffering, even if his belief is false. So, some other idea must motivate A**, most likely:

(i) "Pleasure's being taken in a bad intentional object" entails that an intrinsically bad relation exists, namely, an inappropriate way in which a pleasure and a bad intentional object relate.

Supporters of A** will complete their view by affirming one of the following:

(ii) "Pleasure's being taken in a bad intentional object" entails that pleasure exists, a state of affairs that is good; but its value is outweighed by the disvalue of the intrinsically bad relation's existing.[272]
(iii) "Pleasure's being taken in a bad intentional object" entails that pleasure exists, a state of affairs that is either neutral under the circumstances or that contributes disvalue to the larger state of affairs.[273]

Alternatively, one may deny A**, affirming:

(~A**): Pleasure's being taken in a bad intentional object—that state of affairs—is not intrinsically bad.

And, in explanation of this view, one has two options:

(~i) "Pleasure's being taken in a bad intentional object" entails the existence of no intrinsically bad states of affairs.

(iv) "Pleasure's being taken in a bad intentional object" entails that an intrinsically bad relation exists, namely, an inappropriate way in which a pleasure and a bad intentional object relate. However, its disvalue is counterbalanced or outweighed by the value of pleasure existing.

Is A** true; is pleasure's being taken in a bad intentional object intrinsically bad? Carson says, "I am convinced by the arguments of Brentano and others,"[274] but the literature lacks convincing arguments. Philosophers either assert A** without argument; offer A** as intuitively compelling; or offer some combination of (i)-(iii), themselves controversial, to support it.[275] I will argue, against (iii), that pleasure's existing is good even when pleasure is taken in bad objects. This thesis counts in favor of ~A**. However, I won't discuss (i), according to which "pleasure's being taken in a bad intentional object" entails an intrinsically bad relation; so, I won't assess, (ii), according to which A** is true because the inappropriate relation's existing has enough disvalue to outweigh the value of the pleasure's existing.

Is it good for pleasure to exist even when pleasure is taken in bad intentional objects? I'll now revert back to talking about pleasure (rather than the existence of pleasure), so I'll ask: is pleasure good when taken in bad intentional objects? Consider the corresponding question for (D): is pleasure good when undeserved? Goldstein argues:

> The distastefulness in a mass murderer retiring to days rich in *undeserved* pleasure seems to lie not merely in bad effects; his pleasure may seem intrinsically repugnant. Still, there is good even here. We denounce the pleasure *because* of its good; moral

> degenerates luxuriating in health, happiness, or long life seems repugnant for the same reason.[276]

But those who deny that "there is good even here" can respond effectively: "I don't denounce the villain's pleasures because they're good; I denounce them because he enjoys them; and, for that reason, they're not good."

The following argument may be more successful. Intense pleasure is good when taken in a neutral object (such as the thought of one's pet rock),[277] and intense pleasure is good that lacks an object (as in generalized euphoria). Think of such pleasures that do not accompany laudable or culpable behavior; are neither deserved nor undeserved; and stem from neither error nor insight. These pleasures are good merely because they feel good. But, if so, altering their extrinsic properties shouldn't affect *that* value. Hence, pleasures are good irrespective of what they're taken in; of what accompanies them; of what they depend on; and of whether they're deserved.[278]

Nobody, to my knowledge, has *argued* that pleasures lack value when taken in bad intentional objects. Many philosophers, however, find that view compelling, perhaps because "There is no sign more infallible of an entirely bad heart, and of profound moral worthlessness," as Schopenhauer says, "than open and candid enjoyment of seeing other people suffer."[279] But something good may signal something bad: sadists are pernicious people with ill intentions and contemptible characters-traits who repel good folk;[280] however, their pleasures may still be good (although, for utilitarian reasons, nothing about the sadistic mentality should be praised). Moreover, taking *intense* pleasure in bad objects betrays an especially malevolent

character; this explains why the greater malicious pleasures of others seem worse than their lesser malicious pleasures, despite being better.

So, in assessing the value of pleasures, don't be influenced by the extrinsic features stressed in (A)-(D); such features don't affect the value of the pleasure itself. Nor does the timing of an experience affect its value, so don't discount someone's pleasures or unpleasures just because they are far in the future.[281] Finally, beware motivation; we tend to overestimate the disvalue of highly repellent pains, while we underestimate the disvalue of more tranquil unpleasures. (I argued in chapter 2 that motive power is *extrinsic* to pleasures and unpleasures.) For example, many people would incorrectly judge sharp, agitating physical pains as worse than stifling bouts of depression.

(6) Don't use the peak and end rule.

Subjects consistently assess their pleasurable and painful episodes by their zenith intensity and finish.[282] Kahneman *et. al.* call this the "peak and end rule." Privileging an episode's end is a type of temporal bias, while peaks rarely represent episodes as a whole; so, we shouldn't use this rule. However, I don't whether people "thinking hard" about pleasures and unpleasures, as opposed to having them, will tend to make this mistake.

(7) Don't confuse pleasure intensity with sensation intensity.

As Sidgwick says, don't "confound intensity of *pleasure* with intensity of *sensation*: as a pleasant feeling may be strong and absorbing, and yet not so pleasant as another that is more subtle and delicate."[283]

(8) Don't be misled by deep, fitness-enhancing desires.

Humans, like all animals, desire to do many things that augment reproductive success: mating, eating, avoiding being eaten or beaten, and so on. Selective pressures, no doubt, have favored animals for whom "fit" activities give pleasure and "unfit" activities cause pain—that's why eating is more fun than being eaten. But nature should have selected animals who also desire fitness-enhancing activity for its own sake. Such tendencies might stem from the innate instincts of insensate ancestors. If so, then humans may naturally overestimate the hedonic value of satisfying "biologically deep" desires. Some sexual encounters, for example, may be less pleasurable than more acquired enjoyments, even if one's desire for sex is greater.

(9) Be cautious about assessing types of experience you haven't had or don't appreciate; consult others.

A father may underestimate the pleasure his children get from the latest musical fad, which repels him; people who have never been depressed may greatly underestimate those horrors. Recognize what pleasures and unpleasures you don't understand well; talk to those more experienced before assessing them.

(10) Correct for your imaginative weaknesses.

Hedonic assessors often use imagination, which rarely involves mental images.[284] One needs great imagination, Glover says, to appreciate what it is like to be a hungry person in a poor nation. "And even in face to face relationships, some people are better than others at discerning the extent to which their friends' feelings and responses resemble or do not resemble their own."[285]

All of us tend to make certain imaginative mistakes. Some of them, says Sidgwick, result from "the nature of the represented feeling."

> different kinds of past pleasures and pains do not equally admit of being revived in imagination. Thus, generally speaking, our more emotional and more representative pains are more easily revived than the more sensational and presentative: for example, it is at this moment more easy for me to imagine the discomfort of expectancy which preceded a past sea-sickness than the pain of the actual nausea: although I infer—from the recollection of judgments passed at the time—that the former pain was trifling compared with the latter. To this cause it seems due that past hardships, toils, and anxieties often appear pleasurable when we look back upon them, after some interval; for the excitement, the heightened sense of life that accompanied the painful struggle, would have been pleasurable if taken by itself; and it is this that we recall rather than the pain.[286]

And, I would add, one cannot revive *intense* pleasures and unpleasures in imagination, though one can speak of them with some authority.

Other imaginative errors, Sidgwick tells us, result from "the general state of the mind at the time of making the representation."

> *E.g.* it is a matter of common remark with respect to the gratifications of appetite that we cannot estimate them adequately in the state of satiety, and that we are apt to exaggerate them in the state of desire.

Also, says Sidgwick, we exaggerate the disvalue of pains we fear. And:

> Further, when feeling any kind of pain or uneasiness we seem liable to underrate pain of a very dissimilar kind: thus in danger we value repose, overlooking its *ennui*, while the tedium of security makes us imagine the mingled excitement of past danger as almost purely pleasurable. And again when we are absorbed in any particular pleasant activity, the pleasures attending dissimilar activities are apt to be contemned: they appear coarse or thin, as the case may be . . .

Indeed they do. Sidgwick's discussion of hedonic assessment is the best ever written.[287]

Easily Overlooked Hedonic Contributors

(11) Nonhuman experiences

All suffering is bad because it feels awful. So, Bentham is right about nonhumans: "the question is not, Can they *reason*? nor Can they *talk*? but, Can they *suffer*?"[288] As Singer says, "there can be no moral justification for regarding the pain (or pleasure) that animals feel as less important than the same amount of pain (or pleasure) felt by humans."[289]

(12) Dreams

Dreams, I assume, are experiences; they are not mere data-gatherings later accessed as though one recalls having had them.[290] Human adults dream about three hours per night, so roughly 17% of our conscious lives are dreamt. Normally, we can't remember even the gist of our dreams, so much hedonic information is lost—information crucial for making many comparative hedonic assessments.

(13) A sense of purpose

Humans, like all animals, need to engage in purposeful or meaningful behavior to prosper psychologically.[291] Almost any type of behavior may be purposeful. Breathing, for example, may have meaning for an asthmatic or a yogi; playing the keyboards has meaning for others. Hence, Mill errs in saying, "The main constituents of a satisfied life appear to be two . . . tranquility and excitement."[292] To be satisfied, one also needs a sense of purpose.

Having meaning in one's life is pleasurable; it contrasts with feeling empty and worthless. The lazy life of a rich playboy is not nirvanic if it includes an empty feeling of pointless indirection—a feeling which sometimes follows a great triumph before a new goal is set. "Someone is happy," says Rawls, "when his plans are going well, his more important aspirations being fulfilled, and he feels sure that his good fortune will endure."[293] Hedonists may agree.[294]

(14) Small hedonic differences

According to Stocker, "someone who chooses holidays on the basis of pleasure might want a pleasurable holiday, but not care whether it is more pleasurable than simply 'very pleasurable indeed.'"[295] Why mightn't she care? Several explanations are possible: a) if caring more won't net her a more pleasurable holiday, she might see no point in it; b) she might feel undeserving of a holiday more pleasurable than "very pleasurable indeed;" c) her attitudes might not, in general, be that fine-grained, perhaps for reasons of cognitive economy; she might care about only *large* important differences; d) she might not think clearly about how better pleasures feel in relation to milder pleasures. Avoid such mistakes when implementing IOI and HF: care more about longer, greater pleasures; assign them higher numbers.

(15) Differences too small to recognize introspectively

Philosophical reflection and psychological research have buried the "transparency thesis," according to which one can always come to believe, by reflection, the whole truth and nothing but the truth about one's current or very recent mental states. Since transparency fails, hedonic intensity might change imperceptibly, as Parfit suggests:[296] an attentive subject might fail to realize that her conscious state has changed, even though it has improved or worsened. Moreover, such cases probably occur, given the analogy between introspection and outer perception: Weber's Law specifies input thresholds for the outer modalities below which perceived differences aren't consciously registered.

Quinn's puzzle of the self-torturer, inspired by Parfit, can be solved by appealing to imperceptible changes. Says Quinn,

> Suppose there is a medical device that enables doctors to apply electric current to the body in increments so tiny that the patient cannot feel them. The device has 1001 settings: 0 (off) and 1 . . . 1000. Suppose someone (call him the self-torturer) agrees to have the device, in some conveniently portable form, attached to him in return for the following conditions: The device is initially set at 0. At the start of each week he is allowed a period of free experimentation in which he may try out and compare different settings, after which the dial is returned to its previous position. At any other time, he has only two options—to stay put or to advance the dial one setting. But he may advance only one step each week, and he may *never* retreat. *At each advance he gets $10,000.*[297]

The puzzle arises as follows: each week, the self-torturer gladly advances the dial in exchange for the ten grand. "No harm done, and I can quit my job!" But eventually, he feels so much pain that he would return his riches to end it. Where does he err, and why is it erring?

If pain can worsen imperceptibly, then the self-torturer shouldn't think, each time the dial advances, that no harm is done. The medical device, Quinn says, applies "electric current to the body in increments so tiny that the patient cannot feel them." But the self-torturer does feel them—at least some of them—even if he doesn't realize it; that explains why enough increments cause serious pain.

(There will even be behavioral evidence that particular changes cause harm: at 1,000, the self-torturer moans in pain, though at 0 he is quiet; the increases in volume will be measurable. This evidence is inconclusive, for one might moan louder even though the level of pain

remains constant.[298] Someday we might also have neurophysiological evidence of imperceptible hedonic changes.)

If some twists of the dial harm the self-torturer very slightly at each moment, then they harm him significantly over the course of his life. I don't know whether advancing to 1 is among the changes that imperceptibly do harm; but, since it might be, staying put at 0 would be wise (unless one is poor, in which case the $10,000 might outweigh the chance of permanent harm; further changes, however, would soon be inadvisable due to diminishing marginal utility).

Our feelings may be constantly improving or worsening imperceptibly. However, "don't overlook imperceptible hedonic changes" will be important practical advice only when such changes can be reliably detected. HF, as stated, doesn't even allow for such subtle distinctions: one can assign J a different number than K only if J and K are closest in intensity to different base experiences, but the base experiences, I've been assuming, are perceptibly different. HF could be modified to recognize imperceptible distinctions by enjoining us to assign new numbers to the base every half-second rather than every five seconds; then we could assign different numbers to new, imperceptibly different experiences. However, at this time, our crude comparative methods can't capitalize on such fine distinctions.

(16) Background feelings

Most qualia occur in the theatrical background, unreflected upon, upstaged by more striking or salient experiences. The uniform sounds of a mill or waterfall (or, more common to my experience, of an air conditioner

or dishwasher), as Leibniz says, may be perceived but unnoticed because they are "too unvarying."[299]

Mill contrasts "the occasional brilliant flash of enjoyment," with "its permanent and steady flame."[300] This "steady" flame may be unnoticed in the background despite its continuous value. Mill's steady flame, I suggest, is Sidgwick's "more indefinite kind of pleasure, which is an important element of ordinary human happiness,—the 'well-feeling' that accompanies and is a sign of physical well-being"[301] (and, I would add, of psychological well-being). In Plato's *Philebus*, Socrates condemns out of hand those who "think that at such times as they are not feeling pain they are feeling pleasure."[302] But one will feel pleasure at such times if one's background experience is valuable.

Background pleasures, like foreground pleasures, may differ phenomenologically: unreflected upon joy differs from unreflected upon contentment or pride, and of course instances of these vary in intensity. (Don't think of these feelings as moods: feelings are qualitative while moods—for example, being cheerful or irritable—consist at least partly in behavioral dispositions.) Qualitatively distinct background pleasures also arise from different sense-organs. Oliver Sacks, for instance, quotes a patient who lost his sense of smell:

> Sense of smell? I never gave it a thought. You don't normally give it a thought. But when I lost it—it was like being struck blind. Life lost a good deal of its savour—one doesn't realise how much 'savour' *is* smell. You *smell* people, you *smell* books, you *smell* the city, you *smell* the spring—maybe not consciously, but as a rich unconscious background to everything else. My whole world was suddenly radically poorer.[303]

This patient understood the value of background smells only after he stopped having them.

Nagel, I believe, mischaracterizes the phenomenon:

> There are elements which, if added to one's experience, make life better; there are other elements which, if added to one's experience, make life worse. But what remains when these are set aside is not merely *neutral*: it is emphatically positive. Therefore life is worth living even when the bad elements of experience are plentiful, and the good ones too meager to outweigh the bad ones on their own. The additional weight is supplied by experience itself, rather than by any of its contents.[304]

But a background hum of good feeling supplies the additional weight, not "experience itself." For if experience itself had "emphatically positive" value, then all experiences would be very good in one respect; but dark episodes of depression lack any hedonic consolation.

Background feelings may also be unpleasant: "a neutral state," says Aristotle, "is painful to many people because of their nature."[305] Such people are psychologically unwell.

Negative background feelings also vary phenomenologically; one's background state, for example, may be depressed, anxious or bitter. Such feelings come easily into the foreground: unpleasures, more than pleasures, tend to turn the mind inward.

Conclusion

I outlined two general methods for comparing hedonic value: Ideal Observer Imitation and Hedonic Figuring. These methods alone give little guidance, so I offered the following principles and advice to supplement them:

Value Principles:

(1) More intense pleasures are always better qua experience than less intense pleasures.

(2) Some pleasures and unpleasures differ lexically.

Practical Strategies:

(3) Simplify comparisons by "canceling out" experiences.

(4) Compare representative portions of alternatives.

Mistakes to Avoid:

(5) Don't be influenced by features extrinsic to experience.

(6) Don't use the peak and end rule.

(7) Don't confuse pleasure intensity with sensation intensity.

(8) Don't be misled by deep, fitness-enhancing desires.

(9) Be cautious about assessing types of experience you haven't had or don't appreciate; consult others.

(10) Correct for your imaginative weaknesses.

Easily Overlooked Hedonic Contributors:

(11) Nonhuman experiences

(12) Dreams

(13) A sense of purpose

(14) Small hedonic differences

(15) Differences too small to recognize introspectively

(16) Background feelings

Each of these principles helps implement both IOI and HF, except: HF presupposes (1); (2) helps correct a weakness inherent to HF; (3) and (4) are superfluous to HF; IOI presupposes (5); and HF, as formulated, cannot accommodate (12).

So, to answer the question, "Which of two alternatives is hedonically superior?" one should compare them twice, using IOI and HF (as well as HF can be implemented) along with the auxiliary principles. One can be guided by these principles, once internalized, without reviewing them. If IOI and HF yield the same answer, and if others using these methods concur, then that answer is well-supported. If IOI and HF yield sharply different results, favor the former (as I did regarding lexicality): human judgment, though fallible, provides solid evidence for hedonic assessment, while HF assumes, without evidence, that hedonic value reliably conforms to arithmetic methods.

All this messy advice, of course, falls pitifully short of a workable decision procedure for hedonic assessment; however, nobody who grasps the complexities of the topic would think that goal in sight.

Chapter 7: A Set of Solutions to Parfit's Problems

In "Future Generations," Part Four of *Reasons and Persons*, Parfit searches for a satisfactory theory of well-being.[306] Utilitarians believe that well-being exhausts value theory. However, principles governing other normative domains, such as rights and distributive justice, may supplement theories of well-being.

Parfit cannot find a theory that solves each of his ingenious problems. These problems involve comparing—and hence ranking—outcomes or states of affairs. Does Parfit unwittingly discredit his own approach to ethics? As Temkin tells us,

> some Aristotelians and Kantians might relish insuperable difficulties with ranking outcomes. Already convinced that the question, "what ought to be the case?" receives too much attention, they might welcome its relegation to the scrap heap of the unanswerable and irrelevant, enabling the "genuinely" important questions—that is, "how ought one *to be*?" or "what ought one *to do*?"—to receive more (their rightful?) consideration in the domain of practical reasoning.[307]

But I will show how a theory of hedonic well-being solves Parfit's problems. One may affirm the general form of these solutions even if such a theory is only part of well-being.

A Quasi-Maximizing Theory

First, let me briefly describe the relevant parts of the theory.

1. (Lexicality) If pleasure S is sufficiently more intense than pleasure I, then S is *lexically superior* to I, meaning that no finite duration of I is as good as some relatively short duration of S. Let's say, somewhat arbitrarily, that S is lexically superior to I just in case a year or more of S is better than any finite duration of I. For example, a year of ecstasy is better than any finite enjoyment of muzak and potatoes, so ecstasy is lexically superior to such pleasures.

2. (Duration) According to this principle, someone's feeling pleasure is better than someone's feeling slightly more intense pleasure for less than 1% as long.

3. (Intransitivity) According to *Transitivity*, if A is better than B, and B is better than C, then A is better than C. Lexicality and Duration entail that Transitivity is false. The proof goes like this. Duration entails that 1 year of ecstasy (=A) is worse than 100 years of pleasure slightly less intense (=B); B is worse than 10,000 years of pleasure slightly less intense (=C); C is worse than 1,000,000 years of pleasure slightly less intense (=D); and so on to Z, which is 10^{50} years of extremely mild pleasure. Given these premises, Transitivity entails that A is worse than Z. However, according to Lexicality, A is better than Z. I reject Transitivity.

4. (The Prudence Principle) If the value of alternatives could be accurately represented with numbers, then *being better than* would be transitive, for "being a higher number than" is transitive. But since Transitivity fails, this quantitative picture of value fails. Hence, I cannot say that hedonic value ought to be maximized, since "maximize" is quantitative. However, the theory I propose closely resembles maximizing theories, as I shall explain.

According to the Prudence Principle, alternative B is hedonically better than alternative C if having all the experiences in B would be better for one than having all the experiences in C.[308] And, as Rawls says, total utilitarianism—which enjoins maximizing value—uses the idea of an impartial sympathetic spectator, who

> gives free reign to his capacity for sympathetic identification by viewing each person's situation as it affects that person. Thus he imagines himself in the place of each person in turn, and when he has done this for everyone, the strength of his approval is determined by the balance of satisfactions to which he has sympathetically responded. When he has made the rounds of all the affected parties, so to speak, his approval expresses the total result.[309]

Were such an observer to find that his "balance of satisfactions" favors B over C—and so, he would prefer to have all the experiences in B rather than all those in C—then this would provide excellent evidence, given the Prudence Principle, that B is hedonically better than C. Hence, my theory, given the Prudence Principle, yields many of the judgments of total utilitarianism.

I call the view expressed by 1-4 *Quasi-Maximizing*: "Maximizing" because of the Prudence Principle; "Quasi" because of Intransitivity.

Parfit's Nonparadoxical Problems

Principles of well-being, Parfit tells us, may take either a person-affecting or an impersonal form. Person-affecting principles hold that "what is bad must be bad for someone"[310] because what matters is "whether our acts will be good or bad for those people whom they affect." These principles cannot solve Parfit's *Non-Identity Problem*. If a 14 year old conceives a child now, let's suppose, then her child would have a life worth living, but if she waits, her child would be much better off. Assume that nothing else is relevant to her decision. Person-affecting principles cannot explain why the young woman shouldn't conceive now, as doing so wouldn't harm any people who are ever actual; the child with a higher quality of life would never exist.[311] On the Quasi-Maximizing theory—more specifically, according to the Prudence Principle—waiting would be better since one would be better off having the *later* child's experiences.

An adequate theory of well-being, Parfit concludes, must take an impersonal form: "We must appeal to a principle which is about the quality and quantity of lives that are lived, but is not about what is good or bad for those people whom our acts affect."[312] However, Parfit rejects each impersonal theory he considers. According to the The Hedonistic Impersonal Total Principle,

> If other things are equal, the best outcome is the one in which there would be the greatest quantity of happiness—the greatest net sum of happiness minus misery.

This principle and its non-hedonic versions entail Parfit's second problem, the *Repugnant Conclusion*:

> For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living.[313]

A total principle of well-being thus entails the repugnant view that enough lives of piddling value are preferable, in terms of welfare, to ten billion of the very best lives. The Prudence Principle, conjoined with Lexicality, denies the Repugnant Conclusion: ten billion lives of very high quality should include at least a year of intense pleasure, so having those pleasures would be better for one than having any duration of very mild pleasure.

According to the Average Principle, ". . . it is worse if there is a lower average quality of life, per life lived."[314] On this principle, the Repugnant Conclusion is false, for the lives of very high quality are better, on average, than those barely worth living.[315] However, the Average Principle is untenable: it entails, for instance, that a population in agony would be improved in terms of welfare if someone were born whose agony was a tiny bit less.[316]

On the Average Principle, only "quality" or average well-being matters; on the Total Principle, only "quantity" or total well-being matters. Each view is too extreme. The best theory, Parfit thinks, will value both.[317]

But merely valuing both, he recognizes, won't avoid the Repugnant Conclusion; for even if its smaller, better-off population gets extra points for quality, so long as no limit is placed on quantitative value, enough lives barely worth living would be better.[318] So, Parfit considers valuing quality and quantity but limiting how much lives worth living can contribute to a world's quantitative value. Let that limit be whatever value ten billion very good lives have. On this view, the two populations in the Repugnant Conclusion would be of equal quantitative value, but the smaller population would be better overall because of its greater qualitative value.

Consider two variants of this view. On the first, lives worth living *that occur, say, within any hundred year period* cannot contribute more to the world's quantitative value than the limit allows. Now consider two alternatives:

K: 20 billion people, all of whom are very happy and live in the same century.
L: 20 billion people, all of whom are very happy, ten billion of whom live in one century, and ten billion in another.

On this view, C is twice as good as B. But this is absurd; timing as such shouldn't matter. (This objection is mine.)

Parfit further develops the first variant. "It would always be bad if an extra person has to endure extreme agony. And this would be just as bad, however many others have similar lives."[319] Thus, even if we limit how much lives worth living can contribute to a world's quantitative value, we shouldn't limit how much disvalue horrible lives can contribute. Now, given this asymmetry, consider these two future populations:

D includes: (i) Earthlings like Earth's present population; (ii) vastly many people living concurrently, all of whom have a very high quality of life, except that one person in each group of 10 billion has a painful disease that makes his or her life not worth living.
E: Just like D, except that each group exists in a *different* future century.

On the first variant, D would be very bad and E would be very good, "even though, in both outcomes, there would be the very same number of extra future people, with the same very high quality of life for all except the unfortunate one in each ten billion."[320] This is the *Absurd Conclusion.*

Let's extend the Prudence Principle as follows to cover alternatives that are equally good in terms of hedonic value: "two alternatives have equal hedonic value if my having the experiences of one is neither better nor worse for me than my having the experiences of the other." On this extension, the two alternatives in the Absurd Conclusion are equally good; and so, by appealing to it, the Quasi-Maximizing theory rejects the Absurd Conclusion.

On the second variant, *all* lives cannot contribute more to a world's quantitative value than the limit allows. So, once that limit is reached, it would be considered bad in terms of welfare for ten billion and one additional persons to exist, ten billion of whom have excellent lives and one of whom has a life barely worse than neutral. But such an addition would not be bad. Moreover, the view is absurd given that, before the limit is reached, adding those persons to the population would be considered a substantial improvement. Again, timing as such shouldn't matter.

Parfit considers two more ways to avoid the Repugnant Conclusion. Each denies that there is a single scale of value. On the Appeal to the

Valueless Level, the mass of lives barely worth living in the Repugnant Conclusion cannot be as good as ten billion blissful lives because lives below a certain quality have no value.[321] On the Lexical View, *Mediocre* lives always improve a state of affairs, but no number of them are as good as one *Blissful* life. Parfit criticizes these views similarly; I will focus on the Lexical View, which Lexicality entails. That view, Parfit says, entails an unacceptable variant of the Absurd Conclusion:

> (A) Suppose that, in some history of the future, there would always be an enormous number of people, and for each one person who suffers, and has a life that is not worth living, there would be ten billion people whose lives *are* worth living, though their quality of life is not quite as high as the Mediocre Level. This would be *worse* than if there were no future people.[322]

The conclusion of (A) is too strong; no theory of well-being entails that a state of affairs "would be worse than if there were no future people," for values outside the domain of well-being might be absent if persons are. For example, without future people, there would be no intellectual achievement, and perhaps this loss would outweigh the ill-fortunes of those who live. So, (A) should conclude, "This would be *worse in terms of well-being* than if there were no future people."

How does the Lexical View entail (A)? Parfit says,

> The existence of ten billion people below [the Mediocre Level] would have less value than that of a single person above the Blissful Level. If the existence of these people would have less value than that of only one such person, its value would be more than outweighed by the existence of one person who suffers, and has a life that is not worth living.[323]

So, if any number of Mediocre Lives are worse than one Blissful Life, then such lives should be outweighed by "one person who suffers, and has a life that is not worth living." For this argument to succeed, this bad life must be very bad; it must be the unfortunate analogue of Bliss. Such a life is worse than the phrase "who suffers and has a life not worth living" suggests; suffering might be compensated, at least partly, by the good things in life, while "a life not worth living" might connote a life barely worse than neutral. If Parfit's argument is sound—if the one life is *very* bad—then indeed the Lexical View entails that such a life could outweigh the ten billion Mediocre Lives in terms of well-being. However, I don't find this implication absurd; I accept it.

"On the Lexical View," says Parfit, "when we consider lives above the Mediocre Level, quantity could always outweigh quality."[324] On my view, this is because having a much longer duration of Mediocre plus experiences would be better for one than having a shorter duration of very intense pleasures. So, Parfit says, the Lexical View entails a variant of the Repugnant Conclusion:

> (R) If there were ten billion people living, all with a very high quality of life, there must be some much larger imaginable population whose existence would be *better*, even though its members have lives that are barely above the Mediocre Level.[325]

(R) strikes me as true because I have internalized the Quasi-Maximizing theory. I'll offer two arguments for (R). I don't aim to prove it; I merely want to show that a theory's entailing it doesn't count against the theory.

First, a principle analogous to (R) is plausible:

(R*) If there were ten billion people living, all with agonizing lives, there must be some much larger imaginable population whose existence would be *worse*, even though its members have lives that are barely below the Bad Mediocre Level.

By "the Bad Mediocre Level," I mean, "the level of well-being characterized by experiences such that the following is true: having these experiences for vastly long would be neither better nor worse than having the experiences of ten billion horrible lives." Lives barely below the Mediocre Level, I take it, are awful.

Insofar as (R*) is plausible, (R) should seem acceptable. Pleasures and unpleasures are normatively analogous in this case.

My second argument has the form: m and n are of equal value; k is better than n; so, k is better than m. The premises support the conclusion, even though, on my view, they don't deductively entail it.

(1) The following two worlds are of equal hedonic value: (m) a world containing ten billion Blissful people, all of whom live on Earth; (n) a world containing ten billion Blissful Earthlings *and* a great many additional people in distant galaxies whose lives are hedonically neutral.
(2) The n-world would be improved in terms of hedonic well-being if all its inhabitants became Mediocre plus people. (Those who quantify hedonic value might say, "Such a change would raise both average and total utility.")
(C) The Mediocre plus world is better than the m-world. In other words, (R) is true.

People who loathe (R) should offer sound principles that justify rejecting this argument.

Why is (R) considered repugnant? The Average Principle entails its denial; (R) seems less repugnant once one internalizes Parfit's refutation of that principle. And even if average utility deserves some weight, enough Mediocre plus lives should outweigh that influence.

Perhaps (R) seems false because each Blissful life is better than each Mediocre plus life. One might reason as follows: "Behind a veil of ignorance, I would rationally prefer a Blissful world to a Mediocre plus world, given that I will be have to be someone, and I would rather be Blissful than just above Mediocre." But this decision procedure, as Parfit points out, entails the absurdity that Hell One is worse than Hell Two. In Hell One, ten people suffer great agony for fifty years. In Hell Two, ten million people suffer great agony for fifty years minus a day.[326] So that procedure cannot be trusted.

Hence, (R) is acceptable, if not true.

The Second Paradox

Creative work in philosophy typically arises out of existing ideas; so what one person does would most likely have been done by someone else, given a few more years. But Parfit has devised a problem so clever and original that it may be an exception to this rule. He calls it "the Second Paradox."[327]

The Second Paradox is a set of possible worlds ordered so that they seem to get better and better, yet the last is worse than the first. The paradoxical conclusion, derived with Transitivity, is that the last possibility is better than the first. Parfit does not reject Transitivity; he resolves the paradox differently. I will argue that Parfit's and Temkin's resolutions of the paradox are inadequate, and that the Second Paradox supports Intransitivity. My ambitions here are great: I want to show that denying Transitivity is the *only* acceptable resolution of the paradox.

How the Second Paradox Goes

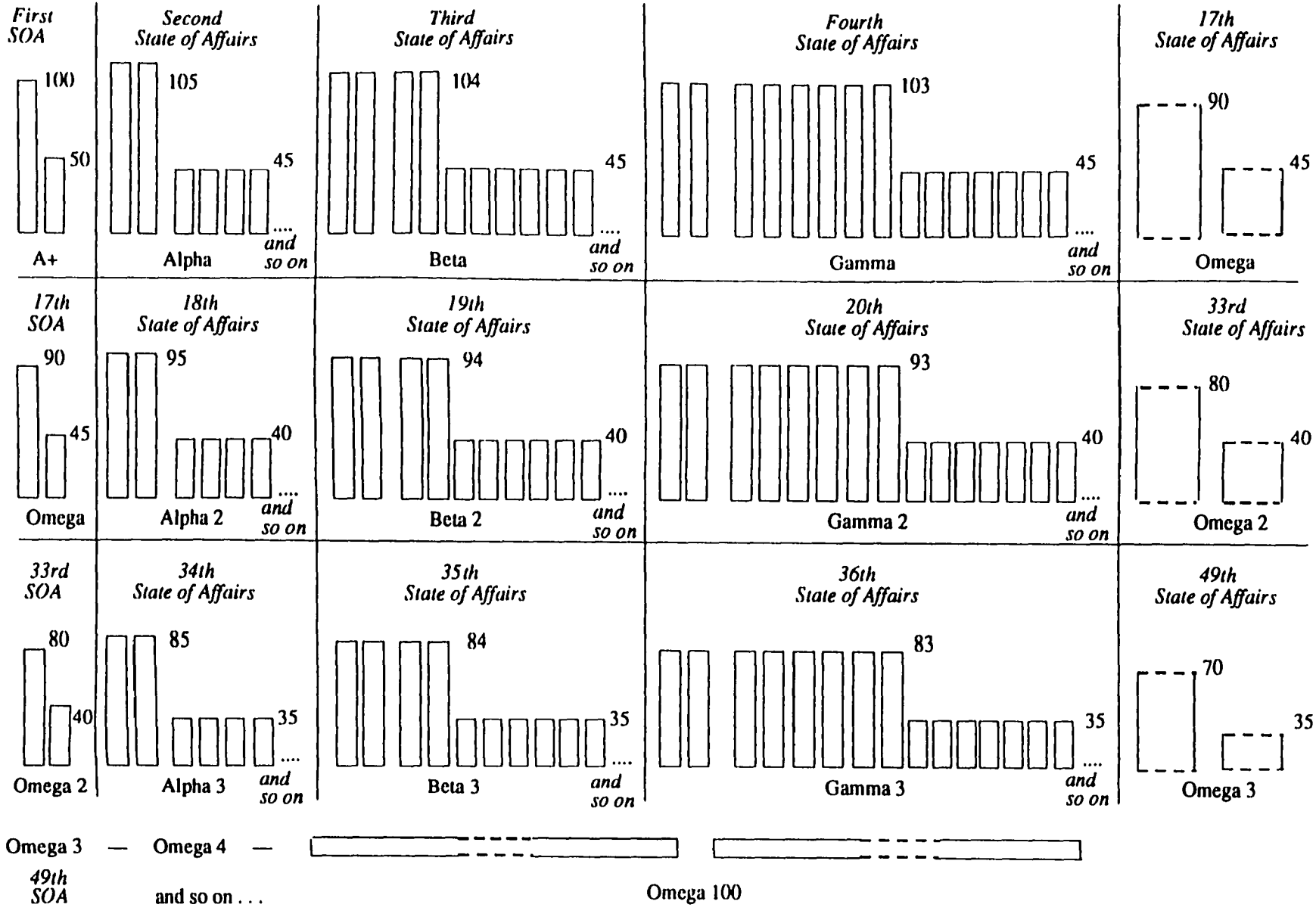
The Second Paradox is tedious to work through. However, once grasped, it dazzles the philosophical imagination.

The first state of affairs is A+. (See the diagram on the next page.) A+ contains two groups of 10 billion people: one group whose lives are at '100,' an ecstatic level, and another whose lives are at '50,' a level of pleasure well worth enjoying. The last state of affairs is Omega 100, a world that contains many, many lives each of which is barely worth living at each moment. In Omega 100 muzak and potatoes are the only pleasures in life. Although A+ is better than Omega 100, A+ is transformed into Omega 100 via only changes for the better.

Each change from A+ to Omega 100 takes one of two forms.

The first kind of change occurs as A+ becomes Alpha. This happens by raising both groups in A+ to a 105 level of pleasure and adding many, many groups of 10 billion people whose lives, at 45, are well worth living.

A Visual Guide to Parfit's Second Paradox



The width of the blocks indicates the number of people living; the height shows the intensity of their pleasures. Dotted lines indicate that the block is much wider than shown, and wide blocks have been condensed in succeeding rows to make the diagram size manageable.

The world is much improved going from A+ to Alpha because all the people in A+ benefit from the change, especially those in the 50 group, and the only 'cost' of this benefit is adding people to the world who are glad to be alive.

The second kind of change occurs as Alpha is transformed into Beta. This occurs by lowering the two better-off groups in Alpha from 105 to 104 but raising as many worse-off groups from 45 to 104. (Even after this change many groups are at 45.) This kind of change occurs down the Greek alphabet until we reach Omega. In Omega, many groups are at 90 but many more still at 45.

Omega is transformed into Alpha 2 by improving all the lives in Omega to 95 (including the lives that were at 45) and adding many more groups at 40. This repeats the first sort of change. Alpha 2 is transformed into Beta 2 by lowering the better-off groups to 94 but raising the same number of worse-off groups to 94. This repeats the second kind of change. By the time we reach Omega 2, the better-off groups are down to 80, though there are many more of them, while many groups are still at 40. At Alpha 3 all the people in Omega 2 are promoted to the level of 85 and many groups at 35 are added.

So at each Omega the average quality of life is lower than it was at the previous Omega, and the population has been greatly increased. At Omega 100, everyone's life is barely worth living at each moment. We want to say both that Omega 100 is worse than A+ and that each change from A+ to Omega 100 is for the better. Each change seems for the better because the quality of life is lowered only for those who are better off, and then only when this loss is more than offset by gains for the worse off.

The Second Paradox may also be formulated in terms of painful lives. To do this, change the numbers in the Second Paradox to negative numbers; then the worlds get worse until they are better (the last state of affairs better than the first), contradicting Transitivity.

Parfit's Suggestion

Parfit resolves the paradox by claiming that Alpha is better than Beta: 20 billion people at 105 plus many more at 45 is better than 40 billion people at 104 plus many more at 45 (20 billion fewer at 45 than in Alpha).

Parfit defends this by appealing to "Perfectionism." Perfectionism is the view that "even if some change brings a great net benefit to those who are affected, it is a change for the worse if it involves the loss of one of the best things in life."[328] So, Alpha is better than Beta because in Alpha the best things are better. In Alpha the luckiest 20 billion listen to Mozart; in Beta, 40 billion listen to Haydn.[329]

However, the "best things in life" need not be lost as one's well-being declines; one could simply appreciate them less. Hence, Perfectionism is relevant as a defense of Parfit's resolution only if we interpret "the best things in life" as the most intense pleasures. Yet on this interpretation, Perfectionism is wildly implausible. Perfectionism entails that a duration of the best pleasure is better than a *much* longer duration of pleasure *very* slightly less intense. As Parfit admits, "[Perfectionism] conflicts with the preferences most of us would have about our own futures." Furthermore, it

entails that it could be bad for the middle class to become rich at a tiny cost to the wealthy. Therefore, Perfectionism is false.

Temkin's suggestion

Temkin suggests that many, if not all, of the steps in the Second Paradox may be worse in terms of inequality. Consider the comparison between A+ and Alpha. The fact that Alpha includes many more worse off people than A+ may represent an egalitarian respect in which Alpha is worse. Perhaps this respect outweighs other factors and yields the result that A+ is all things considered better than Alpha. Or consider the comparison between Alpha and Beta. The inequality between the better off groups and the worse off groups may be worse in Beta because Beta has even more better off people for the worse off people to resent (or that we can resent on their behalf). And again, perhaps this suffices to license the judgment that Beta is all things considered worse than Alpha.[330]

Temkin's proposals depend on the thesis that certain states of affairs are intrinsically bad in terms of equality. Temkin never argues this, but he offers the "fundamental intuition underlying egalitarianism—that it is bad, unfair or unjust, for some to be worse off than others through no fault of their own . . ."[331] Do such intuitions apply to the Second Paradox? Is it unfair, for example, that in Beta so many people prosper at the 104 level, while others only prosper at the 45 level? Nothing deters us from supposing that, in our version of the Second Paradox, the different groups of 10 billion people live in different galaxies and cannot affect each other. In such a

scenario, I think, fairness and justice do not come into play. Saying that a universe is made worse by the mere fact that isolated populations within it have varying degrees of welfare makes an ethical judgment too much on the model of certain aesthetic judgments. It would lessen the value of a Picasso (= prospering population) to add a section onto it painted by a lesser artist (= less prosperous population). But if a remote population prospered more than we on Earth, the resulting inequality would no more lessen the universe's value than the prospering of a possible but nonactual population. Egalitarian considerations should not affect how we evaluate the Second Paradox. At the very least, they do not outweigh benefits enjoyed by billions of people. There is a treatment of the paradox more palatable than this.

A Different Proposal

By denying Transitivity, we may hold that the alternatives in the Second Paradox get better and better *and* Omega 100 is worse than A+. The pleasures in A+ are lexically superior to the pleasures in Omega 100, and so A+ is better than Omega 100; but Parfit finds a path from A+ to Omega 100 involving only changes for the better.

The Mere Addition Paradox

Mere Addition is "when, in one of two outcomes, there exist extra people (1) who have lives worth living, (2) who affect no one else, and (3) whose existence does not involve social injustice."[332] The *Mere Addition Paradox* arises for these three states of affairs:

A: 5 billion people, all of whom have a very high quality of life.

A+: The 5 billion A-people and, by Mere Addition, 5 billion people whose lives are worth living though considerably worse than the lives of the A-people.

B: 10 billion people whose lives are about four-fifths as good as the lives of the people in A. The average quality of life is higher in B than in A+.[333]

According to Parfit, A is better than B, B is better than A+, but A is not better than A+. The paradox, on Parfit's view, arises because these beliefs are inconsistent with Transitivity. I reject Transitivity in favor of Lexicality and Duration. However, those latter principles don't apply to Mere Addition Paradox, and inductive evidence suggests that counterexamples to Transitivity are rare. So, I would put the problem as follows: "These three beliefs entail a violation of Transitivity, so they need support and explanation."

Temkin defends principles suggesting why (A, B, and A+) might violate Transitivity,[334] but this issue doesn't arise on a Quasi-Maximizing view, which rejects Parfit's thesis that A is better than B. On the Prudence

Principle, B is better than A: having the experiences of the 10 billion people would be better for one than having the experiences of the 5 billion people whose lives are better by 25%. This thesis is like (R), according to which some number of Mediocre plus lives are better than ten billion Blissful lives, but perhaps more plausible, since the B-lives are well above the Mediocre Level. I defended (R) with two arguments. I could defend the idea that B is better than A with two similar arguments.

Parfit argues, against this view, that "B is better than A" leads to the Repugnant Conclusion:

> There is a possible outcome C whose relation to B is just like B's relation to A. In C there are twice as many people, who are all worse off than everyone in B. . . . If we conclude that B is better than A, we must conclude that C is better than B. On the same argument, D would be better than C, E better than D, and so on down the Alphabet. The *best* outcome would be Z: an enormous population all of whom have lives that are barely worth living.[335]

I believe that C is better than B, D is better than C, and so on. However, I deny that Z is better than A; I deny that Z is best.[336] In fact, no state of affairs among A-Z is best; no state of affairs is better than each of the others. A-Z violates Transitivity for the same reason that the Second Paradox does: the A-pleasures are lexically superior to the Z-pleasures, but the path from A to Z involves only changes for the better. The Omegas in the Second Paradox closely resemble A-Z.

Conclusion

After a long and inventive, but ultimately unsuccessful, search for a satisfactory theory of well-being, Parfit despairs that

> With more unsolved problems, we are further away from the Unified Theory. We are further away from the theory that resolves our disagreements, and that, because it achieves these aims, might deserve to be called the truth.[337]

But having more unsolved problems needn't take us further away from the truth; solving one problem can point the way to solving the others and thus making theoretical progress. I believe that is the case here. The Second Paradox strongly suggests that betterness is not transitive; this allows us to affirm that B is better than A in the Mere Addition Paradox without entailing the Repugnant Conclusion, and if B *is* better than A—if 10 billion lives are better than 5 billion lives of somewhat higher quality—then duration can swamp intensity when the difference in intensity is small; while the Repugnant Conclusion suggests that intensity swamps duration when the difference in intensity is large. All this points to something like the theory of hedonic value I advocate, which affirms Lexicality, Duration, Intransitivity and the Prudence Principle.

In his Concluding Chapter, Parfit says,

> As I argued, we need a new theory about [well-being]. This must solve the Non-Identity Problem, avoid the Repugnant and Absurd Conclusions, and solve the Mere Addition Paradox. I failed to find a theory that can meet these four requirements. Though I

> failed to find such a theory, I believe that, if they tried, others could succeed.[338]

My Quasi-Maximizing theory, I've argued, provides defensible solutions to Parfit's problems. Many philosophers, however, won't deny Transitivity. I offer those philosophers an alternative set of solutions, which I find unpalatable. Transitivity is incompatible with Lexicality and Duration. Duration can hardly be denied, but perhaps Lexicality is false, and so a long enough duration of mild pleasure is preferable to a year, or even a trillion years, of ecstasy. If so, then the Impersonal Total Principle can solve Parfit's problems. On this principle, the Non-Identity Problem doesn't arise; the Repugnant Conclusion is true; the Absurd Conclusion doesn't follow; the Second Paradox is resolved by affirming that Omega 100 is better than A+; and the Mere Addition Paradox is resolved by affirming that A is worse than B and then—in response to Parfit's objection—by again embracing a repugnant conclusion, that A is worse than Z.

One Grand Conclusion: All Pleasures are Intrinsically Good; All Unpleasures are Intrinsically Bad

> in so far as they are pleasant, are [pleasures] not also good, leaving aside any consequences that they may entail? And in the same way pains, in so far as they are painful, are bad?[339] (Socrates)

> pain . . . is a bad thing *in itself*. It does not matter who experiences it, or where it comes in a life, or where in the course of a painful episode. Pain is bad; it should not happen. There should be as little pain as possible in the world, however it is distributed across people and across time.[340] (Broome)

My essay supports these views, adapted to pleasures and *unpleasures* (which include most, if not all, pains). I have argued for the grand conclusion that all pleasures are intrinsically good and all unpleasures are intrinsically bad (using "pleasure" and "unpleasure" to refer to experiences) by arguing for the following, more modest theses:

1. All pleasures are good in some way and all unpleasures are bad in some way due to how they feel. (ch. 1)
2. Experiences themselves are pleasant or unpleasant. (ch. 2)
3. Pleasures and unpleasures have agent-neutral value. (ch. 3)
4. There is reason to create people who will feel pleasure. (ch. 4)
5. Pleasures are good irrespective of what they're taken in; of what accompanies them; of what they depend on; and of whether they're deserved. (in ch. 6)

6. A theory which includes the thesis that pleasures are intrinsically good can solve Parfit's problems for moral theory. (ch. 7)

If any of these are false, then so is "the grand conclusion." For this reason, there is no short road to that finale. Perhaps, on top of my arguments, the elegance of the thesis—*all pleasures are intrinsically good and all unpleasures are intrinsically bad*—counts in its favor. But mere elegance, I think, doesn't count for much.

Endnotes

---

[1] See John Rawls, *A Theory of Justice*, p. 20. Reflective equilibrium seems more complex as Rawls characterizes it on pp. 48-50.

[2] Monroe Beardsley, "Intrinsic Value," p. 7.

[3] "If a person conceives her practical situation in terms provided for her by a specific ethical outlook, that will present her with certain apparent reasons for acting. On a better understanding of Aristotle's picture, the only standpoint at which she can address the question whether those reasons are genuine is one that she occupies precisely because she has a specific ethical outlook. That is a standpoint from which those seeming requirements are viewed as such, not a foundational standpoint at which she might try to reconstruct the demandingness of those requirements from scratch, out of materials from an independent description of nature." (John McDowell, *Mind and World* , p. 80)

[4] Peter Singer, *Animal Liberation*, 1990, p. 211 (p. 231 in the 1975 edition).

[5] Samuel Clark's approach is more bullying: "These things are so notoriously plain and self-evident, that nothing but the extremist stupidity of mind, corruption of manners, or perverseness of spirit, can possibly make any man entertain the least doubt concerning them." (*A Discourse Concerning the Unchangeable Obligations of Natural Religion*, in *British Moralists I*, p. 194)

[6] Christine M. Korsgaard, *The sources of normativity* , p. 18. Compare Peter Railton: "Our evaluative notions bear the stamp of their origins in religious, teleological conceptions of the world. This might lead one to skepticism about such discourse." ("Naturalism and Prescriptivity," p. 158)

[7] Henry Sidgwick, *The Methods of Ethics*, p. 396. Compare Christine M. Korsgaard: "Pufendorf and Hobbes ask how nature, an indifferent and mechanical world of matter in motion, can come to be imbued with moral properties." (*The sources of normativity* , p. 22)

[8] See the analogous problems for mathematical Platonism presented in Philip Kitcher, *The Nature of Mathematical Knowledge*, ch. 6.

[9] Derek Parfit, *Reasons and Persons* , p. 452. The next quotation is from the same page.

[10] Christine M. Korsgaard, *The sources of normativity*, p. 46.

[11] David Lewis, "Extrinsic Properties," p. 197. Moore, Chisholm, Korsgaard and Feldman, among others, conceive intrinsic value like this. See G. E. Moore, "The Conception of Intrinsic Value"; Roderick M. Chisholm, "Objectives and Intrinsic Value," p. 262; Roderick M. Chisholm, "Defining Intrinsic Value," p. 99; Christine M. Korsgaard, "Two distinctions in goodness," *Creating the Kingdom of Ends* , p. 251; Fred Feldman, "On the Intrinsic Value of Pleasures," p. 457.

[12] For other (and in my opinion, less motivated) uses of "intrinsic value," see John O'Neill, "The Varieties of Intrinsic Value."

[13] Harman's *basic* intrinsic values, however, have no bad inner aspects. See Gilbert Harman, "Toward a Theory of Intrinsic Value," p. 799. For further discussion, see Fred Feldman, *Doing the Best We Can*, p. 30.

[14] G. E. Moore, *Principia Ethica*, p. 36.

[15] According to Christine M. Korsgaard, Moore's contributive values "do the same job" as Kant's conditional values. ("Two distinctions in goodness," *Creating the Kingdom of Ends* , p. 270)

[16] See *ibid.*, p. 252 and elsewhere.

[17] Ronald Dworkin, *Life's Dominion*, p. 74.

[18] David Hume, *An Enquiry Concerning the Principals of Morals*, Appendix 1.

[19] Roderick M. Chisholm, *Brentano and Intrinsic Value* , p. 95.

[20] David Hume, *op. cit.*

[21] Monroe Beardsley, "Intrinsic Value," pp. 12-13, p. 13.

[22] Christine Korsgaard, *The sources of normativity*, p. 109.

[23] But not, I think, Noah Lemos or Pepita Haezrahi. See Lemos, *Intrinsic Value*, p. 75 (as well as Part II) and Haezrahi, "Pain and Pleasure: Some Reflections on Susan Stebbing's View That Pain and Pleasure Are Moral Values," p. 71.

[24] See John Stuart Mill, *Utilitarianism*, ch. 4, para. 3 and G. E. Moore, *Principia Ethica*, pp. 91, 223. These are quoted below.

[25] J. J. C. Smart, "An outline of a system of utilitarian ethics," p. 31.

[26] See Derek Parfit, *Reasons and Persons* , pt. 1.

[27] See Shelly Kagan, *The Limits of Morality* , pp. 13-14.

[28] Christine Korsgaard, "Two distinctions in goodness," *Creating the Kingdom of Ends*, p. 271. The next quotation is from the same page. Also, according to Korsgaard, Plato argues that the just soul is intrinsically good because its internal structure makes its possessor both happy and master of himself. (*The sources of normativity*, pp. 109-110)

[29] Ronald Dworkin, *Life's Dominion*, p. 74. The next quotation is from the same page.

[30] G. E. Moore, *Principia Ethica*, p. 28 (where the whole sentence appears in italics). The next quotation is from the same page.

[31] Michael Stocker, *Plural and Conflicting Values*, p. 323.

[32] James Griffin, *Well-Being*, p. 355, fn. 33.

[33] See John Rawls, "Outline of a Decision Procedure for Ethics," pp. 177-197, especially pp. 178-80, 187-189, and David Lewis "Dispositional Theories of Value—II."

[34] Aristotle, *The Nicomachean Ethics*, X. 2. (The next quotation is from VII. 13.) Epicurus and Diogenes Laertius made basically the same argument. See Whitney J. Oates, "The Life of Epicurus," p. 63.

[35] David Hume, Treatise of Human Nature, III. 1. ii, p. 471.

[36] John Stuart Mill, *Utilitarianism*, ch. 4, para. 3.

[37] Jan Narveson, *Morality and Utility* , p. 284.

[38] C. I. Lewis, *An Analysis of Knowledge and Valuation* , pp. 374-375.

[39] Thomas Nagel, *The View From Nowhere* , p. 146.

[40] G. E. Moore, *Principia Ethica*, p. 91. Also see p. 223.

[41] The quoted phrase is from James Griffin, *Well-Being*, p. 3.

[42] See Alan Gewirth, *Reason and Morality*; R. M. Hare, *Moral Thinking*; David Gauthier, *Morals By Agreement*. According to Richard Brandt, Firth takes "is morally wrong" to mean "would be disapproved of by any person who was factually omniscient, impartial and devoid of emotions toward particular persons but otherwise normal," and Westermarck takes "It is wrong to do A" to mean, "I have an impartial disposition to disapprove of acts like A." (*Facts, values and morality* , p. 2.) Brandt cites Edward Westermarck's *The Origin and Development of the Moral Ideas* , pp. 1, 17-18 and *Ethical Relativity*, pp. 14-15 and 141-142.

[43] Peter Railton, "Naturalism and Prescriptivity," p. 171.

[44] See Ronald de Sousa, "Arguments from Nature."

[45] See Roderick M. Chisholm, *Brentano and Intrinsic Value* and Noah Lemos, *Intrinsic Value*.

[46] J. J. C. Smart, "An outline of a system of utilitarian ethics," pp. 6, 7, 31.

[47] Peter Railton discusses this sort of reasoning in "Naturalism and Prescriptivity," pp. 167, 169.

[48] Derek Parfit makes this point about evolutionary accounts of attitudes. (*Reasons and Persons*, p. 308)

[49] George Sher, *Desert* , p. 133.

[50] See Henry Sidgwick, *The Methods of Ethics*, p. 213.

[51] Thomas Nagel, *The Last Word*, p. 105.

[52] That is the question.

[53] Monroe Beardsley, "Intrinsic Value," p. 6. Beardsley calls this "The Dialectical Demonstration."

[54] Monroe Beardsley, "Intrinsic Value," p. 4.

[55] John Rawls, *A Theory of Justice* , p. 51.

[56] "[Realists] say that they know moral truths by 'intuition,' but I cannot find that they mean anything by this except that they do have moral opinions." (Jonathan Bennett, *The Act Itself*, p. 12)

[57] This is a standard criticism of ideal observer theories. Compare Alan Gewirth, *Reason and Morality*, pp. 20-21.

[58] Christine M. Korsgaard, *The sources of normativity*, pp. 92-93.

[59] *Ibid.*, p. 93.

[60] *Ibid.*, p. 97.

[61] Thomas Nagel, "Universality and the reflective self," p. 202.

[62] Christine M. Korsgaard, *The sources of normativity*, pp. 97-98.

[63] John Stuart Mill, *Utilitarianism*, ch. 1, para. 4.

[64] Christine M. Korsgaard, *The sources of normativity* , p. 100. The next two quotations are from p. 101 and p. 102.

[65] James Griffin, "Modern Utilitarianism," p. 359. This passage is repeated in his *Well-Being*, p. 352, fn. 26.

[66] Christine M. Korsgaard, *The sources of normativity*, p. 103.

[67] *Ibid.*, pp. 103-104.

[68] *Ibid.*, pp. 120-121. (The next quotation is from p. 121.) Korsgaard offers this sort of argument in at least two other places: *Creating the Kingdom of Ends* , pp. ix-x, and "Two distinctions in goodness," *Creating the Kingdom of Ends*, pp. 272-273.

[69] Christine M. Korsgaard, *The sources of normativity*, p. 121.

[70] *Ibid.*, p. 122.

[71] Jonathan Bennett suggested this interpretation to me.

[72] *Ibid.*, p. 124.

[73] *Ibid.*, p. 41.

[74] Christine Korsgaard, "Aristotle and Kant on the Source of Value," *Creating the Kingdom of Ends*, pp. 240-241. She expresses similar thoughts in "Two distinctions in goodness," *Creating the Kingdom of Ends*, on p. 260 and especially pp. 261-262.

[75] *Ibid.*, p. 259.

[76] Christine M. Korsgaard, *The sources of normativity* , p. 92.

[77] Christine M. Korsgaard, "Two distinctions in goodness," *Creating the Kingdom of Ends*, p. 259.

[78] George Sher, *Desert*, p. 58. The next quotation comes from the same page.

[79] Christine Korsgaard, "Two distinctions in goodness," *Creating the Kingdom of Ends*, pp. 267-268.

[80] *Ibid.*, p. 261.

[81] *Ibid.*, p. 259.

[82] James Rachels, *The End of Life* , p. 46.

[83] Thomas Nagel, *The View From Nowhere*, pp. 145-146. And on pp. 157-158 he says: "Without some positive reason to think there is nothing in itself good or bad about having an experience you intensely like or dislike, we can't seriously regard the common impression to the contrary as a collective illusion. Such things are at least good or bad *for us*, if anything is."

[84] See Christine M. Korsgaard, *The sources of normativity* , pp. 33-42, 48.

[85] Nagel says much the same thing: "No objective view we can attain could possibly overrule our subjective authority in such cases." (*The View From Nowhere*, p. 158)

[86] *Ibid.*, p. 144. Nagel's point also helps to rebut David Gauthier's thesis that if objective value exists, then the best explanation of our behavior must refer to it. (*Morals By Agreement*, p. 56)

[87] J.L. Mackie, *Ethics* (Penguin Books, 1977), p. 38.

[88] Paul M. Churchland, *A Neurocomputational Perspective* , p. 303.

[89] William P. Alston, *Epistemic Justification*, p. 58.

[90] William P. Alston, "Epistemic Desiderata," p. 527. The next quotation comes from p. 539.

[91] Norton Nelkin, "Reconsidering Pain," p. 332.

[92] Christine M. Korsgaard, *The sources of normativity* , p. 146. Korsgaard should not say "*very*" strong. When the pain is mild, so is the impulse.

[93] Derek Parfit, *Reasons and Persons*, p. 501.

[94] Richard J. Hall, "Are Pains Necessarily Unpleasant?" p. 647.

[95] Michael Tye, *Ten Problems of Consciousness* , p. 113. Tye doesn't characterize disorder. Do representations of bodily disorder have to be painful? Mightn't an experience during an LSD trip be pleasurable but represent bodily disorder?

[96] George Pitcher, "Pain Perception," p. 371. Similarly, see D. M. Armstrong, *A Materialist Theory of Mind* and K. Wilkes, *Physicalism*.

---

[97] R. Melzack and P. D. Wall, *The Challenge of Pain*, p. 206, quoted in Nikola Grahek's excellent essay, "Objective and subjective aspects of pain," p. 252. Grahek gives other examples of pain without bodily damage on p. 252 and p. 260.

[98] Norton Nelkin, "Reconsidering Pain," p. 332.

[99] *Ibid.*, p. 337. The exclamation mark suggests that emotion is essential to pain, but that is not Nelkin's considered view.

[100] See Nelkin, "Reconsidering Pain," pp. 327, 329, 334-335, 338.

[101] Christine M. Korsgaard, *The sources of normativity*, p. 147.

[102] Richard Brandt, *A Theory of the Good and the Right*, p. 38. Gilbert Ryle advances a behaviorist forerunner of this view in *Dilemmas*; for criticism of Ryle, see T. L. S. Sprigge, *The Rational Foundations of Ethics* , pp. 131-132.

[103] Thomas Nagel, *The View From Nowhere*, pp. 156-157.

[104] The analogous objection to Dislike—"People dislike their unpleasure because it's unpleasant . . ."—admits of similar treatment.

[105] Henry Sidgwick, *The Methods of Ethics*, p. 125.

[106] *Ibid.*, p. 127. Sidgwick also says that "exciting pleasures are liable to exercise, even when actually felt, a volitional stimulus out of proportion to their intensities as pleasures." But I can't think of any convincing examples of exciting pleasures.

[107] This might be a counterexample to R. M. Hare's claim that, "If I am suffering, I have a motive for ending the suffering." (*Moral Thinking*, p. 93)

[108] Richard J. Hall, "Are Pains Necessarily Unpleasant?" p. 646.

[109] See C. D. Broad, *Five Types of Ethical Theory* , pp. 237-238.

[110] Derek Parfit, *Reasons and Persons*, p. 493.

[111] *Ibid.*, p. 501.

[112] Kurt Baier and Roger Trigg make this point. See Baier, *The Moral Point of View*, p. 273 and Trigg, *Pain and Emotion* , p. 114.

[113] Roger Trigg, *Pain and Emotion*, p. 151.

[114] Has this experiment been performed, or is it an urban legend? Robert Van Gulick calls it "an old fraternity gag." Paul M. Churchland uses this type of example without citation. (*Matter and Consciousness*, p. 77)

[115] Paul M. Churchland, *Matter and Consciousness*, p. 52.

[116] T. L. S. Sprigge ably defends this view in *The Rational Foundations of Ethics*, ch. 5.

[117] Richard J. Hall, "Are Pains Necessarily Unpleasant?" pp. 643, 648, 653, 655.

[118] For criticism, see Guzeldere on HOP and Byrne on HOT. (Guven Guzeldere, "Is Consciousness the Perception of What Passes in One's Own Mind?" and Alex Byrne, "Some Like It HOT: Consciousness and Higher-Order Thoughts.")

[119] C. D. Broad, *Five Types of Ethical Theory* , p. 229.

[120] G. E. Moore, *Principia Ethica*, pp. 12-13, 78; J. S. Feibleman, "A Philosophical Analysis of Pleasure," p. 252; Carolyn Morillo, *Contingent Creatures*, especially p. 97; and perhaps David Chalmers, *The Conscious Mind*, p. 17.

[121] Henry Sidgwick, *The Methods of Ethics*, p. 127; Georg Henrik von Wright, *The Varieties of Goodness*, pp. 67-69; William P. Alston, "Pleasure," p. 344; Kai Nielsen, "Hedonism and the Ends of Life," *Philosophical Journal*, p. 24; Jonathan Glover, *Causing Death and Saving Lives*, p. 63; Rem B. Edwards, *Pleasures and Pains: A Theory of Qualitative Hedonism*, pp. 83-86; James Griffin, "Modern Utilitarianism," p. 333; Derek Parfit, *Reasons and Persons*, p. 493; Paul M. Churchland, *Matter and Consciousness*, p. 52; T. L. S. Sprigge, *The Rational Foundations of Ethics*, p. 130; Richard J. Hall, "Are Pains Necessarily Unpleasant?" p. 646; Noah Lemos, *Intrinsic Value*, p. 67; Norton Nelkin, "Reconsidering Pain," p. 329; Christine M. Korsgaard, *The sources of normativity*, p. 148; and Fred Feldman, "On the Intrinsic Value of Pleasures," p. 449.

[122] David Hume, *A Treatise of Human Nature*, p. 629 (appendix).

[123] Christine M. Korsgaard, *The sources of normativity*, p. 148. Brandt makes a similar criticism. (*A Theory of the Good and the Right*, p. 37)

[124] Norton Nelkin, "Reconsidering Pain," p. 331.

[125] Henry Sidgwick, *The Methods of Ethics*, p. 127. For a similar view, see Kai Nielsen, "Hedonism and the Ends of Life," p. 24.

[126] Derek Parfit, *Reasons and Persons*, p. 501.

[127] Richard Brandt, *A Theory of the Good and the Right*, pp. 37-38; Richard J. Hall, "Are Pains Necessarily Unpleasant?" section III.

[128] Richard J. Hall, "Are Pains Necessarily Unpleasant?" p. 651.

[129] *Ibid.*, p. 652.

[130] Email communication, October 29, 1996. Also, when I gave this paper at Western Washington University, Philip Montague enthusiastically agreed.

[131] R. M. Hare, "Pain and Evil," p. 88.

[132] Roger Trigg, *Pain and Emotion*, p. 81.

[133] Christine M. Korsgaard, *The sources of normativity*, p. 147.

[134] T. L. S. Sprigge, *The Rational Foundations of Ethics*, p. 141, p. 142.

[135] For a similar view of pleasure, see Carolyn R. Morillo, *Contingent Creatures*, especially pp. 63-65.

[136] T. L. S. Sprigge, *op. cit.*, p. 148.

[137] Carolyn R. Morillo, *Contingent Creatures*, p. 42.

[138] J. N. Findlay, *Values and Intentions*, p. 177.

[139] Carolyn R. Morillo, *op. cit.*, p. 165.

[140] Jonathan Bennett taught me this view. He says that he learned it from Hume.

[141] (2) follows Peter van Inwagen's response to the argument from design. (*Metaphysics*, ch. 8)

[142] David O. Brink endorses one such view (without developing it) in "Rational Egoism and the Separateness of Persons," p. 112.

[143] Irwin Goldstein, "Why People Prefer Pleasure to Pain," p. 360.

[144] David Gauthier, *Morals By Agreement*, p. 52. George R. Carlson seems to agree in "Pain and the Quantum Leap to Agent-Neutral Value."

[145] Henry Sidgwick, *The Methods of Ethics*, p. 386, fn. 4. (For the idea that this conflict can't be resolved by argument, see p. 498 and elsewhere in the Concluding Chapter.) Compare Gauthier: "The reconciliation of morality with rationality is the central problem of modern moral philosophy." (*Moral Dealing: Contract, Ethics and Reason*, p. 150)

[146] For Singer, see p. 234 of *How Are We To Live?* Parfit leaves open whether I have a reason to act morally if I have no such desire. This suggests that he has no argument for resolving the prudence/morality clash. (*Reasons and Persons*, pp. 121-122)

[147] Jonathan Bennett, "On Maximizing Happiness," p. 69. Nagel: ". . . the personal standpoint must be taken into account directly in the justification of any ethical or political system which humans can be expected to live by. This is an ethical and not merely a practical claim." (*Equality and Partiality*, p. 15)

[148] Philip Pettit and Michael Smith, "Parfit's P," pp. 83-84.

[149] Bertrand Russell: "'Reason' has a perfectly clear and precise meaning. It signifies the choice of the right means to an end that you wish to achieve. It has nothing whatever to do with the choice of ends." (*Human Society in Ethics and Politics*, p. viii) Rawls uses what he calls the standard concept of rationality in social theory: "a rational person is thought to have a coherent set of preferences between the options open to him. He ranks these options according to how well they further his purposes; he follows the plan which will satisfy more of his desires rather than less, and which has the greater chance of being successfully executed." (*A Theory of Justice*, p. 143) According to Foot, "Irrational actions are those in which a man in some way defeats his own purposes, doing what is calculated to

be disadvantageous or to frustrate his own ends." ("Morality as a System of Hypothetical Imperatives," p. 162) Harsanyi: "rational behavior is simply behavior consistently pursuing some well defined goals, and pursuing them according to some well defined set of preferences or priorities." ("Morality and the theory of rational behavior," p. 42) Simon says, "Reason is wholly instrumental. It cannot tell us where to go; at best it can tell us how to get there. It is a gun for hire that can be employed in the service of any goals we have, good or bad." (*Reason in Human Affairs*, pp. 7-8) Nagel's endorsement is tentative: "If there is such a thing as practical reason, it does not simply dictate particular actions but, rather, governs the *relations* among actions, desires, and beliefs—just as theoretical reason governs the relations among beliefs and requires some specific material to work on." (Thomas Nagel, *The Last Word*, p. 107)

[150] According to Sen, "there are two predominant methods of defining rationality of behavior in mainline economic theory. One is to see rationality as internal consistency of choice, and the other is to identify rationality with maximization of self-interest." (Amartya Sen, *Ethics & Economics*, p. 12) In attributing the ethical view to Sidgwick, I rely on Parfit's impression, quoted in the text. For Moore, see *Principia Ethica*, sects. 59-61. Smart says, "Let us use the word 'rational' as a term of commendation for that action which is, on the evidence available to the agent, *likely* to produce the best results . . ." (J. J. C. Smart, "An outline of a system of utilitarian ethics," pp. 46-47) Brandt says, "I shall preempt the term 'rational' to refer to actions, desires or moral systems which survive maximal criticism by facts and logic." (Richard Brandt, *A Theory of the Good and the Right* , p. 10) And Hare: "I agree, therefore, with what I take to be Brandt's view, but put into my own words, that the rational action will be what is preferred when our *present* preferences have been exposed to facts and logic." (R. M. Hare, *Moral Thinking*, pp. 104-105; also see p. 214) Gibbard: "The rational act is what it makes sense to do, the right choice on the occasion." (Allan Gibbard, *Wise choices, apt feelings*, p. 7)

[151] R. M. Hare, *Moral Thinking*,, p. 190.

[152] John Rawls, *A Theory of Justice*, p. 142.

[153] David Lewis, "Prisoners' Dilemma is a Newcomb Problem," essay 26 of *Philosophical Papers Volume II*, pp. 303-304.

[154] Derek Parfit, *Reasons and Persons*, p. 130.

[155] Amartya Sen, *Ethics & Economics*, p. 15.

[156] David Gauthier, *Moral Dealing: Contract, Ethics and Reason*, p. 152.

---

[157] For Sidgwick's own, complex view, see *The Methods of Ethics*, p. 498. Sidgwick on ancient Greece: pp. 91-92; Spinoza: p. 89; Butler and Clarke: pp. 119-120 (for more on Butler, see pp. 205-206); Bentham: p. 119.

[158] Derek Parfit, *Reasons and Persons*, p. 129.

[159] For example, "Ethical Egoism" is ch. 6 of James Rachels' *The Elements of Moral Philosophy*.

[160] Derek Parfit, *Reasons and Persons*, p. 3.

[161] Parfit says that since it would take at least a book to explain his central concepts adequately, he won't waste time doing less. (*Ibid.*, p. ix)

[162] *Ibid.*, p. 192. The next quotation comes from p. 123.

[163] *The Methods of Ethics*, pp. 420-421. (The previous quotation is from p. 420.) Similarly: "Reason shows me that if my happiness is desirable and a good, the equal happiness of any other person must be equally desirable." (p. 403)

[164] *Ibid.*, p. 420; see also pp. 497-498.

[165] *Ibid.*, p. 382.

[166] Compare Singer: I shall use Sidgwick's phrase ["the point of view of the universe"] to refer to a point of view that is all-embracing, while not attributing any kind of consciousness or attitude to the universe, or any part of it that is not a sentient being. From this perspective, we can see that our sufferings and pleasures are very like the sufferings and pleasures of others; and there is no reason to give less consideration to the suffering of others, just because they are 'other.'" (*How Are We to Live?*, p. 222) By assuming an "all-embracing" perspective, Singer assumes that egoism is false.

[167] David O. Brink, "Rational Egoism and the Separateness of Persons," p. 102. Brink doesn't endorse this argument.

[168] Christine M. Korsgaard, *The sources of normativity* , p. 143.

[169] Henry Sidgwick, *The Methods of Ethics*, p. 498.

[170] Parfit attributes this view, with citations, to Butler, Sidgwick, Wiggins, Madell, Swinburne, Perry and Wachsberg. (*Reasons and Persons*, pp. 307-308)

[171] David O. Brink, "Rational Egoism and the Separateness of Persons," pp. 128, 105. The next quotation is from p. 105.

[172] *Ibid.*, p. 108.

[173] See Sarah Hrdy's classic work on competitive infanticide in *The Langurs of Abu* and "Infanticide Among Animals: A Review, Classification, and Examination of the Implications for the Reproductive Strategies of Females."

---

[174] Bishop Butler admonishes us not to confuse power and authority in "Fifteen Sermons." (*British Moralists I.*, p. 353, originally in *The Analogy of Religion*, p. 402)

[175] Jonathan Bennett suggested this position to me.

[176] Jan Narveson, *Morality and Utility*, p. 270.

[177] See Henry Sidgwick, *The Methods of Ethics*, p. 418; Thomas Nagel, *The Possibility of Altruism*, chs. xi and xii; Derek Parfit, *Reasons and Persons*, p. 140.

[178] Derek Parfit, *Reasons and Persons*, p. 140.

[179] I'd like to thank John O'Leary-Hawthorne for his help with this argument.

[180] Christine M. Korsgaard, *The sources of normativity*, p. 135. Korsgaard italicizes "social nature."

[181] As Nagel explains, this departs from his view in *The Possibility of Altruism*. (*The View From Nowhere*, p. 159) Sidgwick: "And it may be observed that most Utilitarians, however anxious they have been to convince men of the reasonableness of aiming at happiness generally, have not commonly sought to attain this result by any logical transition from the Egoistic to the Universalistic principle." (*The Methods of Ethics*, p. 498)

[182] *The View From Nowhere*, p. 162.

[183] *Ibid.*, p. 160.

[184] *Ibid.*, p. 161.

[185] See Derek Parfit, *Reasons and Persons*, pp. 502-503.

[186] See Peter Unger, *Living High and Letting Die: Our Illusion of Innocence* .

[187] Similar arguments are developed in Jan Narveson, "Moral Problems of Population," R. I. Sikora, "Utilitarianism: the Classical Principle and the Average Principle," pp. 413-416 and Thomas Schwartz, "Obligations to Posterity," pp. 10-12.

[188] See Derek Parfit, *Reasons and Persons*, ch. 16.

[189] James Woodward disagree with Parfit's handling of many Non-Identity cases, but he still thinks that premise 1 is false: "there are at least some Non-Identity cases in which one's reasons for making a certain choice can only be explained by reference to Q," where Q states that "If in either of two outcomes the same number of people would ever live, it would be bad if those who live are worse off, or have a lower quality of life, than those who would have lived." (James Woodward, "The Non-Identity Problem," p. 806; also see pp. 811-812 of Woodward's "Reply to Parfit.")

[190] Jan Narveson, "Utilitarianism and New Generations."

[191] Jonathan Bennett, "On Maximizing Happiness," p. 62.

[192] R. I. Sikora suggests that Bennett's principle is subject to Non-Identity style counterexamples in "Is It Wrong to Prevent the Existence of Future Generations?" pp. 162-163, fn. 18.

[193] Michael Tooley, *Abortion and Infanticide*, p. 272.

[194] *Ibid.*, p. 272. The next two quotations are from p. 272 and p. 243.

[195] For a similar example, see Parfit, *Reasons and Persons*, p. 375.

[196] See Tooley, *op. cit.*, sect. 7.33 (especially p. 262) and p. 268.

[197] *Ibid.*, p. 272. What does Tooley mean by "those natural resources . . . that make it possible for one to lead a satisfying life?" He does not mean that each natural resource is needed to lead a satisfying life, for then his application of the principle would be invalid. (On p. 273, Tooley uses the principle to show why a woman shouldn't have a handicapped child, even though such a child could lead a satisfying life.) I interpret the principle to mean "those natural resources . . . that typically make an important contribution to well-being."

[198] Parfit, *Reasons and Persons*, pp. 487-490. John Bigelow and Robert Pargetter take a different view in "Morality, Potential Persons and Abortion."

[199] David Gauthier, *Moral Dealing: Contract, Ethics and Reason*, p. 5.

[200] David Gauthier, *Morals By Agreement*, p. 16.

[201] Rawls talks about *rightness as fairness* as a program to apply Rawlsian methods outside the sphere of justice. See *A Theory of Justice*, pp. 17, 111.

[202] Derek Parfit makes a similar criticism. (*Reasons and Persons*, p. 393)

[203] Parfit considers only that version of (b) according to which the contractors do not know whether they will ever exist. This is incoherent because the contractors "cannot assume that, in the actual history of the world, it might be true that [they] never exist." (*Reasons and Persons*, p. 392) But the contractors can assume that they never exist *in society*. For this reason Parfit's arguments fail to show that contractualism cannot be applied to questions of population size.

[204] Derek Parfit, *Reasons and Persons*, p. 422. Also, see Parfit's other objections to the average principle in this section.

[205] This comes from Jonathan Glover, *Causing Death and Saving Lives*, p. 69.

[206] To support something like 3, Bennett says "someone might accept a principle enjoining the preservation of every species, or every animal species, or every instance of extreme physical complexity, or every form of life which is capable of moral reflection. . ." ("On Maximizing Happiness," p. 65) Bennett, however, believes 3 because he wants our great

biological and spiritual adventure to continue. (p. 66) Compare Gregory Kavka on "the collective enterprises of man" in "The Futurity Problem," pp. 196-198. Tooley says that 3 might be believed because one wants humankind's understanding of reality to advance, because one wants improved human interaction, or because one wants greater justice or fairer distribution of goods. (*Abortion and Infanticide*, pp. 257-258) In defense of something like 4, Bennett says, "I share Leibniz's liking for rich, organic complexity, and so the discovery that our world has more of it than we realized would be good news indeed." (*op. cit.*, p. 64) If so, then it would better for more happy people to exist because then there would be more rich, organic complexity.

[207] John Seabrook, "All in the Genes," p. 81.

[208] R. I. Sikora has a similar example. See "Utilitarianism: the Classical Principle and the Average Principle," pp. 414-416, and "Is It Wrong to Prevent the Existence of Future Generations?" p. 114 and elsewhere.

[209] This example is taken from Rob Reiner's movie, *This Is Spinal Tap*.

[210] Derek Parfit, *Reasons and Persons*, p. 487.

[211] R. M. Hare would support the Strong Thesis. See "Abortion and the Golden Rule."

[212] R. I. Sikora agrees. See "Is It Wrong to Prevent the Existence of Future Generations?" pp. 140-142.

[213] John Broome argues for Transitivity in *Weighing Goods*, pp. 10-12.

[214] See Larry S. Temkin, "Weighing Goods: Some Questions and Comments," pp. 361-363 and "A Continuum Argument for Intransitivity," pp. 193-194. Temkin formulates Transitivity in terms of "all things considered better than" instead of "intrinsically better than," but this doesn't matter, since he discusses only properties intrinsic to A, B and C.

[215] See Larry S. Temkin, "Weighing Goods: Some Questions and Comments," p. 363, and especially "A Continuum Argument for Intransitivity," sect. 4.

[216] Thomas Kuhn: "all revolutions involve, among other things, the abandonment of generalizations the force of which had previously been in some part that of tautologies. Did Einstein show that simultaneity was relative or did he alter the notion of simultaneity itself? Were those who heard paradox in the phrase 'relativity of simultaneity' simply wrong?" (*The Structure of Scientific Revolutions, second edition*, pp. 183-184)

[217] Derek Parfit, *Reasons and Persons*, p. 160.

[218] See Larry S. Temkin, "Intransitivity and the Mere Addition Paradox," pp. 180-183 and his discussion of the same material in "Rethinking the Good, Moral Ideals and the Nature of Practical Reasoning," sect. L.

[219] Larry S. Temkin, "Weighing Goods: Some Questions and Comments," p. 361, fn. 22.

[220] What happens to these poor suffering souls after the year in A, the century in B, and so on? Three variations are death, normal life and Heaven (God's apology). But I don't think what happens *later* affects the *intrinsic* disvalue of any of these possibilities.

[221] My response follows Derek Parfit, "Overpopulation and the Quality of Life," p. 160, fn. 12. Temkin ('A Continuum Argument for Intransitivity,' sect. 5) rebuts this objection differently.

[222] G. E. Moore, *Principia Ethica*, p. 28.

[223] "I'll let the theory take care of that one."—John O'Leary-Hawthorne.

[224] Edmund Gurney, *Tertium Quid*, Vol. I, p. 181. Compare Dostoevsky on pleasure: "In certain moments, I experience a joy that is unthinkable under ordinary circumstances, and of which most people have no comprehension. Then I feel that I am in complete harmony with myself and the whole world, and this feeling is so bright and strong that you could give up ten years for a few seconds of that ecstasy—yes, even your whole life" (from Geir Kjetsaa, *Fyodor Dostoevsky: A Writer's Life*, p. 149).

[225] Henry Sidgwick, *The Methods of Ethics*, pp. 123-124, fn. 1.

[226] Parfit prefers "the Century of Ecstasy" to "the Drab Eternity" of muzak and potatoes, and Griffin concurs. If so, then Parfit and Griffin would prefer A to Z. See Parfit, "Overpopulation and the Quality of Life," pp. 160-161. (This repeats his view in *Reasons and Persons*, pp. 498-499.) And see Griffin, *Well-Being*, p. 86.

[227] I'll state dogmatically why I believe this. There should be possibilities involving pain such that Y's greater duration outweighs X's greater intensity, Z's greater duration outweighs Y's greater intensity, but X's greater intensity outweighs—and not merely balances—Z's greater duration. Let me try to identify such possibilities in the first counterexample, which consists in the J-R headaches. R is a day of pain slightly worse than unconsciousness. R is preferable to J, which is five minutes of agony. I want the possibility earliest in the alphabet that is preferable to J, J, and an intermediary possibility. I'm not sure which is first preferable to J, but suppose P is. P is the six hour headache. Now consider J, P and the possibility equidistant from them:

J = five minutes of agony
M = forty minutes of a headache that is bad but not nearly as bad as J's
P = six hours of a headache that is bad but not nearly as bad as M's.

M's duration outweighs J's intensity; P's duration outweighs M's intensity; but J's intensity outweighs P's duration. So, M is worse than J; P is worse than M; but J is worse than P.

[228] The first such argument, to my knowledge, appeared in Donald Davidson, J. C. C. McKinsey and Patrick Suppes, "Outlines of a Formal Theory of Value, I," p. 146. The authors say, "We owe the inspiration for this example to Dr. Norman Dalkey of the Rand Corporation."

[229] In my discussion, X, Y and Z are possibilities. One cannot trade possibilities, but one can trade the means to making them obtain. I ignore this wrinkle in the text.

[230] Compare Robert Nozick, *The Nature of Rationality*, p. 140, fn.

[231] As reported by Temkin, "A Continuum Argument for Intransitivity," p. 209.

[232] See James Griffin, *Well-Being*, ch. VI for a nice discussion of measuring value.

[233] See Plato, *Protagoras*, 356b and 357 a-e; John Stuart Mill, *Utilitarianism*, ch. V, fn., three paragraphs before the end; Richard Brandt, *A Theory of the Good and the Right*, p. 255; James Griffin, *Well-Being*, pp. 75, 104. For three related, relevant arguments, see Henry Sidgwick, *The Methods of Ethics*, p. 94; G. E. Moore, *Principia Ethica*, p. 78; David Wiggins, "Weakness of Will, Commensurability, and the Objects of Deliberation and Desire," p. 267 (p. 255 as reprinted in *Essays on Aristotle's Ethics*).

[234] Nagel says, in *The View From Nowhere*: "What we aim to discover is not a new aspect of the external world, called value, but rather just the truth about what we and others should do and want." (p. 139) "The objective badness of pain, for example, is not some mysterious further property that all pains have, but just the fact that there is a reason for anyone capable of viewing the world objectively to want it to stop." (p. 144) Korsgaard says, "To talk about values . . . is not to talk about entities, either mental or Platonic, but to talk in a shorthand way about relations we have with ourselves and one another." (*The sources of normativity*, p. 138)

[235] Thomas Nagel, *The Possibility of Altruism*, p. 138.

[236] R. M. Hare, *Moral Thinking*, p. 46. Peter Railton says, slightly more cautiously, "I *am* assuming that when a choice is faced between satisfying interest X of A vs. satisfying interest Y of B, answers to the question 'All else equal, would it matter more to me if I were A to have X satisfied than if I were B to have Y satisfied?' will be relatively determinate and stable across individuals under conditions of full and vivid information." ("Moral Realism," pp. 190-191, fn.)

[237] John Stuart Mill, *Utilitarianism*, ch. 2, para. 8.

---

[238] Compare John Rawls, *A Theory of Justice*, pp. 186-187; J. J. C. Smart, "An outline of a system of utilitarian ethics," p. 32.

[239] Henry Sidgwick, *The Methods of Ethics*, pp. 142-143.

[240] John Rawls, *A Theory of Justice*, p. 557. A similar list of questions appears in J. J. C. Smart, "An outline of a system of utilitarian ethics," p. 34.

[241] Jeremy Bentham, *The Principles of Morals and Legislation*, IV. V. 5.

[242] John Stuart Mill, *Utilitarianism*, ch. V, fn., three paragraphs before the end.

[243] Henry Sidgwick, *The Methods of Ethics*, p. 413; also see pp. 84, 94, 110, 123, 133, 381. Compare James Griffin: "If the goodness of life consists in a lot of short-term pleasures or experiences, then to rank two courses of life we should clearly need to do a lot of totting-up." (*Well-Being*, p. 104)

[244] Compare William P. Alston, "Pleasure," p. 346; Richard Brandt, *A Theory of the Good and the Right*, p. 255; Hare, following Griffin and Harsanyi, *Moral Thinking*, p. 123.

[245] I learned about this sort of example from Derek Parfit. For a related discussion, see Parfit's *Reasons and Persons*, p. 431.

[246] On similar matters, see Peter Vallentyne, "Infinite Utility: Anonymity and Person-Centeredness" and Peter Vallentyne and Shelly Kagan, "Infinite Value and Finitely Additive Value Theory."

[247] See my discussion of A vs. Z in ch. 5.

[248] Francis Hutcheson, *A System of Moral Philosophy*, pp. 421-422.

[249] See Hutcheson, quoted above; John Stuart Mill, *Utilitarianism*, ch. 2; Anthony Quinton, *Utilitarian Ethics*, pp. 40-41; Rem B. Edwards, *Pleasures and Pains: A Theory of Qualitative Hedonism*, pp. 40-41, p. 113; Susan L. Feagin, "Mill and Edwards on the Higher Pleasures," p. 251; Fred Berger, *Happiness, Justice and Freedom*, pp. 38, 40; Elizabeth S. Anderson, "John Stuart Mill and Experiments in Living," p. 13.

[250] See Geoffrey Scarre, *Utilitarianism*, p. 92.

[251] John Stuart Mill, *Utilitarianism*, ch. 2, para. 8. Mill makes the same point in para. 6. Rem B. Edwards argues for qualitative hedonism similarly in *Pleasures and Pains: A Theory of Qualitative Hedonism*, ch. 6.

[252] John Stuart Mill, *Utilitarianism*, ch. 2, para. 6.

[253] Mill himself mentions these last three. (*Utilitarianism*, ch. 2, para. 6)

[254] Compare Sidgwick on Plato: "The philosopher, [Plato] argues, has tried both kinds of pleasure, sensual as well as intellectual, and prefers the delights of philosophic life; the sensualist ought therefore to trust his decision and follow his example. But who can tell that the philosopher's constitution is not such as to render the enjoyments of the senses, in

his case, comparatively feeble?" (*The Methods of Ethics*, p. 148; for Plato, see the *Republic*, Book IX, 582a-583a)

[255] John Stuart Mill, *Utilitarianism*, ch. 2, para. 4. Compare Rutherford's Leibniz: Donald Rutherford, *Leibniz and the Rational Order of Nature*, p. 50.

[256] See ch. 5.

[257] Actually, in ch. 5 I only commit myself to the thesis that fifty years of ecstasy is *not worse than* any finite duration of mild pleasure.

[258] Henry Sidgwick, *The Methods of Ethics*, p. 123.

[259] Similarly, see Richard Brandt, *A Theory of the Good and the Right*, p. 254 and pp. 264-265.

[260] With thanks to Jonathan Bennett.

[261] Thomas L. Carson believes, similarly, that pleasure is bad if its object *would be bad* had it all the empirical characteristics which the person believes it to have. ("Happiness, Contentment and the Good Life," pp. 388-389.)

[262] C. D. Broad, *Five Types of Ethical Theory*, p. 234.

[263] Michael J. Zimmerman, "On the Intrinsic Value of States of Pleasure," pp. 34-35.

[264] *The Nicomachean Ethics*, X. 5. Combining elements of (a) and (b), Haezrahi says, "pleasure and pain are morally neutral and dependent in their moral worth upon the moral standing of the thing or the activity they accompany." (Pepita Haezrahi, "Pain and Pleasure: Some Reflections on Susan Stebbing's View That Pain and Pleasure Are Moral Values," p. 72)

[265] Roderick M. Chisholm, *Brentano and Intrinsic Value*, p. 67. Chisholm refers us to Franz Brentano, *The Foundation and Construction of Ethics*, p. 172. Philosophers cite Plato's *Republic* in connection with (c), presumably referring to 585a-e, but there Plato doesn't commit himself to anything much like (c).

[266] This comes from the first paragraph of Kant's *Grounding for the Metaphysics of Morals*, p. 7. Also see, for example, W. D. Ross, *The Right and the Good*, p. 136.

[267] I have in mind the following theses: (A') unpleasures taken in bad intentional objects are good; (B') unpleasures accompanying noble acts of self-sacrifice are good; or perhaps (B") unpleasures accompanying bad behavior or activity are good; (D') deserved unpleasure is good. (Nothing seems to correspond to (C); unpleasures are not thought to be good because they depend on true beliefs, though sadly contemplating some unhappy truths might be thought good, as a special case of (A'). This asymmetry deserves further thought.) Those who say things like (A') typically use "displeasure" rather than "unpleasure,"

perhaps because "displeasure" connotes disapproval accompanied by mild unpleasure; *suffering* taken in bad objects always seems bad.

[268] Noah Lemos, *Intrinsic Value*, p. 73.

[269] *Ibid.*, p. 46.

[270] "An opinion, therefore, or belief may be most accurately defin'd, A LIVELY IDEA RELATED TO OR ASSOCIATED WITH A PRESENT IMPRESSION." (David Hume, *Treatise of Human Nature*, p. 96)

[271] Some philosophers prefer to say that pleasure's being taken in a bad intentional object *includes* a bad intentional object (rather than entails its existence). I'll use the language of entailment, but nothing turns on this.

[272] This Franz Brentano's view. See *The Foundation and Construction of Ethics*, p. 196.

[273] In ch. 1 I discussed contributive value.

[274] Thomas L. Carson, "Happiness, Contentment and the Good Life," p. 387. Carson actually writes, "I am convinced by the arguments of Brentano and others who hold that *Schadenfreude* is bad."

[275] Michael J. Zimmerman acknowledges that he has no argument on these matters. ("On the Intrinsic Value of States of Pleasure," pp. 29, 35) Also see G. E. Moore, *Principia Ethica*, p. 210; W. D. Ross, *The Right and the Good*, p. 137; C. D. Broad, *Five Types of Ethical Theory*, p. 234; Pepita Haezrahi, "Pain and Pleasure: Some Reflections on Susan Stebbing's View That Pain and Pleasure Are Moral Values," pp. 74-75; Thomas L. Carson, "Happiness, Contentment and the Good Life," p. 387; Roderick M. Chisholm, *Brentano and Intrinsic Value*, ch. 6; John O'Neill, "The Varieties of Intrinsic Value," p. 132; Noah Lemos, *Intrinsic Value*, pp. 73-77.

[276] Irwin Goldstein, "Pleasure and Pain: Unconditional, Intrinsic Values," p. 269. Noah Lemos gives the same argument on p. 44 of *Intrinsic Value*.

[277] Several philosophers who disbelieve this argument's conclusion agree with this premise: see Michael J. Zimmerman, "On the Intrinsic Value of States of Pleasure," p. 31; Thomas L. Carson, "Happiness, Contentment and the Good Life," p.387; Roderick M. Chisholm, *Brentano and Intrinsic Value*, p. 66; Noah Lemos, *Intrinsic Value*, p. 76 (see (7): "Pleasure in the merely neutral is intrinsically good").

[278] Compare Irwin Goldstein: "Since at least *some* pleasure is good intrinsically simply because of its pleasurableness, pleasure should always be good intrinsically, whatever the society, and so be an unconditional value." ("Pleasure and Pain: Unconditional, Intrinsic Values," p. 273)

[279] Arthur Schopenhauer, *The Basis of Morality*, pp. 156-157.

[280] Although, as Smart says, "Our repugnance to the sadist arises, naturally enough, because in our universe sadists invariably do harm. If we lived in a universe in which by some extraordinary laws of psychology a sadist was always confounded by his own knavish tricks and invariably did a great deal of good, then we should feel better disposed towards the sadistic mentality." ("An outline of a system of utilitarian ethics," pp. 25-26)

[281] See Derek Parfit, *Reasons and Persons*, pt. 2, ch. 8, especially sect. 62.

[282] See Daniel Kahneman's 1994 Tanner Lecture, "The Cognitive Psychology of Consequences and Moral Intuition" (unpublished). Also see Kahneman, B. L. Fredrickson, C. Schreiber, and D. A. Redelmeier, "When More Pain is Preferred to Less: Adding a Better End."

[283] Henry Sidgwick, *The Methods of Ethics*, p. 94.

[284] Karl Popper emphasizes imaginative moral reasoning that uses visual images. (*The Open Society and its Enemies*, Vol. II, pp. 232-233)

[285] Jonathan Glover, *Responsibility*, p. 93.

[286] Henry Sidgwick, *The Methods of Ethics*, p. 144. The next two quotations are from pp. 144-145 and p. 145.

[287] See Sidgwick's two chapters on empirical hedonism in *The Methods of Ethics*.

[288] Jeremy Bentham, *The Principles of Morals and Legislation* (1789), XVII. 1. IV, fn.

[289] Peter Singer, *Animal Liberation* (1990), p. 15.

[290] According to Norman Malcolm and Daniel C. Dennett, dreams are not experiences. See Malcolm's *Dreaming* and Dennett's "Are Dreams Experiences?" essay 8 of *Brainstorms*.

[291] Peter Singer discusses this in *How Are We to Live?*, ch. 10.

[292] John Stuart Mill, *Utilitarianism*, ch. 2, para. 13.

[293] John Rawls, *A Theory of Justice*, p. 409.

[294] For example, see Kai Nielsen's hedonist in "Hedonism and the Ends of Life," pp. 25-26.

[295] Michael Stocker, *Plural and Conflicting Values*, p. 284.

[296] See Derek Parfit, *Reasons and Persons*, sect. 29.

[297] Warren S. Quinn, "The Puzzle of the Self-Torturer," p. 79.

[298] The point about behavioral evidence was suggested to me by Alastair Norcross's "Comparing Harms: Headaches and Human Lives," pp. 142-143.

[299] G. W. Leibniz, *New Essays on Human Understanding*, p. 53. Leibniz says the "motion of a mill or a waterfall" on p. 53, but he means the sound that such motion makes; on p. 116 he refers to the water-mill's noise.

[300] John Stuart Mill, *Utilitarianism*, ch. 2, para. 12. Compare Karl Duncker: happiness, he says, "imparts its radiance to everything it happens to shine upon." ("On Pleasure, Emotion, and Striving," p. 405)

[301] Henry Sidgwick, *The Methods of Ethics*, p. 179. Similarly, on p. 125 he says: "so long as health is retained, and pain and irksome toil banished, the mere performance of the ordinary habitual functions of life is, according to my experience, a frequent source of moderate pleasures, alternating rapidly with states nearly or quite indifferent."

[302] Plato, *Philebus*, 44a. Socrates again condemns this idea in the *Republic*, Book IX, 583c-584c.

[303] Oliver Sacks, *The Man Who Mistook His Wife For a Hat And Other Clinical Tales*, p. 159.

[304] Thomas Nagel, "Death," *Mortal Questions*, p. 2.

[305] Aristotle, *The Nicomachean Ethics*, VII. 14.

[306] Parfit uses "well-being" and "beneficence" interchangeably, but his topic is well-being. Also, he sometimes says "human well-being" (pp. 370, 393, 394), but a satisfactory theory of well-being should also apply to nonhuman animals.

[307] Larry S. Temkin, "Intransitivity and the Mere Addition Paradox," p. 186.

[308] I discussed this principle in ch. 6.

[309] John Rawls, *A Theory of Justice*, pp. 186-187.

[310] Derek Parfit, *Reasons and Persons*, p. 363. The next quotation is from p. 371.

[311] See *Ibid.*, pp. 357-361.

[312] *Ibid.*, p. 447. The next quotation is from p. 387.

[313] Derek Parfit, *Reasons and Persons*, p. 388. Authors before Parfit charged classical or total utilitarianism with entailing what he calls the Repugnant Conclusion. See John Rawls, *A Theory of Justice*, pp. 162-163 and J. Brenton Stearns, "Ecology and the Indefinite Unborn," pp. 616-617.

[314] *Reasons and Persons*, p. 420.

[315] But the Average Principle *does* entail a variant of the Repugnant Conclusion. See Bill Anglin's ingenious essay, "The Repugnant Conclusion," *Canadian Journal of Philosophy*, Vol. VII, No. 4, December 1977, pp. 745-754.

[316] Derek Parfit, *Reasons and Persons*, p. 422. Parfit makes other strong objections to the Average Principle in this section.

[317] *Ibid.*, p. 405. According to Parfit, the Average Principle is but "one version" of the view that quality alone has value (pp. 402, 405), so he wouldn't identify quality with average well-being, as I do. Perhaps a view that enjoins maximizing the well-being of only the

best-off would entail, in Parfit's view, that quality alone has value. But I am not sure exactly what Parfit means by "quality."

[318] *Ibid.*, p. 402.

[319] *Ibid.*, p. 406.

[320] *Ibid.*, p. 411.

[321] *Ibid.*, p. 412.

[322] *Ibid.*, p. 415. Here and elsewhere I substitute "Mediocre" for "Valueless," with Parfit's permission (see p. 416).

[323] *Ibid.*, p. 528, fn. 40.

[324] *Ibid.*, p. 528, fn. 40.

[325] *Ibid.*, p. 415.

[326] *Ibid.*, p. 393.

[327] Parfit discusses the Second Paradox in *Reasons and Persons* pp. 433-437 and "Overpopulation and the Quality of Life" pp. 156-164. I follow Parfit's account in the latter essay in which he introduces his proposal for resolving it. I am reversing Parfit's order of discussion of the paradoxes because my arguments on how to resolve the Second Paradox will strengthen my suggestion for resolving the Mere Addition Paradox.

[328] Derek Parfit, "Overpopulation and the Quality of Life,'" p. 163.

[329] *Ibid.*, p. 164. The next quotation is from the same page.

[330] Larry Temkin suggested this to me in correspondence, but his published work also bears on these issues, especially *Inequality*, chs. 7 and 9.

[331] Temkin, *Inequality*, p. 290.

[332] Derek Parfit, *Reasons and Persons*, p. 420.

[333] See *Reasons and Persons*, pp. 419-430. For simplicity, I have omitted Parfit's "Divided B." This won't affect the arguments.

[334] See Larry Temkin, "Intransitivity and the Mere Addition Paradox," pp. 147-151.

[335] *Reasons and Persons*, p. 430.

[336] Temkin anticipates this type of resolution to the Mere Addition Paradox: "*However* A and Z compare to some intermediate world, or set of worlds, this does not entail how *they* compare if preferability is deeply intransitive." (Temkin, "Intransitivity and the Mere Addition Paradox," p. 157, fn. 24)

[337] *Reasons and Persons*, p. 452.

[338] *Reasons and Persons* (1987 reprinting only), p. 443.

[339] Socrates speaking in Plato's *Protagoras*, 351 c.

[340] John Broome, "More pain or less?" p. 117.

Bibliography

Alston, William P., "Pleasure," *The Encyclopedia of Philosophy* , Paul Edwards, ed. (New York: Macmillan, 1964).

—*Epistemic Justification* (Ithaca: Cornell University Press, 1979).

—"Epistemic Desiderata," *Philosophy and Phenomenological Research*, Vol. LIII, No. 3, September 1993.

Anderson, Elizabeth S., "John Stuart Mill and Experiments in Living," *Ethics*, 102 (1991).

Anglin, Bill, "The Repugnant Conclusion," *Canadian Journal of Philosophy*, Vol VII, No. 4, December 1977.

Aristotle, *The Nicomachean Ethics*, translated by David Ross, revised by J. L. Akrill and J. O. Urmson (Oxford University Press, 1980).

Armstrong, D. M., *A Materialist Theory of Mind* (London: Routledge & Kegan Paul, 1968).

Baier, Kurt, *The Moral Point of View* (Ithaca: Cornell University Press, 1958).

Beardsley, Monroe, "Intrinsic Value," *Philosophy and Phenomenological Research*, Vol. 26 (1965).

Benacerraf, Paul, "Mathematical Truth," *Journal of Philosophy*, 70 (1973).

Bennett, Jonathan, "On Maximizing Happiness," *Obligations to Future Generations*, R. I. Sikora and Brian Berry, eds. (Philadelphia: Temple University Press, 1978).

—*The Act Itself* (Oxford University Press, 1995).

Bentham, Jeremy, *The Principles of Morals and Legislation* (Buffalo, New York: Prometheus Books, 1988, originally 1789).

Berger, Fred, *Happiness, Justice and Freedom* (Los Angeles: University of California Press, 1984).

Bigelow, John and Robert Pargetter, "Morality, Potential Persons and Abortion," *American Philosophical Quarterly*, Vol. 25, No. 2, April 1988.

Bradley, F. H., *Ethical Studies* (Oxford University Press, 1876).

Brandt, Richard, *Ethical Theory* (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1959).

—*A Theory of the Good and the Right* (Oxford University Press, 1979).

—*Facts, Values and Morality* (Cambridge University Press, 1996).

Brentano, Franz, *The Foundation and Construction of Ethics*, Franziska Mayer-Hillebrand, ed. and Elizabeth Schneewind, trans. (London: Routledge and Kegan Paul, 1973).

Brink, David, "Rational Egoism and the Separateness of Persons," *Reading Parfit*, Jonathan Dancy, ed. (Oxford: Blackwell Publishers, 1997).

Broad, C. D., *Five Types of Ethical Theory* (London: Kegan Paul, Trench, Trubner & Co., Ltd., 1930).

—*Broad's Critical Essays in Moral Philosophy*, D. Cheney, ed. (London: George Allen & Unwin, 1971).

Broome, John, *Weighing Goods* (Oxford: Basil Blackwell, 1991).

—"More pain or less?" *Analysis* 56.2, April 1996.

Butler, Bishop Joseph, "Fifteen Sermons,"*British Moralists 1650-1800 I. Hobbes-Gay*, D.D. Raphael, ed. (Indianapolis: Hackett Publishing Company, 1991), originally in *The Analogy of Religion* (1736).

Byrne, Alex, "Some Like It HOT: Consciousness and Higher-Order Thoughts," *Philosophical Studies* 86, 1997.

Carlson, George R., "Pain and the Quantum Leap to Agent-Neutral Value," *Ethics* 100 (January 1990).

Carson, Thomas L., "Happiness, Contentment and the Good Life," *Pacific Philosophical Quarterly* 62 (1981).

Chalmers, David, *The Conscious Mind* (Oxford University Press, 1996).

Chisholm, Roderick M., "Objectives and Intrinsic Value," *Jenseits von Sein und Nichtsein*, Rudolph Haller, ed. (Graz: Akademisches Druck; und Verlagsanstalt, 1972).

—"Defining Intrinsic Value," *Analysis* , Vol. 41, No. 2 (March 1981).

—*Brentano and Intrinsic Value* (Cambridge University Press, 1986).

Paul M. Churchland, *Matter and Consciousness* (Cambridge, Mass.: The MIT Press, 1984).

—*A Neurocomputational Perspective* (Cambridge, Mass.: The MIT Press, 1989).

Copp, David, and David Zimmerman, eds., *Morality, Reason and Truth* (Totowa, New Jersey: Rowman & Allanheld, Publishers, 1984).

Davidson, Donald, J. C. C. McKinsey and Patrick Suppes, "Outlines of a Formal Theory of Value, I," *Philosophy of Science* 22 (1955).

Dennett, Daniel, *Brainstorms* (Cambridge, Massachusetts: The MIT Press, 1981).

de Sousa, Ronald, "Arguments from Nature," *Zygon*, vol. 15 (June 1980), reprinted in *Morality, Reason and Truth* (Totowa, New Jersey: Rowman & Allanheld, Publishers, 1984).

Duncker, Karl, "On Pleasure, Emotion, and Striving," *Philosophy and Phenomenological Research* 1 (1941).

Dworkin, Ronald, *Life's Dominion* (New York: Vintage Books, 1993).

Edwards, Rem B., *Pleasures and Pains: A Theory of Qualitative Hedonism* (Ithaca: Cornell University Press, 1979).

Feagin, Susan L., "Mill and Edwards on the Higher Pleasures," *Philosophy*, 58 (1983).

Feldman, Fred, *Doing the Best We Can* (Dordrecht: D. Reidel Publishing Company, 1986).

—"On the Intrinsic Value of Pleasures," *Ethics* 107 (April 1997).

Feibleman, J. S., "A Philosophical Analysis of Pleasure" in *The Role of Pleasure in Behavior*, R. G. Heath, ed. (New York: Harper and Row, 1964).

Findlay, J. N., *Values and Intentions* (New York: Macmillan, 1961).

Foot, Philippa, "Morality as a System of Hypothetical Imperatives," *Virtues and Vices* (Berkeley: University of California Press, 1978).

Gauthier, David, *Morals By Agreement* (Oxford University Press, 1986).

—*Moral Dealing: Contract, Ethics and Reason* (Ithaca: Cornell University Press, 1990).

Gewirth, Alan, *Reason and Morality* (The University of Chicago Press, 1978).

Gibbard, Allan, *Wise choices, apt feelings* (Cambridge, Mass.: Harvard University Press, 1990).

Glover, Jonathan, *Responsibility* (New York: Humanities Press, 1970).

—*Causing Death and Saving Lives* (New York: Penguin Books, 1977).

Goldstein, Irwin, "Why People Prefer Pleasure to Pain," *Philosophy* 55 (1980).

—"Pleasure and Pain: Unconditional, Intrinsic Values," *Philosophy and Phenomenological Research*, Vol. 1, No. 2, December 1989.

Grahek, Nikola, "Objective and subjective aspects of pain," *Philosophical Psychology*, Vol. 4, No. 2, 1991.

Griffin, James, "Modern Utilitarianism" *Revue Internationale de Philosophie*, No. 141 (1982).

—*Well-Being* (Oxford University Press, 1986).

Gurney, Edmund, *Tertium Quid*, vol. I (London: Kegan Paul, Trench & Co., 1887).

Guzeldere, Guven, "Is Consciousness the Perception of What Passes in One's Own Mind?" *Conscious Experience*, Thomas Metzinger, ed. (Schoningh: Imprint Academic, 1995).

Haezrahi, Pepita, "Pain and Pleasure: Some Reflections on Susan Stebbing's View That Pain and Pleasure Are Moral Values," *Philosophical Studies*, Vol. 11 (1960).

Hall, Richard J., "Are Pains Necessarily Unpleasant?" *Philosophy and Phenomenological Research* Vol. XLIX, No. 4, June 1989.

Hare, R. M., "Pain and Evil," *Essays on the Moral Concepts* (Berkeley: University of California Press, 1972).

—"Abortion and the Golden Rule," *Philosophy and Public Affairs* 4 (Spring 1975), reprinted in *Philosophy and Sex*, eds.: Robert Baker and Frederick Elliston (New York: Prometheus Books, 1984).

—*Moral Thinking* (Oxford University Press, 1981).

Harman, Gilbert, "Toward a Theory of Intrinsic Value," *The Journal of Philosophy*, Vol. 64 (1967).

Harsanyi, John C.,"Morality and the theory of rational behavior," *Utilitarianism and Beyond*, Amartya Sen and Bernard Williams, eds. (Cambridge University Press, 1982).

Hrdy, Sarah, *The Langurs of Abu* (Cambridge, Mass.: Harvard University Press, 1977).

—' Infanticide Among Animals: A Review, Classification, and Examination of the Implications for the Reproductive Strategies of Females," *Ethology and Sociobiology* 1 (1979).

Hume, David, *Treatise of Human Nature* (Buffalo, New York: Prometheus Books, 1992, originally 1739).

—*An Enquiry Concerning the Principals of Morals*, J. B. Schneewind, ed., (Indianapolis: Hackett Publishing Company, 1983, originally 1751).

Hutcheson, Francis, *A System of Moral Philosophy*, in *British Moralists*, Vol. 1, L. A. Selby-Bigge, ed. (Oxford University Press, 1897).

Kagan, Shelly, *The Limits of Morality* (Oxford University Press, 1989).

—and Peter Vallentyne, "Infinite Value and Finitely Additive Value Theory," *The Journal of Philosophy* Vol. XCIV, No. 1, January 1997.

Kahneman, Daniel, "The Cognitive Psychology of Consequences and Moral Intuition," 1994 Tanner Lectures on Human Values (unpublished).

—and B. L. Fredrickson, C. Schreiber, and D. A. Redelmeier, "When More Pain is Preferred to Less: Adding a Better End," *Psychological Science* 4.

Kant, Immanuel, *Grounding for the Metaphysics of Morals* in *Immanuel Kant: Ethical Philosophy* , James W. Ellington, trans. (Indianapolis: Hackett Publishing Company, 1983, originally 1785).

Kavka, Gregory, "The Futurity Problem," *Obligations to Future Generations*, R. I. Sikora and Brian Berry, eds. (Philadelphia: Temple University Press, 1978).

Kitcher, Philip, *The Nature of Mathematical Knowledge* (Oxford University Press, 1984).

Kjetsaa, Geir, *Fyodor Dostoevsky: A Writer's Life*, tr. Siri Hustvedt and David McDuff (New York: Viking, 1987).

Korsgaard, Christine M., *The Sources of Normativity* (Cambridge University Press, 1996).

—*Creating the Kingdom of Ends* (Cambridge University Press, 1996).

Kuhn, Thomas, *The Structure of Scientific Revolutions*, 2nd ed. (The University of Chicago Press, 1970).

Leibniz, Gottfried, *New Essays on Human Understanding*, translated and edited by Peter Remnant and Jonathan Bennett (Cambridge University Press, 1981).

Lemos, Noah, *Intrinsic Value: Concept and Warrant* (Cambridge University Press, 1994).

Lewis, C. I., *An Analysis of Knowledge and Valuation* (La Salle, Ill.: Open Court Publishing Company, 1946).

Lewis, David, "Extrinsic Properties," *Philosophical Studies* 44 (1983).

—*Philosophical Papers Volume II* (New York: Oxford University Press, 1986).

—"II-Dispositional Theories of Value" (The Aristotelian Society Supplementary Volume LXIII, 1989).

Mackie, J. L., *Ethics: Inventing Right and Wrong* (Harmondsworth, England: Penguin Books, 1977).

Malcolm, Normal, *Dreaming* (London: Routledge & Kegan Paul, 1959).

McDowell, John, *Mind and World* (Cambridge, Mass.: Harvard University Press, 1994).

McKinsey, J. C. C., Donald Davidson and Patrick Suppes, "Outlines of a Formal Theory of Value, I," *Philosophy of Science* 22 (1955).

Melzack , R. and P. D. Wall, *The Challenge of Pain* (Hammondsworth: Penguin Books, 1982).

Mill, John Stuart, *Utilitarianism*, reprinted in *Mill: Utilitarianism*, Samuel Gorovitz, ed. (Indianapolis: The Bobbs-Merrill Company, Inc., 1971, originally 1861).

Moore, G.E., *Principia Ethica* (Cambridge University Press, 1903).

—*Philosophical Studies* (London: Routledge & Kegan Paul, 1922).

Morillo, Carolyn, *Contingent Creatures* (London: Littlefield Adams Books, 1995).

Nagel, Thomas, *The Possibility of Altruism* (Princeton University Press, 1970).

—*Mortal Questions* (Cambridge University Press, 1979).

—*The View From Nowhere* (Oxford University Press, 1986).

—*Equality and Partiality* (Oxford University Press, 1991).

—"Universality and the reflective self," in Christine M. Korsgaard, *The sources of normativity* (Cambridge University Press, 1996).

—*The Last Word* (Oxford University Press, 1997).

Narveson, Jan, *Morality and Utility* (Baltimore: the Johns Hopkins Press, 1967).

—"Utilitarianism and New Generations," *Mind* vol. 76 (1967).

—"Moral Problems of Population," *Monist* 57 (1973).

Nelkin, Norton, "Reconsidering Pain," *Philosophical Psychology*, Vol. 7, No. 3, 1994.

Nielsen, Kai, "Hedonism and the Ends of Life," *Philosophical Journal*, Vol. 10, No. 1 (January 1973).

Norcross, Alastair, "Comparing Harms: Headaches and Human Lives," *Philosophy & Public Affairs*, Vol. 26, No. 2 (Spring 1997).

Nozick, Robert, *The Nature of Rationality* (Princeton University Press, 1993).

Oates, Whitney J., ed., *The Stoic and Epicurean Philosophers*, (New York: Random House, 1940).

O'Neill, John, "The Varieties of Intrinsic Value," *The Monist*, April 1992 (Vol. 75, No. 2).

Parfit, Derek, *Reasons and Persons* (Oxford University Press, 1984).

—"Overpopulation and the Quality of Life," *Applied Ethics*, Peter Singer, ed. (Oxford University Press, 1986).

Pettit, Philip and Michael Smith, "Parfit's P," *Reading Parfit*, Jonathan Dancy, ed. (Oxford: Blackwell Publishers, 1997).

Pitcher, George, "Pain Perception," *The Philosophical Review*, July 1970.

Plato, *Republic* (translated by Paul Shorey). All Plato is in *Plato: The Collected Dialogues*, Edith Hamilton and Huntington Cairns, eds. (Princeton University Press, 1961).

—*Protagoras* (translated by W. K. C. Guthrie).

—*Philebus* (translated by R. Hackforth).

Popper, Karl, *The Open Society and its Enemies*, Vol. II (Princeton University Press, 1963).

Quinn, Warren S., "The Puzzle of the Self-Torturer," *Philosophical Studies* 59 (1990).

Quinton, Anthony, *Utilitarian Ethics* (La Salle, Ill.: Open Court Publishing, 1973).

Rachels, James, *The End of Life* (Oxford University Press, 1986).

—*The Elements of Moral Philosophy*, 3rd ed. (New York: McGraw-Hill, Inc., 1998).

Railton, Peter, "Moral Realism," *The Philosophical Review*, XCV, No. 2 (April 1986).

—"Naturalism and Prescriptivity," *Social Philosophy & Policy* Vol. 7, Issue 1 (1989).

Rawls, John, "Outline of a Decision Procedure For Ethics," *The Philosophical Review* (60), 1951.

—*A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971).

Ross, W. D., *The Right and the Good* (Oxford University Press, 1930).

Russell, Bertrand, *Human Society in Ethics and Politics* (London: Allen and Unwin, 1954).

Rutherford, Donald, *Leibniz and the Rational Order of Nature* (Cambridge University Press, 1995).

Ryle, Gilbert, *Dilemmas* (Cambridge University Press, 1958).

Sacks, Oliver, *The Man Who Mistook His Wife For a Hat And Other Clinical Tales* (New York: Harper & Row, 1987).

Scarre, Geoffrey, *Utilitarianism*(London: Routledge, 1996).

Schopenhauer, Arthur, *The Basis of Morality*, A. Broderick Bullocks, trans. (London: Swan Sonneschein, 1903).

Schwartz, Thomas, "Obligations to Posterity," *Obligations to Future Generations*, R. I. Sikora and Brian Berry, eds. (Philadelphia: Temple University Press, 1978).

Seabrook, John, "All in the Genes," *The New Yorker*, February 12, 1996.

Sen, Amartya K., *Ethics & Economics* (Oxford: Basil Blackwell, 1987).

Sher, George, *Desert* (Princeton University Press, 1987).

Sidgwick, Henry, *The Methods of Ethics*, 7th ed. (Indianapolis: Hackett Publishing Company, 1906; repr. 1981).

Sikora, R. I., "Utilitarianism: the Classical Principle and the Average Principle," *Canadian Journal of Philosophy*, Volume V, Number 3, November 1975.

—"Is It Wrong to Prevent the Existence of Future Generations?" *Obligations to Future Generations*, eds.: R. I. Sikora and Brian Berry (Philadelphia: Temple University Press, 1978).

Simon, Herbert, *Reason in Human Affairs* (Stanford University Press, 1983).

Singer, Peter, *Animal Liberation* (New York Review Books, 1975).

—*Animal Liberation*, revised edition (New York: Avon Books, 1990).

—*How Are We to Live?* (New York: Prometheus Books, 1995).

Smart, J. J. C., "An outline of a system of utilitarian ethics," *Utilitarianism for and against* (with Bernard Williams) (Cambridge University Press, 1973).

Smith, Michael and Philip Pettit, "Parfit's P," *Reading Parfit*, Jonathan Dancy, ed. (Oxford: Blackwell Publishers, 1997).

Sprigge, T. L. S., *The Rational Foundations of Ethics* (London: Routledge & Kegan Paul, 1988).

Stearns, J. Brenton, "Ecology and the Indefinite Unborn," *The Monist* Vol. 56, No. 4 (October, 1972).

Stocker, Michael, *Plural and Conflicting Values* (Oxford: Clarendon Press, 1990).

Suppes, Patrick, Donald Davidson and J. C. C. McKinsey, "Outlines of a Formal Theory of Value, I," *Philosophy of Science* 22 (1955).

Temkin, Larry, "Intransitivity and the Mere Addition Paradox," *Philosophy and Public Affairs*, Vol. 16, No. 2 (Spring 1987).

—*Inequality* (Oxford University Press, 1993).

—"Weighting Goods: Some Questions and Comments," *Philosophy and Public Affairs* 23 (1994).

—"A Continuum Argument for Intransitivity," *Philosophy and Public Affairs* 25 (1996).

—"Rethinking the Good, Moral Ideals, and the Nature of Practical Reasoning: A Critical Examination of Part Four of *Reasons and Persons*," *Reading Parfit*, Jonathan Dancy, ed. (Oxford: Blackwell Publishers, 1997).

Tooley, Michael, *Abortion and Infanticide* (Oxford University Press, 1983).

Trigg, Roger, *Pain and Emotion* (Oxford University Press, 1970).

Tye, Michael, *Ten Problems of Consciousness* (Cambridge, Mass.: The MIT Press, 1995).

Unger, Peter, *Living High and Letting Die: Our Illusion of Innocence* (New York: Oxford University Press, 1996).

Vallentyne, Peter, "Infinite Utility: Anonymity and Person-Centeredness," *Australasian Journal of Philosophy*, Vol. 73, No. 3; September 1995.

—and Shelly Kagan, "Infinite Value and Finitely Additive Value Theory," *The Journal of Philosophy* Vol. XCIV, No. 1, January 1997.

van Inwagen, Peter, *Metaphysics* (Boulder: Westview Press, 1993).

von Wright, Georg Henrik, *The Varieties of Goodness* (London: Routledge & Kegan Paul, 1963).

Wall , P. D. and R. Melzack, *The Challenge of Pain* (Hammondsworth: Penguin Books, 1982).

Westermarck, Edward, *The Origin and Development of the Moral Ideas* (New York: Macmillan, 1906).

—*Ethical Relativity* (New York: Harcourt, Brace, 1932).

David Wiggins, "Weakness of Will, Commensurability, and the Objects of Deliberation and Desire," *Proceedings of the Aristotelian Society*, NS 79 (1978-79), reprinted in A. O. Rorty, ed., *Essays on Aristotle's Ethics* (Berkeley: University of California Press, 1980).

Wilkes, K., *Physicalism* (London: Routledge & Kegan Paul, 1977).

Woodward, James, "The Non-Identity Problem," *Ethics* 96 (July 1986).

—"Reply to Parfit," *Ethics* 97 (July 1987).

Zimmerman , David, and David Copp, eds. , *Morality, Reason and Truth* (Totowa, New Jersey: Rowman & Allanheld, Publishers, 1984).

Zimmerman, Michael J., "On the Intrinsic Value of States of Pleasure," *Philosophy and Phenomenological Research* (41), 1980-81.

Curriculum Vitae

Stuart Craig Rachels

b. September 26, 1969 in New York City

Degrees Awarded

B.A. with Highest Honors (Philosophy), Emory University, 1991.
Thesis director: Donald Rutherford.

B.A. (Philosophy and Politics), University of Oxford, 1993.
Thesis director: Derek Parfit.

Teaching

As sole instructor:
Philosophy 107, "Theories of Knowledge and Reality" (eight sections).

As teaching assistant:
Philosophy 191, "Ethics and Value Theory" (six sections).
Philosophy 251, "Logic" (three sections).

Academic Honors and Awards

Dean's Scholarship, Emory University, 1989-91

Paul Kuntz Award, Emory University, 1991

Marshall Scholarship, Oxford University, 1991-93

Syracuse University Dissertation Fellowship, 1996-97

Publications

"Counterexamples to the Transitivity of *Better Than*." *Australasian Journal of Philosophy*, March 1998.

"Is It Good to Make Happy People?" *Bioethics*, April 1998.

"Intransitivity." Solicited for the second edition of the *Encyclopedia of Ethics*, edited by Lawrence C. Becker and Charlotte Becker.

"Is Unpleasantness Intrinsic to Experience?" Forthcoming in *Philosophical Studies*

Papers Presented

"Is Pain Intrinsically Unpleasant?" Western Washington University, 1998.

"Counterexamples to Transitivity," University of Colorado, 1998; University of Cincinnati, 1998.

"Reconceiving *Better Than*," Pacific APA Meeting, San Francisco, 1995.

"An Argument Against Physicalism," Mid-South Philosophy Conference, Memphis, 1991.

Miscellaneous

1989 U.S. Chess Champion; numerous publications in chess magazines.

1992-93 MCR President, Hertford College, University of Oxford.