The Dissertation Committee for Jake David Galgon
certifies that this is the approved version of the following dissertation:

Why People Don't Matter

And What to Do About That

Committee:

_____
Galen Strawson, Co-Supervisor

_____
Paul B Woodruff, Co-Supervisor

_____
Marya S Schechtman

_____
Sinan Dogramaci

Why People Don't Matter

And What to Do About That

by

Jake David Galgon, A.B.

Dissertation

Presented to the Faculty of the Graduate School

Of the University of Texas at Austin

In Partial Fulfillment

Of the Requirements

For the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2017

Dedicated to my mom, Florence Ann Roberts

my dad, Gerry Galgon

my brother, Geoff Galgon

and to the memories of Derek Parfit and Randy Poffo

Acknowledgements

I would like to thank the members of my dissertation committee—Galen Strawson, Paul Woodruff, Marya Schechtman, and Sinan Dogramaci—and everyone else who helped me along the way—Kari Rosenfeld, Kathy Hintikka, Andrew Ingram, Simone Gubler, Casey Woolwine, Miriam Schoenfield, Jonathan Dancy, Katherine Dunlop, Jonathan Drake, Alex Grossman, Glenavin White, Elliot Goodine, Elissa Shopoff, Matthias Barker, the PWR Multiverse, the employees of Tyson's, Rice Bowl, Barfly's, and Dolce Vita, and the many others I'm forgetting.

Why People Don't Matter

And What to Do About That

by

Jake David Galgon, PhD

The University of Texas at Austin, 2017

CO-SUPERVISORS: Galen Strawson, Paul B Woodruff

Reductionism about personal identity is the view that facts about personal identity reduce to lower-level facts about things like psychological or physical connectedness. In this dissertation, I give arguments for reductionism and for Derek Parfit's "Extreme Claim" that reductionism requires a radical revision of our ordinary normative thought. After detailing the extent of this revision, I introduce and describe a special sort of self-alienation that is likely to be engendered by a genuine belief in Extreme Claim Reductionism. I argue that this alienation cannot and should not be eliminated, and consider existing attempts to eliminate similar sorts of alienation and note where they seem to fall short of their aim. I then outline a practical strategy for living with Extreme Claim Reductionism and the alienation that accompanies it.

Table of Contents

Chapter One – Why People Don't Matter

Why and in what sense is the person writing this sentence the same person as the one who convinced his mom to let him watch *Jurassic Park* in a theater for his ninth birthday? Or, for that matter, the same person as the one who made tacos for lunch a few hours ago? These are questions about personal identity, and they are deceptively difficult. One popular family of views, which we can call reductionism, holds that questions about personal identity reduce to questions about such things as the psychological or physical features of and connections between the two people in question.

Derek Parfit's 1984 book *Reasons and Persons* advocates an extreme reductionist view of personal identity that, he argues, has radical implications for practical ethics. As far as I can tell, the book marks the furthest-out he ever got. His earlier 1971 paper "Personal Identity" is largely concerned with the view that what matters about personal survival is not identity but certain psychological relations which come in, and matter in, degrees.[1] These relations don't matter in the way that we might have thought heavyweight numerical personal identity would matter, but they do seem to matter. By 2011's colossal *On What Matters* (which is a work in ethical theory that doesn't get into questions of personal identity) he takes it as intuitive and unproblematic that we have some reasons to

[1] E.g. p. 26

value our own present and future good and the good of people we care for over the good of strangers. But in *Reasons and Persons*, Parfit entertains, but remains agnostic about, the "Extreme Claim" that on the reductionist view we have no reason to care any more about our own futures than other people's.[2]

I'm going to defend the Extreme Claim. In fact, I'm going to defend a very strong version of the claim: If reductionism is true, then nothing matters in anything like the way in which personal identity is ordinarily taken to matter.[3] Combining the Extreme Claim with reductionism, as I believe we should, yields what I will call Extreme Claim Reductionism, or ECR for short. In holding to ECR, I am willing to bite all the bullets that need to be bit. I will have to admit, for instance, that I would have more reason to save two strangers from drowning than my closest friend.

This may seem unhinged. If it does, it would not always have been so.

---

[2] *Reasons and Persons* – henceforth *RP* – pp. 307–312. Parfit characterizes the Extreme Claim in several (perhaps incompatible) ways. Initially, the Extreme Claim is that "we have *no* reason to be concerned about our own futures" (p. 307).

But Parfit also counts Perry's view that he has some reason to prevent his own future pains, but no *more* than he would have to prevent a stranger's, as a version of the Extreme Claim (p. 308).

Swinburne's claim that psychological connectedness and continuity do not matter on their own also counts (p. 308—note that Swinburne believes that there *is* a further identity fact, so it is untroubling that connectedness and continuity do not matter on their own).

The version of the Extreme Claim that I will be defending seems to me most similar to the one attributed to Swinburne: Neither psychological connectedness and continuity nor any of the other relations that we might think identity reduces to matter in an identity-like way in the absence of a deep further fact. But unlike Swinburne, and like Parfit, I believe that there is no deep further fact.

[3] My version of the claim, like Swinburne's but unlike some that Parfit considers, is thus explicitly a conditional.

Before the rise of virtue ethics, intuitionism, particularism, care ethics, and the like over the course of the twentieth century, impartialist ethical systems like utilitarianism, Moorean consequentialism, and Kantianism reigned supreme. There seems to me to be an undeniable truth in these grim, impersonal, uncompromising theories. With Parfit (or at least with the Parfit of *Reasons and Persons*) I believe that reflection on the nature of persons supports an impartialist view. For instance, Kant's command that we never make exceptions of ourselves seems particularly plausible when temporally extended selves disappear or dissolve under scrutiny. Could anything be more arbitrary than making an exception of something that can hardly be said to exist?

My basic argument for Extreme Claim Reductionism will take this form: Parfit's own reasoning suggests a highly permissive view of which other people should matter for us in the way that our ordinary future and past selves do, which would even include such persons as causally unconnected duplicates.[4] If causally unconnected duplicates matter for us in the same ways that our past and future selves do, and if causally unconnected duplicates can ground none of the special reasons that we ordinarily take our past and future selves to ground, then ECR is true. I will argue that causally unconnected duplicates do not, in fact, ground any of these special reasons. ECR threatens the justificatory foundations

[4] A paradigm case of a causally unconnected duplicate, if the notion is confusing: A person on a far-off planet who, by sheer chance, is exactly microphysically and psychologically like me.

of partial reasons, responsibility, desert, and the intrinsic value of things like survival and friendship even more than the view Parfit explicitly adopts. But all this comes later. First, I need to give my reading of Parfit's view and of how he gets to it.

Parfitian Reductionism

So: Why and in what sense is the person writing this sentence the same person as the one who convinced his mom to let him watch *Jurassic Park* in a theater for his ninth birthday? The *Reasons and Persons* view is that, once we know all the physical and psychological facts about two people (or, if you like, two person-stages) there is no deep "further fact"[5] as to whether they are, in fact, the same person (or, if you like, stages of the same person).[6] We can say that they inhabit the same biologically continuous body and that they share various personality traits and links of memory and intention. If we want to say, as a further matter, that they are the *same person*, we're reporting something like a conventional fact.[7] He calls his view (a form of) reductionism and notes that, at

[5] The "further fact" language is Parfit's. He uses it, e.g. in *Reasons and Persons* p. 210, "Personal Identity" pp. 3–4, and "Is Personal Identity What Matters" p. 13.

[6] Parfit seems to be loose when talking about what sort of entities—persons or person-stages or whatever else—we ought to be comparing when we ask questions about personal identity. We might think, with Lewis ("Survival and Identity") that these details matter for the sorts of metaphysical claims that Parfit is entitled to make. But I do not think that they should matter for the sorts of practical ethical considerations that I am focusing on.

[7] "Experiences, Subjects, and Conceptual Schemes" p. 218: "We can imagine cases in which questions about our identity would be indeterminate: having no answers. These

least at this broad level of specificity, it is shared by a number of others.[8]

Of course, in ordinary cases, people's physical and psychological features usually behave in relatively predictable ways, and it's easy to give non-controversial verdicts about personal identity. But this needn't always be so. Parfit considers cases in which people undergo brain transplants and personality-altering surgeries, split like amoebas, live for centuries, or put themselves through teletransporters that reconstruct perfect copies of their brains and bodies out of new matter in distant locations.[9]

I'll rehearse two of his arguments from cases as illustrative examples of how the arguments go.

First, The Combined Spectrum:[10] A surgeon intends to perform a procedure that alters your brain and body. At the end of the procedure, all that's

---

questions would also be in the following sense *empty:* they would not be about different possibilities, but only about different descriptions of the same course of events." Parfit explicitly means this statement to be compatible with earlier expositions of his view (p. 217).

Another example from "Is Personal Identity What Matters?" pp. 6–7: "According to Constitutive Reductionism, the fact of personal identity is distinct from these facts about physical and psychological continuity. But, since it just consists in them, it is not an independent or separately obtaining fact. It is not a further difference in what happens. To illustrate that distinction, consider a simpler case. Suppose that I already know that several trees are growing together on some hill. I then learn that, because that is true, there is a copse on this hill. That would not be new factual information. I would have merely learnt that such a group of trees can be called a 'copse'. My only new information is about our language. That those trees can be called a copse is not, except trivially, a fact about the trees."

[8] In note 43 to *RP* part three (p. 518) he names Grice, Ayer, Mackie, Perry, Lewis, and Shoemaker. By now, there are more names that we could add to the list.

[9] These cases are spread across part three of *RP*.

[10] *RP* pp. 437-443.

left will be a duplicate of a different person. (In Parfit's case it's Greta Garbo.) If the surgeon only barely begins the procedure, perhaps changing your hair color and removing a few memories, it seems clear that the resulting person is you. If she completes the procedure, or stops just short of completion so that all that remains of you are a few moles and the apparent memory of watching *Jurassic Park*, it seems clear that the resulting person is *not* you in the ordinary sense. If there is an all-or-nothing further fact about personal identity, then there must be some critical point in the procedure where a few cells are replaced and you go out of existence, even though there could perhaps never be any evidence for where this point is. Moreover, if the further fact is what matters, then it is presumably a catastrophe to replace those few critical cells, and not only a very little bit worse than stopping the operation a second earlier. Rather than accept these conclusions, Parfit argues that we should be reductionists.

Second, Fission:[11] Consider a surgery that would split you, like an amoeba, into two bodies. Both of the resulting people would have apparent memories of your life. It makes little sense to say that you would be identical with both resulting people, since they are not identical with one another. It is also bad to say that you will be identical to either one over the other, since neither has a stronger claim to being you than the other does. Parfit claims that it is best to say that ordinary identity concepts fail and that—if we have to say anything at

[11] There are versions of this argument in *RP* pp. 253–266 and "Personal Identity" pp. 4-14.

all about personal identity—we should say that you are identical with neither. This means that you do not survive the operation in the usual sense, since after the operation there will be no one alive who is you, as survival would seem to require. But surely this strange double-survival is not as bad as death, and so personal identity is not what matters.

There seem to be no clear answers about personal identity in cases like fission and the combined spectrum, and certainly none that respect the notion that personal identity is a normatively significant, all-or-nothing further fact. And so, says Parfit, we ought to reject heavyweight further fact personal identity and talk only about the physical and psychological relations between persons or person stages. That is: We ought to be reductionists.

But what sort of reductionists? Jettisoning personal identity does not immediately reveal the answers to the practical questions one might have about Parfit's imagined cases. Should you fear the fission operation as being as bad as death, given that you will not be either of the resulting people? Is it *prudent* to go through the teletransporter? What if conventional travel is riskier? (How much riskier would it have to be before teletransportation became preferable?) Should it matter to you now which of your post-fission duplicates will suffer some misfortune, if you know that one of them has to? And so on.

Anti-reductionists can say about these cases: What matters is whether the teletransported copy, or the surgically-altered person on the operating table, or

the amoeba-like copy, or whatever else, *is you*. They can say this even if we do not (or cannot) know whether these people are you. The reductionist cannot appeal to such identity facts. But if they are a Parfitian reductionist, there are other facts they can appeal to in deciding how teletransportation, division, etc. stack up against ordinary survival.

Parfit's positive practical view is this: What actually matters in the way in which personal identity is ordinarily taken to matter, if anything does in fact matter in that way, is:

> Relation R: "psychological connectedness and/or psychological continuity, with the right kind of cause."[12]

We can ask: What is psychological connectedness? What is psychological continuity? And what is the right kind of cause? Parfit defines connectedness as "the holding of particular direct psychological connections,"[13] where direct psychological connections include such things as experiential memory and intention as well as persistent beliefs, desires, and other psychological features.[14]

None of the sorts of connection that R comprises are meant to presuppose personal identity. Where they seem to, as perhaps in the case of memory, we are

[12] He gives this definition several times, e.g. on *RP* p. 215, p. 262, and p. 271.
[13] *RP* p. 206
[14] *RP* p. 205

to substitute analogues that do not. So rather than experiential memory, which might require that the subject of a remembered experience be the same person as the rememberer, we can talk about experiential "quasi-memory," which requires only that the remembered experience did happen and that it is causally linked, in the right way, to the quasi-rememberer.[15]

Continuity is defined in terms of connectedness: It is "the holding of overlapping chains of *strong* connectedness" and is, in theory and practice, less important than connectedness.[16]

Crucially, the right sort of cause of psychological connectedness and continuity for Parfit is any cause whatsoever.[17]

So, Parfitian reductionism makes two claims: First, that reductionism is true—that is, that there are no deep further facts about personal identity. Second, that what matters in the way that identity is ordinarily taken to matter, if anything does, is R.

Ultimately, Parfit claims that R probably cannot matter in exactly the way

---

[15] *RP* p. 220. Quasi-memory, quasi-intention, etc. are introduced to handle the charge that analysing personal identity in terms of psychological relations is circular. E.g. Butler: "And one should really think it Self-evident, that consciousness of personal Identity presupposes, and therefore cannot constitute, personal Identity; any more than Knowledge, in any other Case, can constitute Truth, which it presupposes" ("Of Personal Identity" p. 305–306). Galen Strawson argues that arguments like Butler's, at least insofar as they are meant as objections to Locke's original psychological account of personal identity, rest on a misunderstanding of Locke's view and of his "forensic" notion of personhood ("The Secrets of All Hearts…").

[16] *RP* p. 206

[17] *RP* pp. 283-287

that we take heavyweight further-fact identity to matter. As I mentioned above, he even entertains the "Extreme Claim" that R does not matter at all. But he remains agnostic between the Extreme Claim and the Moderate Claim that R simply matters *less* or *differently* than further fact identity might.[18] I will be arguing that the Extreme Claim is true, because R does not, and could not, matter in at all the right way.

The Any-Cause Claim and the No-Cause Claim

What does Parfit mean when he says that what matters is Relation R "with any cause?"[19] It is unlikely that he means "even with no cause," because the relations of psychological connectedness and continuity seem to be necessarily causal; on the most natural reading, two causally unrelated psychological duplicates would not be R-related to one another. But if, as Parfit claims, the *nature* of the causal connection doesn't matter, it is natural to at least wonder whether the *existence* of a causal connection matters. I will call Parfit's view the *any-cause* view and the alternative view on which the existence of a causal connection is unimportant the *no-cause* view. To see whether Parfit can or should adopt the no-cause view, we need to get clear on what the any-cause claim amounts to.

We have seen that two R-related people must, at a minimum, be causally

[18] *RP* p. 312
[19] *RP* p. 283

related. But is any causal connection really as good as any other? Suppose that there exists a Putnam-style Twin Earth somewhere in our own universe, just like ours down to the last detail.[20] Now suppose that I catch sight of my (psychologically identical) counterpart with a high-powered telescope. There now exists a causal connection between my duplicate and me, where before there was none. I doubt that Parfit would want to say that, though we were not R-related before, we become R-related when I see my duplicate through my telescope. It would be a stretch to suggest that the two of us are somehow more psychologically "connected" or "continuous" than we were before.

So it seems that the R relation, and thus the any-cause claim, must require more than the bare existence of a causal connection between two psychologically similar people. Rather than just saying that two R-related people must be causally related, we should probably say that at least some of the causal links they share must be causes *of R*. But if we leave it there, we risk neutering the any-cause claim, because it might be that only very special sorts of causal connections can be causes of R.[21]

Fortunately, Parfit himself seems to hold a view on which R's causal requirements are not stronger than the "connectedness" and "continuity"

---

[20] As in, e.g., "The Meaning of 'Meaning'."

[21] E.g. consider the view that two people do not count as "psychologically connected" unless they are biologically continuous with one another in the ordinary way. A proponent of this view might still claim, as I believe that Parfit must, that what matters is R-with-any-cause-so-long-as-it-is-a-cause-*of-R*. But this theorist's "any-cause" claim would be very weak..

language strictly requires. He takes it for granted that R obtains in many unusual cases (e.g. teletransportation) and then argues as a further substantive point that the fact that R obtains via an unusual cause does not matter.[22] And at no point that I'm aware of does he ever explicitly endorse a view on which one type of cause is markedly better than another.

I will not try to give an exact account of what the R relation demands beyond the bare existence of a causal connection. One option might be to say that two people are R-related just in case their causal connection *explains* their psychological similarity. Another might be to say that they are R-related just in case they share an *information-preserving* causal link. Or an *intentionality-preserving* causal link. There is probably room for disagreement in some cases about whether two people are R-related.[23] But what's right out is requiring something like ordinary bodily continuity or sameness of matter or immaterial soul. There cannot be any restrictions of this severe sort on the sort of causal connection that R requires. Any plausible causal restrictions on R should be mild enough for the arguments of the next few sections to go through.

The any-cause claim states that R is what matters in the right way, if

[22] *RP* pp. 283–287

[23] E.g. consider a case where (in a deterministic universe) a scientist designs and create two people, one of whom comes into existence with the apparent memories of experiences that the scientist knows the other will eventually have. They are causally related, and their causal connection explains the similarities and (if it's an appropriate use of the term) connections between their psychologies. Are they R-related? I'm not sure we have to decide.

anything does, no matter how it was caused. Since there are no *severe* restrictions on what sorts of causal connection R requires, it is natural to wonder whether the existence of a causal connection is important. That is, it is natural to at least entertain the no-cause claim.

Is the No-Cause View Worth Considering?

The R relation, on the most natural reading, requires there to be some causal link between the two R-related people. Parfit does not, to my knowledge, consider cases in which people come to be psychologically similar through sheer chance, such that they *would* be R-related if their similarity were explained by their causal connection; he limits himself to cases of strangely-caused R-relatedness. Nothing in his view, so far as we have seen, expressly commits him to any particular response to these no-cause cases. We can now ask: What *should* he say about them?

This question may seem trivial, because it is impossible for two R-related people to be causally unrelated. Someone might reply: If R alone is what gives people reasons to care about their future and past selves in a different way than they care about other people, and R cannot obtain between two causally unconnected people, then cases of accidental psychological similarity are missing the grounds for partial concern; accidental duplicates are relevantly like complete strangers and relevantly unlike R-related persons. This response is

unconvincing. Unless there is independent reason to think that the existence of a causal connection matters, we can tentatively treat the presupposition of causal connectedness as an unimportant feature of R. If we do, we can replace R with a new relation R*:

> Relation R*: A cluster of psychological relations just like R, except that any relation in R that presupposes a causal connection is replaced in R* by the most analogous relation that does not presuppose a causal connection. Instead of psychological *connectedness*, we would speak in terms of psychological *similarity*, and so on.[24]

I mean for R* to be defined in such a way that two people are R*-related just when, and to the degree that, they *would* be R-related if they shared the right sort of causal connection. (In fact, if you'd like, you can take this as an alternative definition of R*.) R* obtains in every case in which R obtains, but it also obtains in cases of accidental "connection." I am now both R- and R*-related to the person who woke up in my bed this morning. I would be R*-related, but not R-related, to a person in some distant galaxy if they were psychologically just like the person who woke up in my bed.

The move from R to R* is in one way similar to Parfit's move from

[24] Note that if (contra my reading of Parfit) nothing in R presupposes a causal connection then R and R* are identical.

memory and intention to quasi-memory and quasi-intention. There, while it might be true that memory presupposes personal identity, the presupposition is taken to be unimportant, and so quasi-memory is an acceptable substitute. Here, while it may be true that R presupposes a causal connection, we have not yet seen any reason that the causal component is important to R, and so R* is tentatively an acceptable substitute.

I want to signpost a few other worries about R* and the no-cause view before I go about defending them. First, even if we think that causal connection is not necessary for what matters about R-relatedness, we might be bothered by the apparent fact that it would still be astronomically unlikely for two causally unrelated people to be R*-related by sheer chance. Second, we might think that R* could not matter in the way that R could. The existence of a causal connection, or the type of causal connection, might turn out to matter. If, for instance, agency matters and requires a certain type of causal connection, or if mental content matters and requires a certain type of causal connection, then R* cannot matter in the way that R (with the right kind of cause) might matter.

These are not trivial concerns, but I would note that many of Parfit's own cases—teletransportation, fission, duplication, etc.—are similarly improbable and might also count as the wrong sort of cause for (e.g.) content-preservation or for continuity of agency. And so rather than preemptively heading off objections here, I will first go over his arguments for the any-cause claim, which I will argue

actually bolster the no-cause view. I will then give my own arguments that the no-cause view is true, that R* matters as much as R, and that, since R* does not matter, R does not matter either. I will not argue directly that agency, content, etc. do not matter, but I will give cases in which it would seem that, if agency and content are preserved by R but not R*, that fact does not matter.

Parfit's Argument for his Any-Cause Claim

The argument for the any-cause view considers a few salient alternatives and judges the any-cause view to be the best among them.[25] By the time Parfit makes this argument, he has already given his arguments that some form of reductionism is true.[26] He considers four alternatives for what could matter in the way that personal identity is ordinarily taken to matter: "(1) Physical continuity, (2) Relation R with its normal cause, (3) R with any reliable cause, (4) R with any cause."[27]

By "physical continuity" in (1) I take Parfit to mean something like bodily continuity, because even in teletransportation cases there is always a sort of physical substrate.[28] He claims that this sort of physical continuity doesn't matter. We do not view the prospect of receiving a liver transplant as being as

[25] The argument I'm talking about is on *RP* pp. 282–287.
[26] Using, e.g., the fission and combined spectrum cases I discuss above. Cf. *RP* pp. 253–266, 236–243 or "Personal Identity" pp. 4–14.
[27] *RP* p. 283
[28] I owe this point to Galen Strawson.

survival-threatening as the prospect of receiving a brain transplant. The reason that we value keeping the same brain and not keeping the same liver is that the brain holds our personality, memories, and so on. Cases in which the body is preserved but psychological connectedness and continuity are lost seem as bad as death.

(2) fares little better, because the "normal cause" of R is the same sort of physical continuity that we have already seen does not matter, at least on its own. (What do matter are "the various relations between ourselves and others, whom and what we love, our ambitions, achievements, commitments, emotions, memories, and several other psychological features.")[29]

There are some reasons to prefer keeping our bodies, just as one might prefer to keep her original wedding ring and not a teletransported copy. There might also be strong reasons to prefer bodies similar to our present ones if we prefer bodies with our own primary and secondary sex traits, if we're very attractive, if we are athletes, etc.; but we might just as often prefer new bodies, and here what matters is that our bodies be qualitatively similar, not that they be the very same material bodies. It's true that we ordinarily prefer to keep our bodies intact, but upon reflection we should think that this preference makes sense only insofar as keeping our bodies intact is ordinarily a pretty good way of ensuring the continuance of R. We might decide not to call the results of some

[29] *RP* p. 284

R-preserving processes, like teletransportation, "survival." But if the only reason we don't call the results survival is that R has been preserved through some abnormal cause, we should not think that it matters much that they are not survival.

The argument against (3) works by way of analogy. Suppose that there is a treatment for some disease that is not very reliable. Of course, when selecting a treatment, we should prefer a reliable one. But after the fact, if the unreliable treatment happened to work, "only the effect matters."[30]

Having rejected all of the other alternatives, Parfit adopts the remaining option: What matters is R with any cause.

The Any-Cause Argument Supports the No-Cause Interpretation

Again: Parfit's strategy in arguing for his any-cause claim is to consider a list of alternatives and, having rejected all but one of them, to settle on the remaining one. Of course, this style of argument will fail if the initial set of alternatives is incomplete. Someone might object that we have only learned that the bodily continuity view, the R-with-its-normal-cause view, and the R-with-a-reliable-cause view are all false; we have not learned that the R-with-any-cause view is true.

This line of objection is interesting only if we can identify some plausible

[30] *RP* p. 287

alternative to the any-cause view that Parfit's arguments do not tell against. Even though he argues explicitly against only three alternatives, he seems to do so by way of arguing *for* (or at least pumping intuitions in favor of) two positive claims, which I will call the Turning-Out Claim and the R Claim.

> The Turning-Out Claim: What matters in the right way in personal identity problem cases (and in everyday life) is how things turn out, not how they came to pass.[31]

> The R Claim: What matters in the right way about how things turn out, if anything does, are facts that depend almost exclusively on R, and not on other things like physical continuity or a further fact.[32]

I myself find these two claims persuasive. I do not see what besides R (or R*) could matter in the absence of a deep further fact, and I am persuaded that there *is* no deep further fact. I find that even the concept of real, hardboiled, further-fact identity slips through my fingers whenever I try to get hold of it.

Even if I could get some sort of hold on what the further fact is supposed to be or depend on—perhaps something like a Cartesian soul—I'm not sure that

[31] Again, see the discussion of unreliable causes on *RP* p. 287.
[32] Once again, the list of what matters from *RP* p. 284: "[T]he various relations between ourselves and others, whom and what we love, our ambitions, achievements, commitments, emotions, memories, and several other psychological features."

it would matter. Certainly it would not matter without R (or R*); the prospect of being "reincarnated" with no memory of my current life has never been the least bit comforting. At best, the further fact could only ever matter in the presence of R (or R*) or enable R (or R*) to matter. But it's not clear what the further fact could contribute even in these cases. I am convinced, for example, by Locke's objection to the same-soul criterion of personal identity, which suggests that if souls could exchange "consciousnesses," personal identity would follow "consciousness" rather than soul.[33] And so I doubt that a further fact, even if it were clearly conceivable and not empirically suspicious, could provide what we want it to.

When I think about whether anything could plausibly matter in the way that personal identity is taken to, merely physical facts do not seem up to the task,[34] while psychological facts are more promising. I'd much rather undergo a liver transplant than a brain transplant, and I am completely untroubled by the fact that the atoms in my body are constantly being replaced. I would remain unconcerned if this process were vastly accelerated, so that from minute to

[33] *An Essay Concerning Human Understanding,* p. 338: "But yet to return to the Question before us, it must be allowed, That if the same consciousness (which, as has been shewn, is quite a different thing from the same numerical Figure or Motion in Body) can be transferr'd from one thinking Substance to another, it will be possible, that two thinking Substances may make but one Person. For the same consciousness being preserv'd, whether in the same or different Substances, the personal Identity is preserv'd." I take Locke's "consciousness" to be something like a narrower version of Parfit's R.

[34] Of course, psychological facts might be physical facts. But they are not *merely* physical facts in the way I am using the term.

minute my body would be made of completely new matter.[35]

While I admit that nonstandard cases of R-preservation like the teletransportation case make me feel uneasy, I cannot justify my unease; all of the differences between the nonstandard and the standard cases appear trivial on closer inspection. Imagining dancing at a crowded party also makes me uneasy. Unless I can come up with some explanation of how it could matter that the teletransporter constructs a new body all at once instead of over time, or that it uses radio waves and chemical vats rather than the usual biological processes, I am forced to reject my unease about teletransportation as irrational, just as I reject my unease about dancing. (We can ignore the fact that I am so bad at dancing that it might actually be dangerous, since that's not the source of my fear.)

I can imagine that it might matter that someone psychologically connected to me steps out of the teletransporter on Mars after my body disappears on Earth. I cannot see how it could possibly matter *how* that comes to pass in the absence of a further identity fact whose metaphysical underpinnings I fear the teletransportation process will disrupt. (And again, I'm not sure I can even see how a further fact would matter if there were such a thing.)

I also think that the Turning-Out Claim and the R claim give as much or

[35] This case was suggested to me by Galen Strawson. I would, of course, become concerned if all the bodily processes associated with taking on and jettisoning matter were also accelerated. But that's not the point.

more support to the no-cause view as they give to the any-cause view. The claim that it is the way things turn out that matters straightforwardly supports the no-cause view. The Turning-Out claim is different from the weaker claim that the sort of cause does not matter. In fact, it explains *why* the sort of cause doesn't matter. In other cases, the sort of cause might not matter even though it is not just the way things turn out that matters. For instance, if I care about whether I am responsible for some event (in a broad sense of "responsible") I might only care *whether* an intentional action of mine caused that event to occur, and not *how* my action caused the event to occur. But in personal identity cases, the sort of cause doesn't matter *because* it is the way things turn out that matters. And if it's the way things turn out that matters, why should we care about whether there was a causal link at all? If it is the way things turn out that matters, we should adopt the no-cause view instead of the any-cause view.

The second claim, that it is the R relation that matters about the way things turn out, may seem to favor the any-cause view because R seems to require a causal connection. But the R claim only favors the any-cause view over the no-cause view if we have independent reason to believe that the existence of some causal connection is part of what allows R to matter in the right way. If the existence of a causal connection is an unimportant feature of R, we should replace the R claim with the analogous claim that it is R*, if anything, that matters about the way things turn out.

The Branch Line

Allow me a quick digression before I get to my further arguments for the no-cause claim:

My favorite Parfitian thought experiment has always been the branch line teletransportation case.[36] In an ordinary teletransportation case, I imagine entering a machine that will scan my body, destroying it in the process, and then send the information gathered to another far-away machine that will reconstruct an exact duplicate out of new matter. If R is what matters, then teletransportation is not to be feared, and would be rational to use when ordinary transportation is expensive, dangerous, or otherwise troublesome. In the branch line teletransportation case, the first machine does not destroy my body immediately, but it does enough damage that after a few days the original body will die painlessly. A copy is constructed out of new matter as usual, but the original body is stuck on a "branch line."

Being stuck on the branch line strikes most people I have discussed the case with as an unfortunate fate, even if these people believe that "ordinary" teletransportation is, or is about as good as, survival. But how could it be *worse* for me that my body be destroyed on Thursday rather than Tuesday? In ordinary cases, we would probably take those extra days of life to be a good thing.[37]

---

[36] *RP* pp. 199–201, 287–289.

[37] It would of course be bad if these extra two days are consumed by a painful fear of impending death. But what's at issue is whether this fear is *warranted*. If this isn't convincing, just suppose that it's an unworried Socrates type on the branch line.

The branch line thought experiment is my favorite of Parfit's thought experiments not because its conclusions are the most convincing—in fact, they are probably the hardest to accept—but because it shows most clearly how radical the consequences of reductionism are. I see the branch line case as the last step in an argumentative process that begins with the rejection of the substantial further fact.

The process goes something like this: Cases like fission and the combined spectrum strongly suggest that there is no such thing as a separately existing temporally extended self or a deep and important further fact about personal identity. If there is no deep further fact, then we have to turn to other facts to ground the reasons and values that seemed to depend on the existence of temporally extended persons. The only available facts that could plausibly ground these reasons and values are R facts. If it's only the R facts that matter, then teletransportation is as good as ordinary survival. The branch line case is not significantly worse than teletransportation, since delaying the destruction of the original body could not be a horrible bad. Since it is not significantly worse than ordinary teletransportation, the branch line case is *also* about as good as ordinary survival. But it seems clear, in a way that it may not be in ordinary teletransportation cases, that the person on the branch line *dies*.

Considering the branch line case will not convince anyone who is not already persuaded by a broadly Parfitian reductionist view that such a view is

true. But it should convince anyone who does hold such a view of how serious its normative implications are. If death can be as good as survival because of something that happened on the moon four days ago, then the way that many people conceive the badness of death is mistaken.

I want to pick up where the branch line case leaves off. My cases are designed to show, first, that R* is just as good as R. The second thing they're designed to show is that R* doesn't matter in anything like the right way. Put those two claims together, combine with reductionism, and you get Extreme Claim Reductionism.

Further Argument for the No-Cause View

I'll call my first case

Evacuation: I am a settler on a planet near a distant star. One day, I wake up to distressing news: The star is expected to go supernova at any minute. There are too few ships for a complete evacuation, and settlers are asked to use private or municipal teletransporters. (All of these teletransporters work in the usual way; the original bodies are destroyed in the scanning process and a digital signal is transmitted to another unit which constructs a duplicate out of new matter.) Unfortunately, the increased stellar radiation will interfere with the teletransporters' signals,

and the information that makes it to the reconstruction chambers will be unavoidably incomplete.

We are told not to worry: Modern teletransporters have advanced gap-filling algorithms for just this sort of contingency. We might, we are told, wake up with some changes to our bodies, memories and personalities, but no matter how bad the signal gets we will be sure to wake up as fully functional human beings. This worries me; I have grown used to thinking of teletransportation as just another way of getting around, but I've always been confident that I wasn't risking radical changes to my psychology. Nevertheless, with no other options, I queue up for the municipal teletransporter and cross my fingers that I'll have good luck and that the person who steps out of the pod on a distant planet will not be radically different from me.

Suppose we accept, with Parfit, that "ordinary" teletransportation is about as good as ordinary travel. Then the best case is one in which my signal makes it through the stellar radiation undistorted. In this case, my duplicate will be maximally R-related to me. But what if the signal *is* lossy? In this case, I will probably hope that the gap-filling algorithms fill in the missing bits as close to the original signal as possible. In the limiting case, the gap-filling algorithms will (against all odds) get everything exactly right.

Is this limiting case as good as the case without any signal loss? If R is better than R*, then it may not be. Parts of my duplicate's psychology will be merely R*- rather than R-related to the relevant parts of my psychology, because there will be no causal link—certainly nothing like an explanatory or information-preserving causal link, to be sure—between them.

I believe that the case in which the gap-filling algorithms happen to get everything exactly right is not worse than the "best case" in which the signal is not lossy. I can think of no reason that this could be true except that R* is as good as R. My next thought experiment is meant to get you to believe that it *is* true; the lucky gap-filling case is as good as ordinary teletransportation:

False Alarm: I wake up in a refugee camp on another planet. I seem to remember my whole life, including the moments before entering the teletransporter, but I know this is no guarantee that my signal got through unaltered; the gap-filling algorithms are there to make lossy teletransportation as non-traumatic as possible. I ask one of the counselors if my signal had any gaps that had to be filled, and I am told that it did. I put it out of my mind; my planet is about to be destroyed, which I care about much more.

The supernova fails to materialize. After a few weeks, the anomalous stellar activity dies down, and scientists announce that it was not actually

evidence of an impending supernova. Since the planet was not destroyed, the databases containing everyone's teletransportation data are undamaged. The post-teletransportation settlers have the option of surgical intervention to restore them to their pre-teletransportation selves, with or without their memories of the strange ordeal intact, as they prefer.

Settlers who opt to keep their new personalities are asked to start new lives elsewhere, and new duplicates of their old selves are created on the planet's surface. Other settlers take the surgery and fly back home. Some enter teletransporters but ask that their original data be used for reconstruction. A few, cutting out the middleman, simply kill themselves and leave notes asking to be "revived" on the planet with their original teletransportation data. The end result is that the planet ends up repopulated just as it was before the panic.

My case is unique. Though my signal was lossy, it turns out that the gap-filling algorithms—through an astounding coincidence—reconstructed it exactly as it originally was. When I exited the teletransporter, I was an exact duplicate of the person who entered the teletransporter. But not—or at least not exclusively—*because* of the way he was when he entered it.

Having heard about my case, a surgeon approaches me. She offers to perform the reconstructive surgery that other settlers are getting. She will

remove parts of my brain and body and replace them with *exactly identical* pieces of brain and body. But these new pieces, I am assured, will be copied from the original teletransportation data. After the surgery, I will be a non-accidental duplicate instead of a merely accidental one.

Since I am already a duplicate of the original, albeit an accidental one, surgery costs will not be covered by the government. I go for a walk and consider whether I ought to pay her to perform the surgery.

I want to make two claims about this case. (1) It would be irrational to pay for the surgery. Getting parts of my brain and body replaced with exact duplicates would not make me a more apt successor to the person who entered the teletransporter. The surgery would be successful only at making me poorer. (2) If it would be irrational to pay for the surgery, that could only be because the existence of an R-preserving causal connection does not, in itself, matter. If we were persuaded by the any-cause view, we should transition to the no-cause view, because R* is just as good as R with any cause.

(1) seems almost too obvious to argue for. Imagine our two data files, content-wise identical, the first from the original scan and the second the filled-in one from which the duplicate was constructed. Imagine that the surgeon is sent a printout of each. Two stacks of paper with the same code printed out on them. On any account, performing the surgery based on the second stack of paper is, at

best, as good as not performing the surgery. If the any-cause claim were true, and not the no-cause claim, then performing the surgery based on the first stack would be better. But the thought that having the surgeon read from one (identical) printout rather than the other while performing the surgery would make any meaningful difference strikes me as completely unbelievable and bizarre. It would be equally unbelievable and bizarre even if the *entire signal* had been reconstructed at random.

(2), the claim that the only reason paying for the surgery would be irrational is that R* is as good as R, requires more argument. It might plausibly be objected that the surgery is not better for the *person who undergoes it* because (a) we do not have the same sort of interest in our pasts that we do in our futures or because (b) the surgery disrupts ordinary bodily continuity in a way that matters. Both objections can be warded off by tinkering with the original case.

Suppose that before I enter the teletransporter an oracle descends from on high and tells me about how my signal will be lossy but how it will, by coincidence, be reconstructed exactly as it was. If the oracle were then to offer me the chance to pre-pay to force my duplicate to undergo the surgery, accepting the offer would be just as irrational for me as getting the surgery would be for him, even if I am behaving completely egoistically. Since my concern in this case is for the future, objection (a) fails. Objection (b) fails because I am already resigned to teletransportation, which is total bodily discontinuity. Any further bodily

discontinuity could not plausibly matter more, and so the good of causal connection, if there were any, could be expected to win the day.

If the any-cause view were true, and the no-cause view false, then it would be better, at least from my pre-teletransportation perspective, that my post-teletransportation duplicate get the surgery. Since it is not better, the any-cause view is false, and we should adopt the no-cause view instead.

Some readers might believe my diagnosis of the Evacuation and False Alarm cases but be unmoved by my conclusion because the prospect of accidental duplication is so unlikely. But, in a thought experiment, mere unlikeliness should not bother us. It certainly would not bother Parfit. He is explicit that even impossible cases can make for instructive thought experiments. This is true even of "deeply impossible" cases, but doubly so for "merely technically impossible" cases.[38] If the impossible is kosher, as it always has been in the personal identity literature, then so is the improbable.

If my arguments from this section succeed, then R* is as good as R with any cause. It remains to show that R* doesn't matter.

R* Does Not Matter

On the surface, the claim that R* does not matter seems easy enough to believe. I have no normatively important relation to someone on the other side of

[38] *RP* p. 219

the universe who lived billions of years ago (or will live billions of years in the future) merely because our atoms are arranged in more or less the same way. It would be a mistake to regret things in his past, and if I were punished for them, it would be undeserved. It would also be a mistake to anticipate things in his future. (I mean by "mistake" not that it would be imprudent but that it would rest on a confusion.) If I could somehow make his or his friends' lives better, I would have no more reason to do that than I have to make anyone else's lives go better. If I have a goal, his accomplishing it does not matter to me, except insofar as it may mean that a good thing has happened somewhere. If I put this view forward in a vacuum, I would expect it to be relatively uncontroversial.

I have already argued that R, by itself, does not matter more than R*. And I have endorsed and tried to unpack Parfit's arguments that R, by itself, matters as much as anything can matter in the ways that personal identity is taken to matter. If these claims are right, and R* does not matter, it follows that I have no special normative relation to (what I would ordinarily call) my own past and future and that attitudes of regret, anticipation, etc. rest on a mistake. This claim is harder to accept than the claim that it is a mistake to anticipate the future or regret the past of an accidental duplicate.

If the earlier steps of my argument succeed, the only remaining way to block the Extreme Claim is to argue that R* without R *can sometimes* matter. This is a view worth considering. After all, wouldn't it be some comfort to be told by

the oracle in Evacuation that, though my signal will not get through undamaged, an exact duplicate will be constructed in the other teletransporter pod as usual?

False Alarm has the curious feature that my accidental duplicate *knows* that he is my duplicate (and, if I am visited by the oracle, I know that I will have an accidental duplicate). In this way it is more like an "ordinary" teletransportation case than an "ordinary" accidental duplicate case.

I do not believe that the mere knowledge that I have (or had or will have) a duplicate can make the right sort of difference. Consider:

> Great Big World: Scientists announce that the universe turns out to be much larger than we ever expected. It extends so many light years and aeons in every direction that the numbers involved aren't concisely representable even with tools like Conway's chained arrow notation.[39] It looks more or less the same all the way through, filled with galaxies and stars and planets, and so just through sheer probability we can all expect to have a great number of causally unrelated duplicates spread across time and space, many of whom will be living lives very much like our own.

[39] The chained arrow notation is defined in *The Book of Numbers*. Using Conway's recursively defined notation, we can express numbers that are far too large to encode in ordinary decimal or exponential notation even using all of the matter in the universe by writing down a few numerals and arrows.

I might be happy or unhappy to learn that the universe is this large, but it would be a mistake to be happy or unhappy *for myself* if I hadn't had any money riding on the question.[40]

Of course, knowing that the universe is very large would only justify a belief that I have duplicates somewhere or other. Knowing that I have some *particular* accidental duplicate somewhere or other might give me the psychological machinery necessary to care about him, but it could not give me a *reason* to care about him in a first-personal way. Learning in a deterministic universe that things were so arranged at the beginning of time that I'd have a particular duplicate at spacetime coordinates <t,x,y,z> would not give me any more reason for additional first-personal concern than I would have in the Great Big World. Neither would spotting a duplicate with a sophisticated telescope.

---

[40] It may be worth considering the somewhat parallel case of "quantum immortality." David Lewis believes that if a "no-collapse" / "many worlds" interpretation of quantum mechanics is true then we should expect to live forever. The argument (very roughly) is that there is nothing that it is like to be dead, which means that we cannot properly *expect* such an outcome, but we *can* expect to live out one of the futures in which we survive due to some quantum fluke ("How Many Lives…" especially pp. 16–19). Lewis finds this possibility frightening, because in the overwhelming majority of the cases in which a person's life is saved by a quantum fluke they will be left in very bad shape. (E.g. some fraction of a bullet quantum tunnels past my brain but the rest makes contact.)

From what I can tell, few people who are familiar with the argument share Lewis' fears. I assume that this is not because they are all convinced that the no-collapse view is false. It is more likely that they believe that the jump from "there's nothing that it is like to be dead" to "expect to live forever, so long as there is no collapse" is unwarranted. I agree with this diagnosis. Lewis' expectation seems unwarranted in the same way that it would be unwarranted to expect to live (close enough to) forever upon learning that we live in a Great Big World. In fact, I believe that all ordinary first-personal expectation for the future is fundamentally mistaken in this same way.

(There would be some causal connection between me and my duplicate here, but, as I have argued, it wouldn't be the right kind.)

In Evacuation, if I am visited by an oracle who tells me how things will turn out, then it is not just that I know that I will have a particular duplicate in a particular place at a particular time. I know that he will continue my story; he'll be friends with my friends (or their teletransported duplicates), continue with my work, etc. These are the sorts of thing that Parfit repeatedly suggests matter about R, and it seems that in some special cases R* does as good a job of preserving them as R does. But not always—my causally unrelated duplicates in the Great Big World will *not* continue my story in this same way. Perhaps, then, we can claim that my accidental duplicate in False Alarm does have special and important relations to me even though my duplicates in Great Big World do not.

I do not believe that this reply can work. We can modify a response that Parfit gives on behalf of the Extreme Claim against the objection that we can rationally have special concern for our own futures and pasts for the same sorts of reason that we can rationally have special concern for our loved ones.

He asks: "Why should I care about what will happen later to those people whom I love? The reason cannot be [...] 'Because my loved ones now care about what will happen to them later'. This is no answer, because our problem is also to know why *they* should care about what will happen to them later."[41]

[41] *RP* pp. 310–312. Parfit attributes this argument to Broome.

We can ask: Why should I care that my accidental duplicate will live among people who are causally linked to their past selves? The reason cannot be "because they now care more about those R-related selves to which they are causally linked than any accidental R*-related duplicates they might have." This is no answer, because our problem is also to know why *they* should care more about causally connected "selves."

I have claimed that it would be irrational, before or after the False Alarm, to pay for a surgery that would replace any amount of my duplicate's brain and body with identical parts built from data with a "better" causal history. This would be equally true for all of my acquaintances or their duplicates. It would be an implausible sort of bootstrapping to accept these claims but to argue that nevertheless the fact that my accidental duplicate in False Alarm can step into *these very same* friends' lives gives me a normatively important relationship with him that I do not have with my accidental duplicates in Great Big World.

If these arguments succeed, then we have no special, normatively important relationship to our own pasts and futures that we do not have to any coincidentally R*-related people. By the same token we also have no special, normatively important relationship to our friends' pasts and futures that we do not have to people who are coincidentally R*-related to them. But we have no special relationships with such people. Extreme Claim Reductionism is true.

Normative Consequences of Extreme Claim Reductionism

I want to highlight four areas where ECR undermines our ordinary normative beliefs. (The list is meant to be neither non-overlapping nor exhaustive.)[42]

First, we should reject reasons of partial concern, either for our own pasts and futures or for others'. For example, it would unwarranted for me to care more (or even in a special way) about my own impending torture than about a stranger's, except insofar as it might now be possible for me to prepare for or try to avoid my own torture in a way that I couldn't with a stranger's. This is probably the most-discussed ethical upshot of reductionism and, in a moderate form, it informs many of Parfit's substantial conclusions. If I learn that tomorrow someone will be in great pain, I have no reason to hope that it will not be me and no additional reason, if I learn that it will, to prevent it. This rejection of partial concern follows more or less straightforwardly from the arguments from the past few pages. It may be a welcome conclusion, because someone who could be motivated exclusively by impartial reasons might act selflessly and live without many of the worries that plague the rest of us.

---

[42] Compare Marya Schechtman's "four basic features of personal experience—survival, moral responsibility, self-interested concern, and compensation" (*The Constitution of Selves* p. 2). As I read Schechtman, her "four features" are a proper subset of my "four areas." I've defined the Extreme Claim in terms of the reasons and values that personal identity is *ordinarily taken* to ground, and I take the fact that Schechtman, I, and others have come to similar conclusions about the (supposed) normative importance of personal identity to be good evidence that there are such reasons and values and that my "four areas" are among them.

Second, and relatedly, we should reject everything that depends on any sort of lasting responsibility for actions. Desert, obligation, etc. are at best instrumentally important. Attitudes of resentment and gratitude are unwarranted. If I learn that yesterday someone committed a terrible crime, I have no reason to hope that it was not me and no additional reason, if I learn that it was, to help the victim.

This second conclusion is less welcome than the rejection of partial concern, but I think that it is equally unavoidable. Suppose I now have an exact physical and psychological duplicate whose body has a different history. (This body has been involved in some crimes that his has not, let's say.) As I have argued, undergoing a surgery to replace any amount of my brain or body with exact duplicate parts could not matter, even if the duplicate parts are built from data with a different causal history. This means that neither I nor my duplicate could take on or lose responsibility for some action by getting the surgery. But in the limiting case where my *entire* brain and body are replaced, the surgery would effectively be the same as killing me and replacing me with a teletransported copy of my duplicate, which could not be importantly different from ordinary survival for him *or* for me. It follows that either we are both responsible for each other's actions or we are both not responsible in any morally weighty sense for any past actions. In a Great Big World, I would have many duplicates with many different past lives. I do not think it is plausible that I would be responsible for

all of their actions; discovering that I live in a large universe would not make that sort of moral difference. It is much more plausible that nobody bears responsibility in any deep or morally important sense for any past actions.

Third on the chopping block are uniquely first-personal attitudes like anticipation and reminiscence. This (by my reckoning) is a less-discussed consequence of reductionism, but it is not ignored. Parfit mentions briefly the possibility that anticipation "might be justified only by the non-existent deep further fact."[43] Wachsberg grounds the irrationality of special concern at least partly in the incoherence of anticipation.[44] More recently, Stokes takes the unique phenomenology and apparent respectability of first-personal attitudes like anticipation as evidence that reductionism is not the threat to ordinary morality that people like me take it to be.[45]

I am not as sure as Wachsberg or Stokes that the rationality or intelligibility of attitudes like anticipation is as inextricably tied to the moral facts about impartialism, responsibility, etc. as all that. It seems to me that rational creatures with quite different psychologies, who do not anticipate and remember in the first-personal, metaphysically loaded way that we seem to, might still recognize and be subject to all of the same moral reasons that we are.

Whether or not rejecting anticipation etc. has immediate ethical

---

[43] *RP* p. 312
[44] In *Personal Identity, the Nature of Persons, and Ethical Theory*, especially ch. 2.
[45] In "Will it be me? Identity, concern and perspective."

consequences, it does have a few practical and psychological upshots. Or at least it has for me. When I worry about some event in my future, my worries often take the form "x is going to happen to me." On any reductionist view, this sort of worry can be rephrased as "x is going to happen to someone who bears such-and-such relations to my present self." But I am convinced that none of these relations matter. I can still imagine my future experiences from the inside, and this might still frighten me. But I can imagine experiencing things that will happen to other people just as well. In both cases, my imagining it doesn't make it real, and my fear, to the extent that it takes a first-personal form, is unwarranted.

When I apply this line of thinking to something that is worrying me, I am often able to worry less. In my own case, it is not so much that I find the fearful prospect less fearful as it is that I can more easily turn my attention elsewhere. If I am flying through some turbulence or riding with a speeding driver—two things that have historically terrified me—I am now often able to close my eyes and think: "If there is a crash, that will be unpleasant for the people involved, but it will have nothing at all to do with what's happening now, which is that I am experiencing the pleasant sounds and vibrations of a huge machine propelling itself along at breakneck speeds."

It is often remarked that there is no use in worrying about something if you can't do anything to change it. I have always tried to follow this advice. I

find that I have an easier time following it now that I believe that, besides being useless, many of my worries are fundamentally unwarranted and mistaken.

One might worry that it may be psychologically impossible to stop anticipating and reminiscing in a metaphysically-loaded first-personal way. Or that even if it were possible it would likely lead to a sort of life-destroying myopia. I find both conclusions plausible, but only a little troubling.

The mere fact that ordinary anticipation depends on mistaken assumptions does not mean that we should try to stop anticipating our futures, or that we could succeed if we did. It might be psychologically impossible to feel a certain sort of love for a person or to root for a sports team without on some level believing that they are especially deserving, or to navigate the physical world without tacitly buying into some illusory concepts along the lines of substance and extension, or to enjoy a work of fiction without thinking of the characters as real in some way. Only a person who felt an uncontrollable need not only to be right but to be right at all times about all things in all ways would try to cut these activities from their life because they bring along error.

However: If I start becoming too invested in the Rockets' season, or if I find myself rationalizing too many flaws in a friend, or if I begin sobbing uncontrollably at the death of a character in a TV show, or if I start taking a non-scientific metaphysics too seriously, I can pause, take a step back, and remind myself of the truth. The Rockets don't deserve the championship more

than anyone else just because they are from a nearby city; my caring for someone does not immediately transform them into a better person; the characters in fiction are not real people; the physical world operates in strange ways at the fundamental level.

The truth of Extreme Claim Reductionism is like these truths. If I find myself worrying about the future or fixating on the past in what I judge to be an unhealthy or unproductive way, I can remind myself that my future and past selves bear no normatively important relationship to me and see if that helps.

Admittedly, anticipation and first-personal memory are probably more fundamental to the human experience than rooting for the Rockets is, and so accepting ECR even on an abstract intellectual level may throw a wrench into some of our everyday thought processes. Ever since I started giving the practical implications of ECR serious thought, I've found myself almost obsessively picking apart my thought processes to see if they involve mistaken assumptions about selves or about what matters. It is a strange and self-alienating endeavor.

In the normative sphere, this sort of self-alienation has a unique and disturbing character. But, as I will argue later, some degree of moral self-alienation is unavoidable in a sufficiently reflective life even if one does not accept the Extreme Claim or reductionism about personal identity, and so it has to be grappled with for anyone who takes the time to honestly, scrupulously, and reflectively examine her moral thought. This alienation, and the questions

surrounding it, are the central topics of the remainder of this dissertation.

The fourth and final implication of ECR that I will note here is that the grounds for a lot of what we take to be valuable disappear. Goods like earned accomplishment, lasting friendship, atonement, a positive life trajectory, etc., if they are good for anyone at all, are good for temporally extended persons. The Extreme Claim tells us that temporally extended persons, if they can be said to exist in the first place, are not well suited to be basic units of moral analysis. Something being good for a person is about as plausible as something being good for the composite entity comprising me from ages 8–12, the world's largest octopus, and three minutes of Lebron James's evening on June 12, 2011. The notion that a life well lived could be good beyond the sum of the local goodness of its constituent parts is imperiled.

It's no good to appeal to organic unities or holism about value or to argue that something like a lasting friendship could just be *a good* apart from being good *for* anyone. Nothing that I've said implies that facts about the metaphysics of persons make it strictly *impossible* for things like lasting friendship to matter. A lasting friendship is, of course, a beautiful story, which might make it valuable in some way. Beautiful stories can be (and usually are) told about more than one object. It's not obviously impossible for there to be a good that depends on more than one object; we would ordinarily think that there are many such goods. The problem isn't that persons, because they aren't unified wholes, could obviously

never ground any goods. The problem is that persons *aren't suited to ground the special kinds of goods we think they do*.

Here we can appeal to Parfit's comparison of persons to nations.[46] There's nothing I've said about the metaphysical structure of values and reasons that implies that nothing could ever be good for a nation, or that there could never be a good that depends on a nation for its existence. But goods like lasting friendship are not among them. If we thought that nations were unified wholes, as many believe about people, we might think that something like a lasting friendship between nations could be an important intrinsic good. If we were then to become reductionists about nations, it would be rational to believe not just that lasting friendship among nations is less important than we thought, but to believe that it is not important at all. It could not matter in anything like the same way.

It might be impossible, and would almost certainly be psychologically unhealthy, to stop caring about lasting friendship, atonement, accomplishment, etc. But this does not mean that these things are important—just that we perhaps ought from a practical standpoint to allow ourselves to go on caring about them in our everyday lives.

[46] Eg. in *RP* p. 211, 240.

Conclusion

This chapter is titled "Why People Don't Matter." It would be more accurate, but less snappy, to say that people don't matter *qua person*. Insofar as people have a capacity for great depth and intensity of experience, they matter a great deal. Pain is just as bad on my view as it is on any other. (Worse, perhaps, because no person could ever be properly compensated for her pain.) And there is room in the view for some non-hedonistic values like aesthetic appreciation, knowledge, capacities, etc., if you find any of those compelling. (I don't, but not, or not entirely, because of my views about personal identity.)

But whatever precise reasons and values do manage to survive Extreme Claim Reductionism, it is clear that ECR substantially shrinks the normative realm from what we might ordinarily imagine. We start with a peach and end up with a pit. ECR is a difficult, alienating truth to believe, let alone to internalize and act on, to the point that we might wonder whether and how we should internalize or act on it at all. It is to these questions that I now turn my attention.

Chapter Two – The Truth Will Set You Against Yourself

In Chapter One, I argued that both reductionism about personal identity and Parfit's Extreme Claim are true.[1] These claims together imply that the moral universe is very different from what it ordinarily appears to be. When I reflect on their truth and implications, I experience feelings of loss, unreality, disconnection, and tension. In short, I feel alienated—from myself and from the world around me. I suspect that for normal, reflective human beings, belief in the truth of Extreme Claim Reductionism might always be alienating. But the fact that the truth is alienating does not imply that it is false or even that it cannot be believed. Perhaps, to the extent that alienation is a bad thing, it *shouldn't* be believed; but it is difficult to sincerely reject a belief that the epistemic reasons favor even if the practical reasons favor dropping it.

My goal in this chapter is to show how and why Extreme Claim Reductionism (ECR) is alienating, and why this alienation is something that we shouldn't expect to be able to easily ignore, sidestep, or reason away. If I am successful, what I say might matter even to people who do not believe ECR; as I will argue in Chapter Three, the sort of alienation that belief in ECR engenders

[1] Recall: The "Extreme Claim" language comes from *Reasons and Persons* (pp. 308–312). I have in mind the strongest available version of the claim: Nothing else matters in anything like the way that we ordinarily take the deep further fact of identity to matter, and if there is no such deep further fact, there is nothing left to ground the value of a long life, reasons of self-interest, etc.

threatens to show up, if less frequently, for people with *any* plausible (and sufficiently realist) set of beliefs about what matters. It will be the work of that chapter to outline a practical strategy for living with alienation.

Reasons Alienation

What do I mean when I say that believing in reductionism together with the Extreme Claim causes alienation? As a first pass, I intend "alienation" in its broadest possible sense—that thing that Railton calls "a kind of estrangement, distancing, or separateness (not necessarily consciously attended to) resulting in some loss (not necessarily consciously noted)."[2] I can imagine a belief in ECR causing everything from interpersonal or familial alienation ("Grandma, please don't take this the wrong way, but you're not a person in any deep sense") to moral alienation ("I know that he killed my best friend and ruined my life, but apparently nobody can deserve to suffer") to even a quasi-Marxist alienation from the fruits of one's labor ("All I can really say is that I'm psychologically and biologically related to whoever built this cabinet over the past month, and I guess that doesn't matter").

These and other sorts of alienation will be in the background over the course of the present chapter, but the sort of alienation that I intend to treat most directly is what I'll call

[2] "Alienation, Consequentialism, and the Demands of Morality" p. 134

Reasons Alienation: The sort of alienation that occurs when what I believe I do, should, or even must care about is different from what I believe actually matters, or when the intensity of my care about something is wildly out of proportion with my beliefs about how much it matters.[3]

It might not be immediately clear that Reasons Alienation as I describe it is possible. It implies a gap between one's beliefs about what matters and what one cares about, and it might seem strange to think that care and belief can come apart in this way. My first task, then, is to defend the claim that Reasons Alienation is not impossible. My first line of defense will be a simple argument from cases: I'll describe situations in which it seems clear that a belief in ECR would engender a gap between what a person does, can, and should care about and what they actually believe is important. If these cases are possible, then Reasons Alienation is possible (and, moreover, can be caused by a belief in ECR).

Extreme Claim Reductionism and Reasons Alienation

ECR is the combination of two theses. The first is reductionism about personal identity. I take reductionism to be the claim that there is no "deep

[3] There may, of course, be ways of mattering that are not properly cashed out in terms of reasons, as valuable objects or states might if a Scanlonian buck-passing analysis of value is wrong (*What We Owe to Each Other* pp. 95–98). But "Mattering-Wise Alienation" doesn't have the same ring as "Reasons Alienation."

further fact" about personal identity, but only lower-level facts about things like psychological connectedness, biological continuity, etc. The second thesis is the Extreme Claim, which says that none of these lower-level facts ground *any* of the special normative reasons that we ordinarily take the supposed deep further fact to ground. When we combine the two theses, we get a view on which, because there is in fact no deep further fact, there are in fact none of these special reasons.

In Chapter One, I discussed four sorts of ordinary judgment or attitude that ECR implies are often or always unwarranted—(1) attitudes of partial concern and judgments of partial importance, (2) judgments about lasting responsibility, (3) attitudes like anticipation and reminiscence, and (4) judgments about values grounded in extended periods of people's lives. Each sort of case is a possible source of alienation, as the following four examples are meant to show:[4]

> (1) Arthur loves his boyfriend Andrew very much, and values Andrew's well-being over the well-being of others. But Arthur also believes that ECR is true, and thus believes that his partial concern and care for Andrew are unwarranted. Nevertheless, Arthur *does* care more about Andrew than about other people, and knows that this special care is part and parcel of his love and is, moreover, necessary for the health and

[4] Feel free to skip ahead once you feel like you get the picture.

happiness of the relationship. Even if he could get rid of his partial care, he would not; nevertheless, he is troubled by the knowledge that it is unwarranted.

(2) Bette, who has been poor for all of her life, comes across a huge windfall of money. She uses the money to buy gifts for the people who helped her when she needed it and, in a few cases, to exact a measure of justice on people who did her terrible wrongs. Paying everyone back in this way gives her a deep sense of satisfaction that she wouldn't give up for the world. This satisfaction strikes her as irrational, however, when she reflects on the knowledge that the people she "repaid" are merely psychologically and biologically connected to the people who did right or wrong by her in the past and that such connections cannot ground any sort of desert. The further knowledge that by repaying her debts she contributes to a culture that incentivizes good behavior is not enough to justify her sense of satisfaction. She is not satisfied because of her contribution to a shared expectation of reciprocity; she is satisfied because she was finally able to follow the dictates of her sense of justice, misguided as she ultimately believes it to be.

(3) Cleo has a well-paying job that she hates. She gets through the day by

thinking about how, in less than a year's time, she will have saved up enough money to quit her job and move to a better city. Occasionally she remembers that all of the fun experiences in her new home will be had by someone who is merely psychologically related to her, and not by "her future self" in any deep sense. In order to avoid sinking into a deep depression, she does her best to put such thoughts out of her mind and go on anticipating a better future in the usual way.

(4) Del's life has involved overcoming a series of difficult and painful challenges. They assure themselves that each each challenge is, was, or will be a meaningful chapter in a full life. But reflection on ECR reminds them that their pain will be in no way absorbed or counteracted by any eventual relief or accomplishment, and that it plays no role in grounding some higher good. Their life is, simply, painful and dreary, with a few bright spots between struggles. Del finds this outlook grim and comfortless, and does their best to put it out of their mind.

Arthur, Bette, Cleo, and Del all care, and moreover do well by caring, about things that they know do not actually warrant their care. In each case, I expect that their knowledge of their situations will prove disconcerting, that it will undermine their images of themselves as practically rational, as unified

agents, as genuine and intellectually honest people. In short, their knowledge will be alienating. Unfortunately, there is no obvious way I can see for Arthur et al to transcend the source of their alienation and live in good faith. Their alienation is a natural result of their knowledge of the gap between what they care about and what actually matters; and this gap, for them as well as for us, is one that is impossible, or at least extremely difficult and probably inadvisable, to bridge.

The Possibility of Reasons Alienation: Akrasia

If the above cases are successful, they show that Reasons Alienation is possible. But they do not show *how* it is possible; there remains the problem of explaining why it's not straightforwardly *im*possible given the apparently close link between caring and valuation.

I've defined Reasons Alienation as resulting from the gap between what we believe we do, should, or even must care about on the one hand and what we believe actually matters on the other. "Care about" here is open-ended, and intentionally so. In one perfectly good sense, believing that something matters is a way of caring about it. If that were *all* that caring about something consisted in, or the *only* way of caring about something, then reasons alienation as I describe it would be impossible. It would also be impossible if, as a matter of fact, judgments of importance always had to line up precisely with every sort of care.

Accepting one of these claims according to which Reasons Alienation is impossible would amount to adopting an extreme sort of what is generally called judgment internalism about motivation. I myself am inclined towards externalism about motivation, according which it is possible to judge something to matter without being at all motivated by this judgment. But even if one were to accept a more moderate internalist view, on which normative judgments might entail *some* degree of motivation without fully determining all facts about motivation and care, there is plenty of space for Reasons Alienation.[5]

I believe that we should be no more than moderately internalist about motivation; an extreme internalism on which normative judgments completely fix the spaces of motivation, desire, and care seems to entail that weakness of the will is impossible, since weakness of will involves being motivated to do something other than what one judges to be best.[6]

It is true that philosophers of as high a degree of eminence as you like have held the view that weakness of will really *is* impossible; but remembering

---

[5] For example, a belief in ECR might cause me to be less inclined to promote my own future welfare at the expense of the future welfare of others, but so long as I retain *some* disproportionate self-interest, the door is open for Reasons Alienation.

[6] It might be argued that a Davidsonian analysis of weakness of will, according to which the akratic actor really does judge the thing they do to be "unconditionally" better and is simply failing to keep all of their reasons in mind, could be made compatible with extreme internalism, though I am not sure how convincingly ("How is Weakness of the Will Possible," especially pp. 38–42). In any case, Bratman's case of Sam, the hard-drinking akratic depressive who sees no good reason whatsoever to keep drinking and avoid sleep but does so anyway, seems to me to show decisively that Davidson's analysis cannot account for every apparent case of weakness of will ("Practical Reasoning and Weakness of the Will" pp. 156–157).

some of the other patently false views that eminent philosophers have held—Kant's claim that one should not lie to the murderer at the door about his victim's hiding place comes to mind—should be enough to disabuse anyone of the belief that every view held by a great philosopher ought to be taken seriously. There may, of course, be a problem of *how* weakness of the will is possible,[7] just as there may be a problem of how consciousness is possible; but *that* weakness of the will, like consciousness, is possible (because it is actual) seems so certain that I am unsure how anyone with any experience of the world could sincerely doubt it. Anyone who has ever wanted to stop running during the last lap of an important race, to keep drinking even though the party is dying down, to order a dessert even though it is expensive and unhealthy and will be gone in a minute, or to keep watching TV instead of finishing a dissertation ought to know that weakness of will is real.

If weakness of will implies that extreme judgment internalism about motivation is false, then the door is open at least to the possibility of Reasons Alienation. Actually, we can say more: Weakness of will, insofar as it can be alienating and involves a disconnect between normative judgments and motivation, *is* a sort of Reasons Alienation, although not one that I will focus on after this; the cases I'm most interested in involve more than a temporary conflict

[7] My guess is that Plato may already have had it close to right with his appeal to parts of the soul.

between occurrent motivation and sober judgment. In the next sections, I turn my attention to the question of whether such deeper conflicts are also possible.

The Possibility of Reasons Alienation:

The Wrong Kind of Reasons Objection – an Introduction

The discussion in the next section will be unavoidably technical, and so to make it easier for the reader to avoid getting bogged down in the details, I want to take some time here to sketch briefly the conclusions with which I hope to emerge.

Consider again the case of Arthur and Andrew. In one sense, Arthur has reason to value Andrew only as much as he does other people, because as a matter of fact Andrew *is not more valuable* than most other people. But in another perhaps less direct way Arthur has all the reason in the world to value Andrew more than other people; it is only by so doing that he can maintain the health and happiness of their relationship.

I will be defending a few claims about this sort of case. First, whatever these two sorts of reason end up amounting to, they are both perfectly good; neither is an illusion, and both make legitimate practical demands on Arthur. Second, and relatedly, whatever we might do to make sense of the moral metaphysics surrounding Arthur's case, the demands made on him are *practically* irreconcilable. He cannot privilege one sort of reason and put the other out of his

mind (or, for that matter, arrive at some satisfying split weighting of the two).

The Possibility of Reasons Alienation:

The Wrong Kind of Reasons Objection

In my description of Reasons Alienation I talk about "what I believe I do, should, or even must care about" coming apart from what I judge to be important. And while "do, should, or even must" is meant to be read as a disjunction, the really interesting cases are the ones where someone knowingly does *and* should or must care in a way that is unwarranted (at least in the usual way) by the actual reasons at play.

One might be happy to allow that weakness of will is possible but doubt that there could be cases where we *should* care about things that do not matter. It is natural to think that that mattering involves—or just *is*—warranting care. Nothing besides the fact that something actually matters, we might think, could give us reason to care about it. It is true enough that by caring more for his boyfriend Andrew than for other people Arthur is able to keep his relationship healthy and happy, but this fact does not bear on Andrew's value. As such, we might wonder whether the utility of Arthur's disproportionate care for Andrew can bear on whether he *should* care disproportionately. Practical utility seems to provide a different sort of reason for having a feeling than does the fact that the

feeling fits[8] or is appropriate to its object in the usual way, and we might wonder whether it in fact provides a reason of the *wrong* kind, and thus no reason at all. Likewise, we might wonder whether it is fittingness, and not utility, that gives Arthur the wrong kind of reason to care. After all, the pain of a toxic relationship is immediate, unavoidable, and intense, whereas perhaps "fittingness" is something that should only worry philosophers.

If we are convinced that only either the fittingness of care on the one hand or the utility of care on the other bears on how Arthur should feel about Andrew, then the alienation that Arthur feels over the conflicting pulls of fittingness and utility is ultimately irrational. Call this the wrong kind of reasons objection to the possibility of reasons alienation.

Several versions of the so-called wrong kind of reasons problem have received a flurry of attention in the past decade or two.[9] Much of the contemporary debate surrounding cases like Arthur's involves the special problem they present for the "fitting attitude" or "buck-passing"[10] accounts of

---

[8] By "fit" or "fittingness" I intend what I think is a commonsense notion with broad application. Anger fits injustice, admiration fits excellence, amusement fits humor, caution fits danger, doubt fits unreliable testimony, motivation fits an opportunity to do some good, and so on. This usage of the terms is meant to be in line with their usage in the relevant contemporary debates.

[9] Key papers and chapters that lay the groundwork for the contemporary debate include D'arms and Jacobson' "Sentiment and Value," Rabinowicz and Rønnow-Rasmussen's "The Strike of the Demon," Heironymi's "The Wrong Kind of Reasons," Olson's "Buck-Passing and the Wrong Kind of Reasons," Raz's "Reasons, practical and Adaptive," and others.

[10] The "buck-passing" language originates with Scanlon (*What We Owe to Each Other* p. 97).

value advocated by Scanlon and others.[11] As we will see, the wrong kind of reasons problem has implications beyond reductionist theories of value, and my focus in this section will be on a few of these implications. But first, it's worth getting the problem on the table in the context of the buck-passing debate.

Buck-passing views, in general, reduce a thing's "value" to its possession of lower-level features that give us reasons to respond to it in certain ways. Taking Scanlon's view as a paradigmatic example, a thing's being valuable consists not in its possession of anything like an irreducible Moorean property of goodness but rather in its having certain other features that give us reason to take certain positive attitudes towards it, e.g. admiration or respect.[12] Value is analyzed *in terms of* reasons, effectively shrinking the realm of the normative by making value theory a subset of the theory of reasons as opposed to a separate field with separate metaphysical commitments. Different buck-passing accounts differ in subtle ways in terms of the precise sort of reasons-to-respond that they reduce value to, and so for simplicity, I'll use the verb "value" as an umbrella term to cover the having of any of a number positive orientations towards an object. Thus I will say that something is valuable on a buck-passing account in case we have reason (of the right kind) to value it.

Let's return to Arthur and Andrew. If Arthur takes exceptionally positive attitudes towards Andrew, and is especially disposed to promote Andrew's

[11] Rabinowicz and Rønnow-Rasmussen and Olson, for example, take this approach.
[12] *What We Owe to Each Other* pp. 95–98

welfare, their relationship is more likely to be happy and healthy. Thus, he has certain indirect practical reasons to take these positive attitudes, to promote Andrew's welfare, etc. But none of these reasons seem like they make Andrew valuable![13] They are reasons of the *wrong kind*.

You don't have to be a buck-passer—that is, you don't have to believe that Andrew's value *consists in* his having lower-level properties that give Arthur reason to value him—to get the feeling that there is a deep difference between the reasons of the "wrong" and "right" kinds in Arthur's case. Whether or not some buck-passing theory is true—that is, whether or not things' value *consists in* the existence of reasons to value them—we can still probably say more cautiously that something is valuable *just in case* it is warranted or fitting to value it. The special sorts of reason that Arthur has for valuing Andrew do not seem to be connected up value in this way.

In the context of the contemporary buck-passing debate, to say that a reason for valuing something is of the "wrong kind" means that it's of the wrong kind to account for that thing's value.[14] But, as we've seen, the right / wrong kind of reasons distinction seems to be (or to suggest) an important distinction outside of the confines of the buck-passing debate. It would be natural at this

---

[13] The non-ECR theorist can, of course, hold that Andrew *is* valuable—just not more than anyone else and not for these reasons. He is valuable because of things like his capacity to feel pleasure and pain, his agency, etc.—not because valuing him helps Arthur be a better boyfriend.

[14] Of course, for the buck-passer, this distinction as stated is circular—hence the challenge posed by the WKR problem.

point to wonder: Can (something like) the right / wrong kind of reasons distinction apply to things besides reasons to value?

The answer to this question seems to be yes, although once we widen our scope it becomes much harder to give rules to classify reasons as being of the right or wrong kind. In fact, there seem to be all sorts "wrong kindish" reasons at play for all sorts of things throughout the history of philosophy. Pascal's Wager purports to give us a reason for belief in God that seems wrong-kindish. Kavka's Toxin Puzzle[15] and Newcomb's Problem[16] suggest wrong-kindish reasons to intend. Railton's Kantian demon case suggests a wrong-kindish reason for

[15] In "The Toxin Puzzle," Kavka imagines an eccentric billionaire who offers you a million dollars if at midnight you intend to drink a non-lethal toxin the next day, at which point the money will already be in your account, so that you will not have to actually drink the toxin to keep it. Kavka believes that in this case you would have reason to intend to drink the poison but not reason to drink it, which is meant to introduce a wedge between reasons to intend and reasons to act.

[16] As presented by Nozick in "Newcomb's Problem and Two Principles of Choice," the problem imagines perfectly-accurate action-predicting computer that will put one million dollars into a box if and only if it predicts that you will not open a second box that will definitely have one thousand dollars in it. The practical question is whether you should open both boxes or instead forego opening the thousand dollar box, given that the million dollars will either be there or not by the time you get the chance.

Though they might appear identical in structure, there are subtle differences between Newcomb's Problem and the Toxin puzzle. You would get your money in Kavka's case if you changed your mind at the last minute so long as you had earnestly intended to drink the poison at midnight. On the other hand, the computer in Newcomb's problem would presumably predict any changes of heart, so that it is what you *actually do* as opposed to what you *intend* that determines (or at least perfectly lines up with) how much money you walk out with. Thus the Toxin Puzzle arguably makes clearer or more pressing Kavka's supposed difference between reasons to act and reasons to intend, whereas Newcomb's Problem arguably makes clearer the depth of the tension between the reasons in play.

having a certain sort of moral outlook or character.[17] You might be able to imagine others.

To pick the simplest and best-known of these cases, Pascal famously argues[18] that we ought to believe in God not because there is overwhelming evidence that They exist but because it is decisively prudential to do so; since we must rationally assign some positive probability to God's existence, and since the reward for belief in the case that They exist is infinite while the reward for unbelief is at most finite, belief in God promises the greatest expected payoff. The reason that Pascal gives us to believe in God, if it is a reason at all, seems "wrong-kindish" in very much the same way that Arthur's reasons for disproportionate care do. What *precisely* these cases all have in common has, again, proven to be hard to nail down, but it is easy to get the sense that there must be a general phenomenon to be accounted for.

I say "wrong-kindish" in the preceding paragraphs because, once we broaden our scope beyond the objection to buck-passing, it's not obvious that all of the reasons we were calling the "wrong kind" are, in fact, wrong for anything.

---

[17] Railton's case involves a powerful, all-knowing demon who threatens humanity with bad consequences to the extent that we stray from Kantian morality. He notes that consequentialists would "have reason to convert to Kantianism, perhaps even to make whatever provisions could be made to erase consequentialism from human memory and prevent any resurgence of it" ("Alienation, Consequentialism, and Morality" p. 155). They have these reasons, presumably, in spite of the fact that their epistemic reasons favor (or seem to them to favor) the truth of consequentialism. I return briefly to the demon case in Chapter Three.

[18] *Pensées* 680 (in *Pensées and Other Writings*, pp. 152–158)

They're certainly not all wrong in the distinctive buck-passing way.[19] As such, from here on out, I'll adopt Raz's more neutral language of "standard" (meaning right-kind) and "non-standard" (meaning wrong-kind) reasons,[20] though in so doing I don't mean to commit myself to understanding the distinction exactly as he does.

Now that the stage is set, I'll restate my purposes. I want to remain less-than-fully committed about what exactly the standard / non-standard reasons distinction amounts to, though I will describe one promising sort of strategy that, if successful, could explain why conflicts between standard and non-standard reasons seem so intractable. Note that even if this attempted explanation of the intractability of conflicts between standard and non-standard reasons fails, I hope to establish in the next section at least *that* they are intractable (or at least really really hard to tract).

On what I understand to be the most popular sort of view, and the one to which I myself am inclined, the essential difference between standard and non-standard reasons is that standard reasons are "object-given" and non-standard reasons are "state-given." Parfit,[21] Raz,[22] and others endorse

---

[19] Remember: In the buck-passing context, the "wrong kind" of reasons are of the wrong kind to account for a thing's value. This sort of wrongness isn't available in other contexts.
[20] "Reasons, Practical and Adaptive" p. 1
[21] *On What Matters* vol. 1 pp. 50–51, 420–432
[22] "Reasons, Practical and Adaptive"

versions of this view. We can see how the object / state view works by considering its application in a few different cases.

In the epistemic case, Pascal appeals not to the object of our belief, which is God's existence, but to the *state of believing* in God. He does not need to say anything about evidence for God's existence to run his argument that the state of belief is a good one to be in, and likewise, he is happy to admit that his argument gives no evidence for the existence of God.[23]

In the toxin puzzle and Newcomb's problem, imagined players are in a position to be rewarded if they genuinely intend to take actions that, when the time comes and their rewards are already secured or lost, they would be better off not performing. In both cases cases, the objects of their intention—drinking a non-lethal poison or leaving a box of money unopened, respectively—have nothing to recommend them. But the state of intending to achieve these ends is a very good place to be.

Turning back to Arthur and Andrew, it's much the same. The object of Arthur's affection, Andrew, has nothing in particular to make him an appropriate object of care and partial concern over anyone else. But Arthur's

[23] Indeed, he believes that the possibility of a rational proof of God's existence is explicitly denied in scripture: "Who will then blame the Christians for being able to provide a rational basis for their belief, they who profess a religion for which they cannot provide a rational basis? They declare that it is a folly, *stultitiam* (1 Cor. 1: 18) in laying it before the world: and then you complain that they do not prove it! If they did prove it, they would not be keeping their word" (*Pensées and Other Writings* p. 153).

being in the state of caring for him is one of the surest sources of happiness in his life (and for that matter, in Andrew's).

The key insight of the object / state account is that standard and non-standard reasons do not simply attach themselves to a single object and favor potentially conflicting responses. Cases of that sort are familiar in everyday life and (comparatively) simple to deal with. Suppose I make a promise that would be harmful to keep; on commonsense morality I have reasons of fidelity to keep it and consequentialist reasons not to. I reflect for a moment and decide one way or the other. Genuine moral dilemmas aside (if such things are, as I doubt, possible) we can usually reach a relatively satisfying weighing of the reasons in these cases and move on with our lives. Cases of straightforwardly competing reasons are certainly common causes of inner turmoil, but except in extreme cases they don't seem to generate the sort of persistent, undermining alienation that I'm considering here.[24] Because standard and non-standard reasons attach themselves to objects of different sorts, they can, unlike merely competing reasons, favor not just different responses but fundamentally different *sorts* of response. This fact, on my view, is what makes standard and non-standard reasons practically irreconcilable.

To get clear on what I mean by "fundamentally different sorts of response," we can turn to Parfit's discussion of state-given reasons in *On What*

[24] Of course, if ordinary conflicting reasons did regularly and unavoidably generate Reasons Alienation, that would help my case rather than hurting it.

Matters.[25] Parfit considers a series of cases in which a despot threatens us with a horrible outcome unless we come to have some unjustified belief or unwarranted desire. He argues that it is an illusion that the threat of the horrible outcome bears on our beliefs and desires in at all the way that ordinary reasons for belief and desire do:

> Return now to the claim that, in such cases, we would be responding to our reasons to *have* these beneficial beliefs. We ought, I have suggested, to reject this claim. If we were *causing* ourselves to have these beliefs, this process might be rational, and involve responses to reasons. We would be responding to reasons for *acting*, which would be provided by the facts that would make it good if we had these beliefs.[26]

In other words, our non-standard / state-given reasons for belief are not reasons for belief at all! They are reasons for action—the action of making ourselves adopt certain beliefs if the means to do so are available to us. The point seems to me essentially right, and one that Pascal may have appreciated when he suggested long-term ceremonial religious practice as the road to belief for anyone convinced by his argument.

---

[25] Vol. 1, pp. 420-432
[26] *On What Matters* vol. 1 p. 422

We can easily extend Parfit's analysis to Arthur's case. Arthur, we can say, has no reason whatsoever to care especially about Andrew, and indeed, caring especially about Andrew is irrational. On the other hand, Arthur has every reason to *cause* or *allow* himself to care especially about Andrew, and no reason not to.[27] If Arthur's case works analogously to the epistemic case in this way, as I believe it does, then in Parfit's terminology we can say that his reasons compete but do not conflict. They compete because Arthur "could not successfully respond to both," but do not conflict because they do not "support different answers to the same question."[28]

Hoping that it might help to clarify Parfit's distinction between competing and conflicting reasons, I spent some time trying to think of cases of competing but non-conflicting reasons that weren't among these perhaps confusing standard vs. non-standard reasons cases. I found that all I could come up with were cases where the reasons in question were had by different agents or by the same agent at different times. For example, my friend and I both have reason to win (or, if you like, try to win) a race. Our reasons compete, because we cannot both act on them successfully, but they do not conflict, because his reasons bear

[27] Arthur might have reason not to cause or allow himself to care especially about Andrew if there were something about having appropriate levels of care that recommended itself as an end, e.g. if having appropriate levels of care turned out to be a moral duty for some reason. But any such reasons would be quite different from the standard reason Arthur would have if Andrew were especially good or valuable, and do not affect my argument here.

[28] *On What Matters* vol. 1 p. 425

on his actions and my reasons on my actions. Similarly, I might now, so far as I can anticipate the future, have most reason to ensure some future outcome that, once the time draws nearer, I realize I must ensure does not occur. Again, my reasons compete, but they do not conflict; my old reasons bear on old actions and my new actions on my current ones.

I mention these multi-agent and diachronic cases because I think that they help to illustrate precisely what it is about competing but non-conflicting reasons in the synchronic single-agent case that alienates: Competing but non-conflicting reasons like Arthur's undermine our practical unity. They set us against ourselves.

As Christine Korsgaard wonderfully puts it, "when you deliberate, it is as if there were something over and above all of your desires, something which is *you*, and which *chooses* which desire to act on."[29] With the quibble that I might prefer to think of deliberation as adjudicating between reasons rather than desires, I think she perfectly describes the ideal case of deliberate action. In the case of competing reasons, I think that most of us are able to achieve this ideal most of the time. Even when we make difficult decisions or ones that we will regret, they strike us as wholly *ours*. Korsgaard is concerned with agential unity in particular, but her characterization of ideal deliberation applies equally well to other cases in which we respond to reasons. In the best sort of case, we can step

[29] *The Sources of Normativity* p. 100

back, survey all of the relevant reasons, and arrive at a single response (or set of responses) that we can endorse wholeheartedly as appropriate and correct.

I do not believe that most of us can achieve the Korsgaardian ideal when it comes to competing but non-conflicting reasons. When Arthur faces up to his decisive reasons not to care especially for Andrew on the one hand but to try to cause or allow himself to care especially for Andrew on the other, he is like the two competitors trying to win the race. As a rational actor, he is bound to try to cause or allow irrational cares if he can. As a rational valuer, he is bound not to have these cares. It seems to me difficult if not impossible for the valuer and the actor—likewise the epistemic agent, the intention-former, and any of Arthur's other "parts," if I can be allowed to speak so loosely, to which reasons can individually speak—to reach any sort of satisfying accord. Arthur cannot feel any one way about Andrew and, stepping back, endorse that feeling as being wholly *his*. Arthur is alienated.

The Rationality of Reasons Alienation:

The Commensurability Objection

I've just argued that the best available account of the standard / non-standard reasons distinction can explain how ECR engenders Reasons Alienation. Though I do believe that this picture is essentially correct, it does not need to be one hundred percent true for it to support the intractability of

conflicts like Arthur's. All that's required is for the demands on Arthur to be genuinely practically irreconcilable. Some philosophers have denied (or seemed to deny) this sort of claim, at least in the epistemic case. Chisholm,[30] for example, holds that ethical demands on belief always trump epistemic ones, implying that ethical demands on belief are commensurable and reconcilable with epistemic demands on belief.[31] In this section I'll be making two claims: First, there is good reason to believe that Chisholm is wrong; moral demands are not even in principle commensurable with epistemic demands (likewise the sorts of axiological demands that bind Arthur). Second, and most importantly, even if Chisholm is right, moral demands are at least *practically* irreconcilable with epistemic (likewise axiological) demands.

To support his commensurability claim, Chisholm gives an analysis of epistemic requirements on which they turn out to be a special proper subset of doxastic requirements, i.e. requirements to have some belief, to withhold judgment, and so on. An epistemic requirement, for Chisholm, is just a doxastic requirement imposed by some fact or state of affairs p such that p imposes no *non*-doxastic requirements beyond such requirements as are required by *any* fact

[30] In "Firth and the Ethics of Belief."

[31] As I will use them, "(ir)reconcilable" and "(in)commensurable" are related but distinct terms. Roughly, reasons are (in)commensurable when they can(not) be weighed against one another whereas reasons are (ir)reconcilable when they can(not) be satisfyingly responded to or accounted for together. Genuine moral dilemmas, if there are such things, might thus be thought of as cases of commensurable but irreconcilable reasons.

or state of affairs.[32] On this view, epistemic requirements are analogous to requirements of etiquette or aesthetics, and just like such requirements they are trumped absolutely by ethical requirements. He claims: "The distinguishing feature of *ethical* duty is not to be found in the the considerations that impose that duty. Rather, an ethical duty is simply a requirement that is not overridden by any *other* requirement."[33]

Against Chisholm's view we can say that it makes the difference between standard and non-standard reasons for belief—which pre-theoretically seems like a very real and very deep one—look shallow and ad hoc. This objection is particularly strong in the face of Parfit's view that epistemic reasons are properly reasons for belief while non-standard reasons "for belief" are in fact reasons for the *actions* of trying to cause or allow oneself to have certain beliefs. On Parfit's view, the difference between standard and non-standard reasons for belief seems neither shallow nor ad hoc, which is good reason to prefer it to Chisholm's view.

---

[32] "Firth…" pp. 123–124. Two bookkeeping notes on this retelling of Chisholm's view: First, Chisholm does not actually specify what sort of thing p is, but I think that something like facts or state of affairs are what he has in mind. If not, and he's actually thinking about beliefs or propositions or something else, nothing much is lost. Second, this is actually Chisholm's account of a "purely doxastic requirement." But when he goes on to define "epistemic requirements" as a proper subset of these, he seems to do so in such a way that he repeats his earlier definition, such that epistemic requirements just are purely doxastic requirements. Colleagues with whom I've talked on the subject have likewise failed to see the difference between the two definitions, and in his "Why There Are No Epistemic Duties," Chase Wrenn notes that Chisholm seems to have made the error I'm claiming here (pp. 117 and 133).

[33] "Firth…" p. 127

Still, though some have denied it,[34] there is something attractive about Chisholm's "ethical ought." It's an all-things-considered just-plain-ought, a top level ought, a trump card. Wouldn't it be nice if such things existed? When we are faced with Pascalian wager or a Newcombian box or an Arthurian relationship, shouldn't there be at least one rational way to respond? Can even the most virtuous possible person really be condemned to some form of irrationality when faced with competing standard and non-standard reasons?

I think that the incommensurabilist can offer a partly satisfying response to this sort of worry. We could (though would not be not bound to) say that in these cases there *is* at least one right way to *act*—it's just that the way you're required to act might be to cause yourself to be irrational in your beliefs, cares, etc. We might even be able to say that so acting may be the *best* available response, where "response" is meant to include not only your actions but your beliefs, cares, etc. Since it seems to me generally far, far less important that I believe or care rationally than that I act rightly, it would generally be much better for me to act rightly in making myself believe or care irrationally than to act wrongly in allowing myself to believe or care rationally. Still, even in this best

[34] E.g. Feldman: "I take Hall and Johnson to be suggesting that when you epistemically ought to gather more evidence and you morally ought to do something else, the moral ought "wins" and you just plain ought to do that other thing. It's this that I just don't understand. Of course, by this I mean to suggest that no one else understands it either. It makes no sense" ("The Ethics of Belief" p. 692).

case, my belief or care is not made rational by the rightness of my actions, and so, insofar as I can observe my irrationality, the seeds of alienation are sown.

A conversation with my colleague Jonathan Drake made me aware of a second, more difficult objection to the sort of incommensurabilist view that I favor. (I don't actually remember which one of us came up with the objection or if he ultimately endorsed it.) I am inclined to say, with Parfit, that reasons for belief are one thing and reasons for action, including the action of causing or allowing oneself to believe something if one can, are another. But perhaps it is just a contingent fact about human psychology that our faculties for action and belief formation seem so separate and different. If we were, say, perfect doxastic voluntarists, with just one faculty responsible for both action and belief formation, would we really be tempted to say that reasons for action and belief were incommensurable? (Would such a claim even make sense?) If not, maybe we shouldn't make the deep and perhaps necessary-if-true claim that reasons of different sorts are incommensurable on the basis of contingent facts about human psychology.

Perhaps we can respond that even though it is plausibly a merely contingent fact about human psychology that we apparently respond to reasons for action and reasons for belief in deeply different ways or using distinct faculties, nevertheless reasons to act, believe, and care are still necessarily deeply conceptually different from one another, or work metaphysically in totally

different ways, or are sensitive to wholly different sorts of considerations, and as a result are still deeply incommensurable with one another. On this account, the perfect doxastic voluntarist would simply be using one faculty to respond to two deeply different and incommensurable sorts of reason. I think that a response along these lines is likely to be right, but I am not certain.

Or perhaps it is enough to say in response that reasons for care, belief, action, etc. are incommensurable in principle *for creatures like us.* Suppose there exists some perfectly doxastically and affectively voluntarist alien that handles all reasons-response with a single mental faculty. Suppose too that the metaphysical structures of reasons for action, care, belief, etc. do not render these reasons deeply or necessarily incommensurable. Suppose that there are no such things as genuine dilemmas between reasons of different "sorts" and that in every case this alien can make a wholly rational all-things-considered best response. Well, so what? *We* can't do such a thing, because *we* respond to reasons of different sorts in deeply different ways, and rationality in one realm can't somehow wash clean irrationality in another.

If neither of these responses is satisfying, I do not know a better way to answer the objection, though that is not to say that it cannot be answered. Fortunately, my more modest claim of practical irreconcilability, which ought to be enough to show that Reasons Alienation is a predictable effect of a belief in ECR, doesn't depend on deep incommensurability.

Suppose that Chisholm is right; suppose that reasons for belief, for care, and for action, among others, are perfectly commensurable. Suppose too that there are no such things as a genuine dilemmas, as some people believe there are even between reasons of commensurable sorts. That is, suppose that it is always possible in every circumstance to simultaneously act, believe, and care in right, warranted, and fitting ways. Even then, I believe that knowing the truth of ECR must be alienating for creatures like us. Here's why:

However the moral metaphysics turn out, Arthur will always be able to think: "The man sitting next to me on this couch is merely psychologically and otherwise continuous with and connected to the person who was at the park with me yesterday, just as "me yesterday" was merely connected in these ways to me now; I can owe the man next to me nothing, nor can he owe me; every act of kindness and cruelty and romance and bitterness that he has done to me or I to him is in the past and has nothing to do with the two people here now except insofar as it may have helped shape us; it cannot matter for me now whether he will die tomorrow anymore than it could matter for me whether I will die tomorrow; the man I love, since I love all of him and not just the tiny piece of him in this room, is ultimately a conventional fiction."

Since Arthur believes ECR, he believes everything he's just thought. How could having these thoughts be anything but painful and alienating? He can, of course, assure himself that since reasons are commensurable and since his

reasons to care for Andrew outweigh his reasons not to, he's being perfectly rational in every way. I do not think that this thought will be much consolation. Whether or not he can ultimately endorse his attitudes, they will unavoidably strike him as immediately bizarre, absurd, unwarranted, fantastic.

What, then, does Arthur do? Or any of us?

Chapter Three – How to Live a Lie

In my first chapter, I argued for Extreme Claim Reductionism (ECR), the view that there is no "deep further fact" about personal identity and that nothing matters in anything like the way that we ordinarily take the "deep further fact" to matter. If this is true, then things like partial concern, desert, first-personal anticipation, and the value of lives taken as wholes are left without any plausible grounds or justification.

In my second chapter, I argued that a belief in ECR threatens us with what I called Reasons Alienation, which I defined as alienation resulting from a gap between what one believes one does, should, or must care about on the one hand and what one believes actually matters on the other. I then argued that, in the present case, our Reasons Alienation is the result of ECR undermining many of our standard, object-given reasons to value, act, and believe while leaving untouched many of our non-standard, state-given reasons to cause or allow ourselves to have certain motivations, values, or beliefs. If, as I argued, standard reasons are at least practically irreconcilable with non-standard reasons, we should expect Reasons Alienation to be a difficult-to-avoid fact of life for those of us convinced of ECR.

Thus we face a question: What do we do when our beliefs set us against ourselves in this way?

Consequentialism and Self-Defeat

This question, it turns out, is not entirely new in the philosophical literature. Charges of self-defeat and alienation have been made against consequentialist ethical theories for many years, and philosophers who take these charges seriously have attempted to give accounts of how people should go about trying to be consequentialists. The issues faced by ECR and the Reasons Alienation that it engenders are similar to the ones faced by consequentialism, and can, I think, be met with the same sorts of replies that consequentialists have historically used to defend their theories. In fact, because ECR strips away much of ordinary morality, the remaining core may end up looking much more consequentialist than whatever we started with.[1] (Parfit thought, at least when he wrote *Reasons and Persons*, that even Moderate Claim Reductionism supported act utilitarianism.) If this is so, as I believe it is, then the self-defeat charges that consequentialism faces will be *among* the self-defeat charges that ECR faces. And so, since the relevant literature on consequentialism is as rich as it is, I will in the next few sections consider the cases of consequentialism and ECR side by side.

[1] An argument by way of example: Suppose that, before coming to believe ECR, someone believed in the seven Rossian prima facie duties of fidelity, reparation, gratitude, justice, beneficence, self-improvement, and non-maleficence (*The Right and the Good* p. 21). The only of these duties that could plausibly survive ECR would seem to be beneficence, nonmaleficence, and perhaps a very limited form of justice. This now-former Rossian, whose ethics started out looking very much like commonsense morality, would now have something very close to a consequentialist view.

It is widely believed that some forms of consequentialism are bound to recommend that we do not strive to do the actions that the theories say are right. I share this belief. The following case is meant to motivate this seemingly paradoxical recommendation:

Suppose I get a $500 tax refund and am deciding whether to buy a nice new bike or give it to the JustMilk charity.[2] On any minimally plausible view, a world where fewer infants contract HIV or Malaria or suffer from dehydration is vastly impersonally better than a world where I experience fewer flat tires and chafed thighs. And so, on any minimally plausible impartialist consequentialist theory, I should give the $500 to JustMilk before spending it on a bike, barring any unforeseen effects of either action.[3]

So far so good; I give the money to charity and forego the bike.[4] Now suppose that I run through this sort of calculation whenever I make even minor decisions. I stop going out with friends, because my time could be better spent earning supplemental income to donate or volunteering at a homeless shelter; I

[2] Since it's my dissertation, I'm picking my favorite charitable organization, co-founded by my brother, which aims to manufacture silicone nipple shields with inserts that would be used to deliver nutrients and drugs to breastfeeding infants.

[3] In most of what follows, I assume a maximizing consequentialism, on which I may perform an action only if no other available option has a better outcome. I do not think that this assumption is necessary; I make it because the argument runs most straightforwardly in the maximizing case. Satisficing views make room for permissible actions that do not maximize consequences so long as the outcome is "good enough," but the gap between buying the bike and giving to JustMilk is so massive that it is hard to see how any satisficing view that allows buying the bike could ever prohibit much of anything.

[4] I'm giving myself too much credit here, but this is after all a thought experiment.

stop drinking socially—alcohol is expensive, after all—and eating anything besides brown rice and beans; I take my girlfriend to Taco Bell for our anniversary dinner; I take on a second and then a third job so that I can donate the proceeds; and so on. Before long, my friends have begun to ignore me, my girlfriend has left me, I've lost my jobs, I'm a nervous wreck, my health is failing, and—worst of all, consequentially speaking—I am giving *less* time and money to worthy causes than I was before my consequentialist awakening.

Well, so what? Haven't I just described a case where I'm not *really* performing actions that are all things considered for the best? Aren't I simply failing to perform the consequentialist calculus properly? The answer to the first question is surely "yes," but the second one is trickier. In some cases it seems likely that I have simply ignored some of the predictable psychological and social tolls of my actions, as when I take my girlfriend to Taco Bell for our anniversary, which results in my being dumped and suffering through a period of crushing depression and non-productivity. This sort of case may be enough by itself to recommend that I stop trying to perform the consequentialist calculus, because it might be that I am predictably bad at it.

But in other cases, I might go through the consequentialist calculus perfectly and manage to act according to its dictates but *still* end up bringing about worse consequences than I would have otherwise. There may seem to be a contradiction here, but if the very implementation of the calculus involves costs,

there is not. Acting always for consequentialist reasons might blunt my capacity to care for the people around me or make me difficult or unpleasant to be around, regardless of what actions I perform. Thinking constantly about the suffering that I could be using my time and money to prevent might turn me into a nervous wreck, even if I often reason myself into performing actions of self-care. And there's the obvious point that calculation takes time and mental energy. Again, these are costs associated with the *very application* of the calculus; they could not be avoided by making better calculations. For someone like me, applying a consequentialist calculus might have bad consequences even if I were to calculate correctly in every case!

There is, of course, nothing impossible about creatures that are psychologically equipped to apply and act on a consequentialist calculus with minimal cost, but it seems probable that few human beings are of this sort. For most of us, *trying* in all cases to do the consequentialist thing will make things go suboptimally, or even very badly; in Parfit's terminology, consequentialism is "indirectly self-defeating."[5] I will follow his usage and call any normative theory indirectly self-defeating just in case attempting to follow the theory increases the

[5] *RP* p. 14. Parfit believes that indirectly self-defeating theories are not problematic in the way that what he calls *directly* self-defeating theories are. For these latter theories, individual success guarantees collective failure. This chapter will not explicitly deal with any directly self-defeating theories.

risk (vs. some other available strategy) of failure on its own terms.[6] I will also follow Parfit in calling normative theories "self-effacing" just in case they *admit* that they themselves are self-defeating.[7] That is, a theory is self-effacing just in case it explicitly tells us to try to believe or follow some other theory.

It may not be immediately clear whether an indirectly self-defeating theory could be correct or a self-effacing theory coherent. Williams, who is reluctant to draw a distinction between a false moral theory and a moral theory that should not be adopted, takes the self-defeat objection to be absolutely fatal to utilitarianism.[8] I believe that Railton,[9] Parfit,[10] and others have shown that this reaction is too strong.

Self-effacing consequentialism might seem paradoxical because it might seem to order us not to do the consequentialist thing. "Do not follow this order" is paradoxical in something like the way that the Liar sentence is.[11] If I follow the order, I have not followed it; if I do not follow it, I have followed it. But, in fact, no version of consequentialism that I know actually takes this paradoxical form.

---

[6] More than just moral theories could be indirectly self-defeating. Commands, aims, games—anything with specified success conditions, really—might be indirectly self-defeating.
[7] *RP* p. 24
[8] "A Critique of Utilitarianism" p. 135
[9] "Alienation, Consequentialism, and the Demands of Morality"
[10] *RP* part 1
[11] A theory that took this form would be even more immediately self-defeating than Parfit's directly self-defeating theories, discussed above in footnote five.

Some theories worth calling consequentialist might instead say: "Try your best to make things go well." Let's suppose that following these theories would make things go much worse than they otherwise might. This feature of the theories makes them implausible, since a main intuition supporting consequentialism is that it is important that things *actually do* go well, but it does not make the theories paradoxical. By way of comparison: You tell me that you are feeling stressed and I tell you to focus as hard as you can on not thinking about your responsibilities, which causes you to obsess over them. My advice is bad, because it makes you experience more stress, which is presumably the opposite of what is good for you, but it is not paradoxical or impossible to follow. All you need to do to follow my advice is to focus as hard as you can on keeping your responsibilities off your mind. Though you could do this, you shouldn't.

Better versions of consequentialism say: "Do whatever as a matter of fact makes things go well." If consequentialism is indirectly self-defeating, *trying* to follow such a theory will mean failing to *actually* follow it. This feature presents a practical problem for anyone convinced of the truth of such a theory, but does not make such theories genuinely paradoxical. By way of comparison: "The Game" is a fairly well-known game that one loses whenever one thinks about it. Focusing on doing well at The Game is thus a sure recipe for failure, whereas not caring about winning is generally a good strategy. (Up until I started work on

this paper, I had been doing very well for years.) There are practical problems with playing The Game, but there is no paradox proper; The Game is entirely coherent, and success and failure are perfectly well defined.

Likewise, self-effacing consequentialism—consequentialism that tells us not to try too hard to discover and pursue consequentialist actions at all moments of our lives—presents a problem, but not a paradox or contradiction. The arguments I have just made to support this claim generalize to other self-effacing theories, which is important, because any plausible normative theory compatible with ECR will tell us that we often have strong practical or moral reasons to cause or allow ourselves to ignore many of the normative implications of ECR.

For a reader unsympathetic to consequentialism or ECR, that a theory is self-effacing[12] might seem to be, if not enough to render it paradoxical, then at least some evidence against its truth. But I believe that even this moderated reaction is too strong, because *any* plausible theory of morality has at least the *potential* to be self-effacing in some circumstances—a point to which I'll soon return. If some malevolent intelligence threatens unspeakable harm to all human

[12] Strictly, I should say not that ECR is self-effacing, because ECR makes only the negative normative claim that nothing matters in the way that personal identity is mistakenly taken to matter. ECR does not imply, by itself, that anything does matter. More carefully, I would say that a pairing of ECR together with a plausible moral theory compatible with it will be self-effacing. Such a theory pair is likely to hold, for example, that long-term friendship does not matter intrinsically or finally but that we should let ourselves think and act as if it does.

beings if I do not come to believe and act on some close-but-not-perfect approximation of the true moral theory, then I should certainly do whatever it takes to believe and act on the slightly false moral theory instead of the true one.[13] Importantly, I should do so according to *any* minimally plausible moral theory, not just consequentialism; a theory that told me to try to preserve my true moral beliefs in the face of horrific suffering would be cruel, perverse, and false.

In the terminology of Chapter Two, we should believe that any plausible theory will acknowledge that we might have decisive non-standard state-given reasons to cause or allow ourselves to have irrational beliefs, motivations, or values. That a theory has the potential to be self-effacing in this way in some circumstances thus does not reveal some structural flaw. The fact, if it is a fact, that consequentialism and ECR are self-defeating *more often* in the actual world than are other candidate theories is no evidence against the truth of consequentialism or of ECR.

Can We Get Around Self-Defeat?

I have said that self-effacement and indirect self-defeat present no genuine paradox but that they do present a practical problem. What do I do when I

[13] This case is adapted from Railton's thought experiment about a demon that demands Kantian belief and action of consequentialists ("Alienation, Consequentialism, and the Demands of Morality" p. 155). I describe the case in more general terms in order to make clear that it's not just consequentialism, but rather any plausible moral theory, that is open to self-effacement.

believe a self-effacing theory?

In Chapter Two, I argued that conflicts between standard (object-given) and non-standard (state-given) reasons will generally be practically impossible to reconcile, since the different sorts of reason tend to appeal to different faculties or forms of rationality, or at least to one faculty at different times. In the consequentialist case, we have standard object-given epistemic reasons to have consequentialist moral beliefs alongside competing non-standard state-given reasons to cause or allow ourselves to have non-consequentialist moral beliefs. We also have standard object-given reasons to promote the good alongside competing non-standard state-given reasons to cause or allow ourselves to have a character or motivational profile such that we often fail to promote the good.

Thus, if my arguments in Chapter Two succeeded, we should expect the self-defeating natures of consequentialism and ECR to be pretty stable. However, competition between standard and non-standard reasons is only intractable when there is no way to change our situation so that our reasons become different. For example, if someone threatens to shoot me unless I believe that the moon landing was a hoax, I can get around the irreconcilable conflict between my competing reasons by taking his gun away and driving off. Maybe there is a similarly ideal solution in the consequentialist or ECR cases.

In one sort of ideal case, I might bring my cares in line with my beliefs about what matters. I might reshape my desires, projects, reactive attitudes, etc.

in an entirely new mold in the hopes of minimizing the costs of applying a consequentialist calculus or taking an impersonal view of the world. But it is unlikely that I (or very many other people) could succeed in such a project; indeed, its difficulty is a major reason that consequentialism might be indirectly self-defeating. Even if I could somehow succeed in radically reshaping my personality, it is not clear that doing so would be warranted even on my own terms, as it would only allow me to escape *some* of the traps that make consequentialism and ECR indirectly self-defeating; I might minimize the direct psychological toll of promoting impersonal good over all else, but I might still be prone to calculation errors or be seen as untrustworthy or undesirable as a friend. I might stop caring about people qua person but in doing so come to alienate my friends and family. And so on. This is not even to mention the psychological toll that trying to reshape my personality would likely entail, whether or not I managed to succeed. It seems unwise to try to bring my cares in line with my beliefs.

In another sort of ideal case, I might bring my beliefs in line with my cares. I might succeed in convincing myself that consequentialism or ECR is false and that some other theory is true. If I were to do a good job of picking the new theory out, I might manage to succeed much better on my old terms than I would have if I had continued believing in consequentialism or ECR. The most immediate problem with this strategy is, of course, that it is extremely hard to

make oneself believe something for purely practical reasons when the epistemic reasons seem to rule it out. If I am threatened with some terrible misfortune unless I come to disbelieve some obvious but unimportant truth, it seems clear that I should make myself disbelieve it but equally clear that I will not be able to do so by ordinarily available means.

Moreover, it is not clear that I should come to believe a theory that is *exactly* in line with my ordinary cares. It may be that allowing *some* gap between what I believe and what I personally care about is ideal. It might be best, for instance, to have a tendency to revert to consequentialist reasoning when it comes to some extremely high-stakes decisions.[14] Or there might be something undesirable about psychologies that can shrug off Singer-style thought experiments[15] that are meant to demonstrate the moral irrelevance of factors like physical distance which we may nevertheless be psychologically incapable of

---

[14] Hare argues for this claim, which strikes me as extremely plausible, in his defense of "two level" utilitarianism (*Moral Thinking*, especially ch. 3). On Hare's view, the utilitarian should be able to make both slow, considered, utilitarian judgments and quicker intuitive judgments. He compares this task to that of a commander keeping both strategy and tactics in mind during battle (p. 52). As I will ultimately suggest that we must live with tension between our beliefs, cares, and desires, this sort of multi-standpoint or multi-strategy picture is not too different from my own. But I believe that Hare is too optimistic about the ease of this sort of project. Keeping difficult moral truths in mind alongside ordinary cares is unlike keeping strategy and tactics in mind because strategy and tactics do not directly undermine one another.

[15] In "Famine, Affluence, and Morality," Singer famously compares the choice one has to save famine-threatened lives by donating money with the choice one would have when walking past a drowning child to save it by jumping in and ruining one's clothes. In both cases, lives can be saved at a non-negligible but relatively inconsequential cost. Singer argues that, because there seem to be no *morally relevant* differences between the two cases, we are as just as required to donate our money to famine relief as we would be to save the nearby drowning child.

ignoring in all cases without substantial psychological cost. And so, even if I could practically come to bring my beliefs in line with my cares, it is far from clear that I should.

A middle strategy, wherein I attempt to adjust my desires and beliefs so that they meet somewhere in the middle, inherits all of the problems of the two strategies I've just described. Ultimately, in both the consequentialist and ECR cases, there seems to be no clear way to close the gap between standard and non-standard reasons so that they no longer compete.

In giving these arguments, I do not mean to suggest that *no* creature, or even no human being, could align their cares, beliefs, and desires in such a way as to avoid the possibility of reasons alienation. Perhaps a Buddhist sage could manage it, or a person with deeply impaired capacities for empathy and emotion but a fully functional capacity for moral reasoning. It might also be that technology could provide us with an answer. But for most of us living in the present, myself included, aligning our cares, beliefs, and desires is not a realistic option.

In short, if I believe that consequentialism or ECR is true and self-defeating, then I am *morally and practically required* to accept a disconnect between what I believe matters and what I care about. To the extent that this disconnect is unavoidably alienating, as I have argued that it must be, I am

morally and practically required to live with Reasons Alienation rather than attempting to eliminate it from my life.

Everyday Reasons Alienation

In Chapter Two, I characterized Reasons Alienation as the sort of alienation that results from a disconnect between what I believe I do, should, or must care about on the one hand and what I believe actually matters on the other. I argued from cases that this disconnect is indeed alienating and then attempted to account for this fact (and for the stability of the disconnect) by appealing to the practical irreconcilability of competing standard and non-standard reasons. In this chapter, I elaborated on the specific sorts of competing standard and non-standard reasons at play in the cases of consequentialism and ECR and argued that the tension they present does not undermine the plausibility of either theory. I then argued that we should not expect to be able to change our situations so as to align our standard and non-standard reasons in these cases, which suggests that the disconnect between our beliefs and cares will (and should) remain stable and we will have to live with Reasons Alienation rather than finding a way around it after all.

Some philosophers writing about consequentialism and reductionism about personal identity would reject these claims. They would argue that the disconnect between belief and care need not exist or that it need not be

alienating. My appendix defends my view against these arguments, and for the rest of the chapter proper I will assume that alienation is the real, necessary result of ECR and consequentialism. My eventual "solution" will therefore not pretend to eliminate or even minimize Reasons Alienation but will instead suggest a strategy for living with it. But before I get to this solution, I want to ask my readers to engage in some introspection about sources of possible Reasons Alienation in their own lives and to consider what a tall order it would be to handle even the simplest of these—too tall, I think, for any "good faith" solution to the problem of Reasons Alienation to be successful. After doing so, even readers unsympathetic to ECR or consequentialism may discover that my solution nevertheless has something to offer them.

Many of us—I would think the vast majority—have things in our lives that we care about greatly but that, upon reflection, we could be convinced do not actually matter very much. Probably even more of us have things that we recognize intellectually to be important but that we do not care very much about in everyday life. We all seem to get along fine in the face of these disconnects, but they are alienating nevertheless.

I'm going to recommend a sort of exercise: Think, if you can, of something that matters especially to you that you recognize may not be as objectively important as your care would suggest. If you believe some theory, like ECR or act utilitarianism, that has unintuitive normative implications, the exercise

should be easy. But I think that most anyone who is honest with themselves should be able to manage it. There are any number of things that might fit the bill—the success of a local sports team, the happenings in a fictional universe, the outcome of a months-long multiplayer strategy game you're playing with friends—whatever. Once you've picked one, focus as hard as you're able on its ultimate triviality. Remind yourself that your team doesn't deserve the championship any more than their rivals, that it doesn't actually matter whether your favorite characters live or die (except perhaps insofar as it impacts the quality of the story), and so on. These are obvious truths but they are also—maybe for that reason—easy to keep out of mind.

When I think about facts like these, I feel uneasy. My cares begin to seem silly and unwarranted, and my normative beliefs begin to seem cold and alien. Perhaps strangely, I don't stop caring, and I don't change my beliefs. And I don't know that I could—at least not so quickly. Instead, I feel disconnected, like I am reading a novel in which I—another I—am a character. I find that I cannot fully inhabit the two roles simultaneously. As a character, I lack the reader's clear image of the world of the novel as essentially fiction. As a reader, I lack the character's humanity, his depth of feeling and engagement. Neither role is fully *me*. And I believe that this must be so. I cannot at the same time and from the same perspective identify myself fully with some care and also believe that the

thing that I care about is completely worthless and unimportant. There must be a split.

This might all seem overly-dramatic. After all, I've only recommended meditating on the ultimate unimportance of the events in the Star Trek universe or some such thing. But there are other, tougher thoughts that can do the trick. Try thinking about the ultimate ordinariness of the people that you care most about, and how someone else could easily have taken their places if the timing had been right. Or think about all of the suffering in the world that you put out of mind on a daily basis because you couldn't get through the day if you didn't. Think about how unimportant *you* are in the scheme of things.

Depending on your working theory of value, these exercises may not have much of an effect. You might be content that, though many of the things you care about do not matter much in some objective sense, they nevertheless warrant *your* care. (I think it's implausible that *all* of our cares could survive close scrutiny even on a heavily subjectivist or partialist theory of value, but no matter.) If this is so, all I can think to do is suggest that you try on another theory of value, just for the sake of the thought experiment, and see if you can get a sense of the sort of alienated feeling that I am trying to get across.

What to do?

I have argued at length that it is impossible to reflect honestly on one's

cares and values without experiencing Reasons Alienation and that this is particularly true for those of us whose honest reflection has led us to accept views like hedonistic act utilitarianism or ECR. I expect that for some of of us in this latter group, this alienation is disconcerting, if not outright painful. (It is for me.) It demands a practical response.

A good response cannot be a straight one. That is, a good response cannot and should not attempt to eliminate my experience of alienation, nor should it even necessarily take minimizing alienation as a central aim; as I have said, it is much better to lead a productive, fulfilling, alienated life than a parasitic and miserable one in good faith. On the list of things that contribute directly or indirectly to value—even on the narrow conception of value that survives ECR—good faith is going to rank a lot lower than things like happiness, the capacity for friendship and care, the capacity for theoretical and practical rationality, and so on. These more important components and enablers of value are, in my view, threatened more by an insistence on good faith than they are by Reasons Alienation itself. It is better, if we can manage it, to care irrationally without undermining our capacity for practical reason than it is to blunt our capacity to care or reason for the sake of self-unity or authenticity. Thus alienation is to be managed and accepted, not minimized or eliminated. But how do we manage and accept alienation when we are faced constantly with the plain, glaring fact that we are at odds with ourselves, off balance, irrational?

I do not know a totally satisfying answer to this question, nor do I know if there can be one. But I can describe the strategy which I have tried to apply in my own case and with which I have found some success. I do not know that this is the best strategy for everyone, or even for myself, though I believe that it has much to recommend it. The strategy, to the extent that I can put it into words, involves a sort of compartmentalization of the self, an anti-Korsgaardian *dis*unity of agency, a good faith acceptance of bad faith and irrationality, and the active maintenance of an elaborate, humanizing, and simultaneously totalizing and contingent fiction. Or, better than fiction, kayfabe.

Kayfabe and Wrestling

In professional wrestling terminology, "kayfabe" is the fictional world that the practice of wrestling creates and in which it is supposed to reside.[16] In kayfabe, wrestling is is a sporting competition, not a collaborative performance. In kayfabe, The Rock's purely theatrical People's Elbow finisher is more likely to render a combatant unable to continue than is a torn quadricep, a concussion, or a dislocated shoulder. In kayfabe, The Undertaker is some kind of undead wizard who nevertheless fights (and occasionally loses to) ordinary men and Braun Strowman can flip an ambulance with his bare hands. The concept of

[16] Actually, the word "kayfabe" has a few different related uses—it can be used to talk about the fiction itself, the fiction's status as purportedly real, or the norm that the fiction should not be revealed as fiction.

kayfabe can be extended beyond the world of wrestling, because different activities can create similar fictional worlds—politics, interpersonal relationships, and so on. And so it makes sense to talk not just about kayfabe but about kayfabes.

As I'll use the term, a kayfabe is more than an ordinary fiction in at least two ways.[17] First, a kayfabe is all-encompassing. In a kayfabe, the world of the fiction is just the real world. This is not ordinarily so in most fiction. When I watch an ordinary staged performance, the characters played by the actors are not supposed to be aware of me or of the other audience members, because we are not a part of their world. When I read an ordinary novel, I don't usually think of the characters as standing in some particular spatio-temporal relation to me. This is often true even of novels set in "the real world;" when I entertain such fictions in the usual way, I think of them more as ways things could have been than as ways things *are* (but of course really aren't). So, for instance, when reading a book set in downtown Austin I do not believe, or even pretend, that I could hop on the 7 bus to see the action unfold.[18] Entertaining most fictions is

[17] If you prefer to think of a kayfabe as a just a special kind of fiction, nothing much should be lost.

[18] Galen Strawson has pointed out to me that this may not always be the case for all readers. Walking tours of Dublin that stop at the various places where Leopold Bloom spends time in *Ulysses* might suggest that people like to imagine the world of the novel as being part of our real world. The fact that we sometimes entertain ordinary fictions in this way—that occasionally the novel rises to the level of professional wrestling—does not threaten the point that kayfabe is interestingly different and worth distinguishing from ordinary fiction so long as *most* or *all* ordinary fiction does not invite this sort of

usually just that—entertaining. Properly entertaining a kayfabe is more like make-believe.[19]

Second, and relatedly, a kayfabe is maintained and insisted upon. After an ordinary play, I think nothing of seeing the actors who played the hero and the villain getting on stage and bowing together, because the fiction has come to a close. At a wrestling show, at least before the 1980s, such a thing would be unheard of. In past decades, wrestlers would never publicly admit that the world of wrestling was a fiction. Even today, wrestlers traditionally at least wait for a change of venue to relax their characters and "break kayfabe."

While a kayfabe is more than an ordinary fiction, it is also less than an ordinary lie.[20] A kayfabe, like an ordinary fiction but unlike an ordinary lie, can survive the common knowledge that it is false. If it is common knowledge between me and my friend that I stole his bicycle, there is no point in insisting that I didn't. On the other hand, savvy wrestling audiences have long

---

reification, or so long as this reification is not ordinarily as central to appropriately entertaining such fiction. (Though see the next footnote.)

[19] Kendall Walton has argued that entertaining *all* fiction means engaging in a sort of make-believe, for the reason that to hold otherwise would imply that we experience genuine emotions about things that we know are not real ("Fearing Fictions"). The argument is ingenious, but ultimately I believe that views on which we do in fact experience genuine—not pretend—emotions in response to fiction are more promising. (For one such view, see Noël Carroll's *The Philosophy of Horror* pp. 60–88.) Even if a Waltonian view turns out to be true, there seems to be an important difference between the more active, voluntary, and encompassing game of make-believe played by a wrestling audience and the more passive, non-voluntary, and localized one played by, say, a movie-watcher.

[20] If you prefer to think of a kayfabe as a just a special kind of lie, nothing much should be lost.

understood that wrestling is not real but have only recently begun to tolerate breaks in kayfabe in specific contexts.

A kayfabe, then, is a false image of the real world that is maintained at almost all times, and especially (or most interestingly) one that is maintained in spite of open knowledge of its falsity. Besides the traditional wrestling kayfabe, there are political kayfabes (as when a politician makes some obviously impossible policy a part of their election platform), familial kayfabes (the insistence of a parent that they love all of their children equally is often like this), and others.

What is the purpose of a kayfabe? Why entertain a fiction at almost all times? Why maintain a lie that is known to be false? This is a difficult question, and probably there is no single answer. Sometimes, kayfabes are probably maintained just for the sake of saving face, as in the case of the parent who insists against all evidence that they love their children equally. Other times, when the falsity of the kayfabe is not quite common knowledge, there may be some fun in feeling like one is in on a secret.[21] But to my mind, the most interesting function of a kayfabe is that it allows people to entertain a fiction more fully, deeply, and personally than they might otherwise be able to.

[21] I suspect that this may have the case for much of the history of wrestling—that it was more widely understood by the audience that wrestling was a spectacle than many of the fans or even performers realised, and that many (falsely) believed that they were in on an exclusive secret.

In support of this point, let me tell a story. CM Punk, aka Phil Brooks, was one of the greatest professional wrestlers of the modern era before his early retirement in 2014. In September 2016, after two years of training, Punk had a mixed martial arts fight—a real fight—at UFC 203. He lost, badly. This result was unsurprising. Punk was past his physical prime, was totally unproven, and had gone through much less MMA training than his opponent Mickey Gall. But for a lot of wrestling fans, myself included, the loss felt like a punch in the gut. This guy was a former world champion. He won that title by beating John Cena clean in one of the best matches of the last decade. We know that wrestling is fake, but still—how could he lose? How dare he lose?

Suppose LeVar Burton were to go on Jeopardy and get creamed.[22] Here's a guy who played the brilliant Geordi La Forge on Star Trek: The Next Generation, but he loses at jeopardy? Well, so what? We know that Burton isn't La Forge. We know that La Forge isn't real. Those of us who like Burton's work might feel bad for him, but we wouldn't think that the Star Trek fiction had somehow been undermined. Why the difference? The answer, I think, is kayfabe.

Roland Barthes says this about wrestling:

> When the hero or the villain of the drama, the man who was seen a few minutes earlier possessed by moral rage, magnified into a sort of

[22] I have no reason to believe that this would be what would happen if LeVar Burton were to go on Jeopardy; I have no notion of Burton's knowledge of trivia.

> metaphysical sign, leaves the wrestling hall, impassive, anonymous, carrying a small suitcase and arm-in-arm with his wife, no one can doubt that wrestling holds that power of transmutation which is common to the Spectacle and to Religious Worship. In the ring, and even in the depths of their voluntary ignominy, wrestlers remain gods because they are, for a few moments, the key which opens Nature, the pure gesture which separates Good from Evil, and unveils the form of a Justice which is at last intelligible.[23]

For one type of fan, because this transformation from humble, anonymous husband to godlike figure is central to the magic of wrestling, the ordinary unmagical reality is something to focus in on and keep in mind in order to bring the transformation into sharp relief. But there is another type of fan that does not like to dwell on the fact that these godlike figures are, after all, just ordinary people with a particular talent for performance. This sort of fan knows, of course, that gods don't walk among us, but they would like to in some way believe—to suspect—that they do. It is these fans that kayfabe serves best. I believe that the ECR theorist can learn something from this sort of wrestling fan.

[23] "The World of Wrestling" p. 23

Normative Kayfabe

ECR shrinks the scope of normativity. When someone comes to believe ECR, they come to believe that there are far fewer reasons or sources of value in the world than people ordinarily suspect. Since it is inadvisable either to let this realization do too much to shrink their capacity for care *or* to let their care force them back into a false moral view, the ECR theorist ought to take steps to insure that neither happens and that the experience of the resulting gap between belief and care is no more painful than it has to be. One way to do this, in my view, is for the ECR theorist to enact a sort of normative kayfabe.

In the normative kayfabe, people can deserve praise or blame for past actions. In the normative kayfabe, lives matter as wholes, not just as the sum of individual experiences. Friendships and relationships and the rationality of love and affection are partly grounded in shared histories. The fact that someone has worked long and hard for something makes it all the more worthwhile when it happens. And so on. None of these things, of course, is *true*. But the ECR theorist entertains them in the special way that one entertains a kayfabe: They pretend that they are true and act in accordance with their truth when circumstances allow. They never lose sight of their falsity, but only openly *focus* on this falsity under special circumstances—when it would be pleasant or productive to do so, when engaged in serious conversation, and so on. They build for themselves (or let stand) a working model of the world that operates according to the rules of

the normative kayfabe, and they spend most of their time and their thought at least waist-deep in that world.

There is an apparent disanalogy between the wrestling case and the individual moral case, however. The existence of the kayfabe in wrestling depends on the gap between performer and audience. Wrestlers do not keep kayfabe with one another. They couldn't; a wrestler who believed wholeheartedly in the fiction or even took it too seriously would be an unsafe performer, would undermine stories, would ruin friendships, and so on. But the ECR theorist is just one person. How can someone keep kayfabe with themselves? The best answer I can give, though it's an imperfect one, is that we ECR theorists should allow ourselves to compartmentalize our fictions.

When we care, appreciate, and experience—when we are in a position to be swept up in all of life's apparent richness—the normative kayfabe is our friend. In this, we are like an audience member at a wrestling show, along for the ride and enjoying the fiction. If entertaining and enacting the normative kayfabe helps a person develop deeper care for their friends and family, if it helps them find greater satisfaction in work and in life, if it helps them maintain hope for the future, they ought to do it. In life, as in wrestling, it is unpleasant, taxing, and ultimately unfulfilling to keep the truth in full view. The fans who approach wrestling in this way—so-called "smarks"[24]—seem plainly to be missing out on

[24] In traditional wrestling parlance, a "mark" was an audience member who believed in the reality of the spectacle. The term "smark," short for "smart mark," became used to

much of what the medium has to offer. This charge seems even more telling against those who would adopt a detached, hyper-rationalist approach to life that cares only about how the sausage is made and not at all about its taste.

When we act, the story is more complicated. On the one hand, the best action will be the one most favored by the narrow set of reasons that survive ECR. On the other, as we saw in the consequentialist case, trying to respond only to the reasons that actually exist might not be the best way to act well in the long run. Thus, as an agent, a person ought to entertain the kayfabe only in a limited way. At the risk of straining the analogy, a person should in action be like a wrestler in the ring. They ought to keep reality in mind to the extent that they can act well (and safely), but they ought to do so with an eye to the fiction; they should not do too much to expose or undermine it unless the stakes are high.

How much is the right amount to entertain the moral kayfabe in action? I don't have a good answer to that. I expect that the answer is different for different people, and the question is one that is probably better suited to empirical psychology than to moral philosophy and introspection.

---

describe fans who understood that the spectacle was a fiction but nevertheless remained fans. More recently, as genuinely duped marks are understood on all sides to be a very rare breed, "smark" has taken on a more narrow usage. On the contemporary usage, a smark not only understands that wrestling is staged, but shifts much of their focus from the fiction to what they imagine to be happening behind the scenes. A smark might cheer good writing instead of heroic characters or jeer when a performer botches a move in a way that exposes it as fake. There can of course be a certain pleasure in seeing slivers of reality through cracks in kayfabe, but many wrestlers and fans resent the way in which smarks actively undermine the illusion—with good reason, in my mind.

Unfortunately, it's also a practical question that demands an answer of us, so we've got to do the best that we can with the tools that we have. In answering the question for ourselves, all we can do is step back, take a full view of reality, and try to come up with a plan for living with one foot in a fiction. Something like what I am doing now in writing this chapter. I can't resist extending the wrestling analogy: We must, occasionally, step back and, like the writers of a wrestling show, "book" our own lives. Then, when we step back into everyday life, we do our best to put on a show that everyone can enjoy.

Afterword – It's Not All Bad

The bulk of this dissertation deals with the threat of alienation that I take to be the central practical problem facing the Extreme Claim Reductionist. As such, the reader could be excused for thinking that ECR is all doom and gloom. But this is not so. I focused on the alienation problem because of its philosophical interest and practical urgency. The various positive upshots and silver linings of ECR are, if not necessarily less interesting, at least less pressing; it is easier to deal with good things than bad ones. But it is worth spending some time on those upshots now—not just to provide a relief from pessimism, but because understanding them will be an important part of any fully-worked-out strategy for living with ECR.

In what may be the most-quoted bit of *Reasons and Persons*, Parfit says this about his experience with reductionism:

> Is the truth depressing? Some may find it so. But I find it liberating, and consoling. When I believed that my existence was such a further fact, I seemed imprisoned in myself. My life seemed like a glass tunnel, through which I was moving faster every year, and at the end of which there was darkness. When I changed my view, the walls of my glass tunnel disappeared. I now live in the open air. There is still a difference between

my life and the lives of other people. But the difference is less. Other people are closer. I am less concerned about the rest of my own life, and more concerned about the lives of others.

When I believed the Non-Reductionist View, I also cared more about my inevitable death. After my death, there will no one living who will be me. I can now redescribe this fact. Though there will later be many experiences, none of these experiences will be connected to my present experiences by chains of such direct connections as those involved in experience-memory, or in the carrying out of an earlier intention. Some of these future experiences may be related to my present experiences in less direct ways. There will later be some memories about my life. And there may later be thoughts that are influenced by mine, or things done as the result of my advice. My death will break the more direct relations between my present experiences and future experiences, but it will not break various other relations. This is all there is to the fact that there will be no one living who will be me. Now that I have seen this, my death seems to me less bad.[1]

[1] *RP* p. 281

Parfit is not an Extreme Claim Reductionist; he is agnostic between the Extreme and Moderate Claims. But I believe that the liberatory experiences he describes, or ones very close to them, are available to the ECR theorist.

Parfit claims that coming to believe reductionism has caused at least two positive changes in him. First, it has made him feel less separate from, and thus closer to, other people. Second, it has made him less afraid of death. I'll consider his claims in reverse order.

Death

The second change that Parfit describes—becoming less worried about death—would seem to be equally available to the ECR theorist, and I have already considered it to a degree in Chapter One. If ordinary death is not first-personally worse than going through a teletransporter, and going through a teletransporter is not first-personally worse than ordinary survival, then ordinary death is not first-personally worse than ordinary survival.

To banish death in this way may seem like a Pyrrhic victory. As Parfit notes,[2] when we come to believe reductionism, we do not learn that duplication, teletransportation, or death give us much of what we wanted out of ordinary

[2] "When I come to see that my continued existence does not involve this further fact, I lose my reason for preferring a space-ship journey [to Teletransportation]. But, judged from the standpoint of my earlier belief, this is not because Teletransportation is about as good as ordinary survival. It is because ordinary survival is about as bad as, or little better than, Teletransportation. Ordinary survival is about as bad as being destroyed and Replicated" (*RP* p. 280).

survival. Instead, we learn that ordinary survival does not give us what we wanted out of it. If we think that we were rational to want something more from ordinary survival than it in fact has to offer and correct to believe that death is bad because it deprives us of what we wanted, then reductionism implies that survival is as bad as we thought death was.

If it were indeed a fact that survival is as bad as most people take death to be, it would be a horrifying one. Anyone who could believe it wholeheartedly would probably be driven out of their mind. Fortunately, it is probably not a fact. Rather than being mistaken about the acceptability of survival, we are probably mistaken about the badness of death. We cannot rationally want more from survival than it has to offer, in part because it is impossible that it could have offered more, and so we should not be upset that death does not offer it either.

This impossibility claim may sound too strong. Parfit believes that there *might have been* a deep further fact about identity, even though there is not.[3] There might, for example, have been Cartesian souls. Be this as it may, I doubt that any further fact, regardless of depth, could have been enough to make survival matter in the way that we want it to. As I argued briefly in Chapter One, the R relation would probably be the best candidate for what matters even if there *were* a deep further fact about identity. We would not be better off if there were Cartesian souls, base-level brute facts, or anything else.

[3] *RP* pp. 227–228

Perhaps it sometimes makes some sense to wish for impossible things. Maybe it makes some sense to wish that Gödel's incompleteness theorems were false so that there could be hope for a single formal system that could describe all of mathematics. But in the case of survival, it's not even clear that we know what it is that we want. When I want to survive in the distinctly first-personal way,[4] what I want is for *myself* to continue to exist. I want to *still be here* in a way that is more demanding than the continued existence of my soul or even a brute identity fact seems prepared to guarantee. When I step back and think on what it is that I want, my desire seems not only impossible but fundamentally misconceived. What it is that I want is something that I cannot clearly imagine or describe, since I cannot imagine or describe a way things could be that would satisfy my desire. I believe that it would be a mistake to be disappointed that such a desire would be frustrated. It makes much more sense, and is much easier, to become less afraid of death rather than becoming disappointed in survival.

Friendship

Parfit's other claim is that his belief in reductionism has caused him to feel less self-involved and more concerned with other people. This change might seem to be less available to the Extreme Claim Reductionist. If R had mattered, it would be clear how we could feel for others in the same way that we feel about

[4] As opposed, for example, to wanting to survive for the reasons that my death would cause a great deal of pain and shut off the possibility for a great deal of happiness.

our own pasts and futures, because we can have ties very much like the R relation to other people. I can have memories of shared experiences, intentions for shared projects, and so on. But the Extreme Claim says that R does not matter in the right way. Perhaps, though the Moderate Claim Reductionist can break through their glass tunnel, the ECR theorist is doomed to retreat within it.

I believe that this pessimistic conclusion is too strong. It is true, as I have maintained throughout this dissertation, that ECR undermines our justification for such things as partial benevolence, loyalty, gratitude, and so on. But it would be a mistake to think that every aspect of things like our friendships and relationships depends entirely on such things.

ECR says that nothing matters in the way that we mistakenly take the deep further fact about personal identity to matter.[5] This definition shouldn't be taken to mean that all values or attitudes that people think of as depending on personal identity are valueless or irrational. Some of these values or attitudes might have other grounds, or people might be mistaken for tending to ground them entirely in identity. I believe that friendship, to take the paradigmatic example of care beyond the self, is one such case.

Suppose you were to learn, as Bertrand Russell once hypothesized, that the world had come into existence five minutes ago, with all apparent memory

[5] I put it in these terms rather than talking about "the way that a the deep further fact about personal identity might have mattered" because, as I say, I do not think that any facts could have mattered in the right way. The extreme claim can only be defined in relation to what I have come to believe is likely a confused set of beliefs.

and history just as it is now.[6] The news would, no doubt, be very disturbing. Learning that the world is only five minutes old would be very similar in terms of normative consequences to learning the truth of ECR, at least in the backwards-looking direction. (If ECR seems less shocking, I think it is only because the view is abstract and difficult to understand and believe, whereas we can understand immediately what it would mean for the world to have only just come into existence.)

In the five-minute-old world, you would never have met any of your friends (or "friends"), though by hypothesis you could be sure that they were out there and that they would seem to remember you when you next saw them. Assuming you could get past the shock of learning that the world was new, how would you react to your friends when you saw them? Or how should you?

Let me invent two friends for you, cobbled together from some of my own. Eric, or so you thought this morning before learning the true age of the world, has been your friend since childhood. You've been with each other through good times and bad, and he's helped you through some of the hardest times in your life. Over the years, you've grown different—new personalities, new priorities, new perspectives—to the point that you no longer really enjoy one another's company except insofar as it reminds you of the bond that you share. Then there's Fiona. Fiona and you don't have much of a shared history,

[6] *The Analysis of Mind* pp. 159–160

but every time you see each other there's a certain electricity—the conversation flows freely, and you both feel energetic, happy, and alive.

I think that learning that the earth is only five minutes old might well threaten the foundations of your friendship with Eric. You might decide that there is no compelling reason to see him or otherwise maintain the friendship in the absence of your shared history. But it should not threaten your friendship with Fiona very much at all. It is true that much of what we take to be important in many of our relationships with other people, as I argued in the case of Arthur and Andrew in Chapter Two, does seem to depend on things that would be threatened by ECR or by the five-minute hypothesis. But plenty survives. When you next saw Fiona, even though you would know that you had never met her before, I expect that you would still feel the same electricity, connection, and bond that you would seem to remember. And why wouldn't you? None of *these* feelings or attitudes depend for their justification on anything about a shared history or anything of that sort. They are just pleasant, fulfilling ways that another person makes you feel. Moreover, the fact that she can engender these feelings in you seems to make her—the her that exists now, detached from any history—a worthy object of friendly and affectionate feelings.

I don't want to overstate my case here, lest I undermine the work of Chapter Two. Many of our attitudes and cares, including some of our attitudes about our friends, projects, and so on, will be left without needed justification.

But others will be unaffected. Some will be left without *unneeded* justification. And coming to recognize that these attitudes and cares *never needed* any justification of the sort we might have demanded might be the most hopeful and freeing consequence of accepting ECR.

Freedom

Many of our attitudes seem to require a certain sort of normative cognition in order to be warranted. It is unwarranted and irrational to despair over something that we see is not bad or to feel guilty when we know we haven't done anything wrong. It is likewise unwarranted to feel happiness or relief at bad news (though perhaps, if hedonism is true, it is still in another way *good* to feel happiness in such cases). Other attitudes do not seem to have such a requirement. I enjoy eating peaches, and I don't have to believe that peaches are *good* for my enjoyment to be rational. It is the same with the way that I enjoy being with friends. More controversially, it might be this way with art. And so on.

I do not have any worked-out theory of precisely which attitudes, cares, and feelings demand some normative cognition in order to be appropriate, but plenty seem not to. And there are many more that, though they may make such a demand, and though that demand may not be rationally satisfiable, we are inclined or adept enough to adopt without the experience of alienation being too

painful. It is easy, for example, to have feelings of gratitude or vengefulness even without believing that people can be truly responsible for past actions, and in fact it is hard not to. To the extent that accepting ECR expands the space that these various attitudes, cares, and feelings take up, it can help to free us from the felt responsibility of always reacting in the right ways to the right things.

An analogy: When we first get into philosophy and learn about reasons to be skeptical about one thing or another—morals, say, or numbers—the knowledge is quite troubling. But as we learn how the same sorts of skeptical arguments can be applied across the board to things like intentionality, the existence of an external world, and so on, the news starts to lose its bite. Maybe we *should* be skeptical of morality,[7] but if morality isn't clearly in any more danger than math, basic empirical facts, our ability to have beliefs about the external world, and so on, then it is much less tempting to fall into a nihilistic despair (or abandon).

It's similar (though not *too* similar) in the case of ECR. The Extreme Claim is an extreme claim. It shrinks the normative realm almost beyond recognition. This shrinking has at least two hopeful effects. First, as I've argued with friendship, the stuff that survives is brought into sharper relief, and we can expect it to become a bigger part of our lives. Second, so much of our ordinary perspective is upended that we are given a unique opportunity to reshape it. To

[7] I don't think we should, but nevermind.

a large extent, as I have argued, this reshaping will be alienating; we can't and shouldn't try to be the sorts of people that care in all the right ways about all the right things. But once we've dropped that expectation, why not have a little fun with it?

It would be frivolous, we might ordinarily think, to throw oneself into the world of exploitation cinema, crosswords, etc. (Or professional wrestling.) This might be so. But ECR shows us that many—though definitely not all—pursuits are similarly frivolous. We should, to the extent that we are able, pursue the more worthy pursuits; we should do what we can to combat suffering and build a better world. But we always knew that. In accepting ECR, we can hopefully come to see how a wide variety of relationships, cares, projects, and so on are at least no worse than many others that we might have otherwise felt compelled to value or pursue.

Thomas Nagel says this about the experience of the Absurd: "If *sub specie aeternitatis* there is no reason to believe that anything matters, then that doesn't matter either, and we can approach our absurd lives with irony instead of heroism or despair."[8] It's not quite that easy for the Extreme Claim Reductionist. Things still matter. There is still, for example, state-given reason to prefer irony over despair if you can manage it. More generally, it matters a great deal how we take the news that ECR is true. Some ways of taking it might cause us and the

[8] "The Absurd" p. 727

people around us a great deal of suffering. Others might not. Still, lots of things we probably thought mattered don't, or they don't matter in the way that we might have thought they did. If, as I argued in Chapter Three, we ought to commit ourselves to disjointedness, contradiction, and alienation—if we should build for ourselves and come to inhabit a new normative kayfabe—then we might have reason to hope that this kayfabe, along with all the projects, values, attitudes and relationships that it recommends, will fit us better than our old beliefs did. That is what I hope, at any rate.

Appendix – No Easy Way Out

In Chapter Three, I give general arguments that we should not attempt to escape the Reasons Alienation engendered by Extreme Claim Reductionism (ECR) and even by everyday life. In this appendix, I give specific arguments against existing attempts to answer the problem of alienation for consequentialists and reductionists about personal identity.

For the impartialist consequentialist, objective impartial good is what matters full stop, and hence what warrants care. But, as I have argued, what matters full stop will not (and cannot and should not) be what the impartialist consequentialist actually cares about. The case is similar for the ECR theorist. In both cases, the gaps between what actually matters and actual care will be much bigger than the small fissures that, using the exercise I suggested in Chapter Three, I argued exist for most people. If the exercise worked, you should be skeptical that it is possible to focus on even these small fissures without experiencing some alienation, and all the more skeptical that it would be possible for the impartialist consequentialist or the ECR theorist.

Nevertheless, as I have noted, some philosophers have argued that consequentialism need not be alienating. Following Siderits,[1] I will classify these arguments as either "direct" or "indirect." Taking paradigm examples of each

[1] *Personal Identity and Buddhist Philosophy* (henceforth *PIBP*) pp. 135–137

strategy, I will argue that neither works in the way it intends to. Then, turning my attention entirely to personal identity, I'll argue against Siderits's own "hybrid" response to the personal identity version of the alienation objection.

Railton's Indirect Response

The indirect response to the alienation objection to consequentialism holds that when we do something that we know will have badly suboptimal consequences, we can often console ourselves with the knowledge that the act stemmed from good character, healthy motivations, or similar. So consoled, we will be free from alienation.

Peter Railton defends the indirect response by way of thought experiment. His central example[2] is of a man, Juan, who in a time of stress takes an extra trip to visit his wife Linda instead of donating the cost of the ticket to Oxfam. As Railton imagines the case, donating the money would result in better consequences. But, if Juan had had the sort of character that resulted in him donating the money, he would have accomplished far less good in his life.

Railton stresses that the point is not that forgoing the trip would damage Juan's character—if that were so, it might in fact be worse in consequentialist terms for him to give the money to Oxfam. What Juan does *really is wrong* by

[2] "Alienation, Consequentialism, and the Demands of Morality" p. 159

consequentialist lights.[3] But it is in some sense necessitated by Juan's having the best sort of character for him to have—the sort of character that, we can suppose, Juan has developed over a long lifetime lived just as a consequentialist would recommend.

Railton's claim looks to be that, even if Juan is a consequentialist, there is still no reason for him to experience any alienation over his decision to visit his wife Linda.[4] Perhaps thinking the implication obvious, he does not (as far as I can see) give further argument for it. But one can see how such an argument might go.

One might argue as follows:

1) "Ought" implies some sense of "can."
2) John's character is such that he truly *cannot* give the money to Oxfam in the relevant sense.

[3] As I read Railton, this passage from a few pages earlier (pp. 157–158) must be meant to apply to cases like Juan's: "The objective act-consequentialist would thus recommend cultivating dispositions that will sometimes lead him to violate his own criterion of right action. Still, he will not, as a trait-consequentialist would, shift his criterion and say that an act is right if it stems from the traits it would be best overall to have [...] Instead, he continues to believe that an act may stem from the dispositions it would be best to have, and yet be wrong (because it would produce worse consequences than other acts available to the agent in the circumstances.)"

[4] I don't know if Railton ever makes the claim as explicit as I am doing here, but I don't believe that I am putting words in his mouth. The stated aim of the paper is to show that act consequentialism need not be seriously alienating, and the case of Juan and Linda is positioned as the final, decisive piece of evidence for that claim.

3) Thus, John *is not required* to give the money to Oxfam and need not feel in any way alienated for failing to do so.[5]

I do not believe that this sort of argument can work. There is no single sense of "can" which could simultaneously make premises 1 and 2 true, and if there is, it would only be enough to secure Juan's action a very weak sort of permissibility that we should not expect to free him from feelings of alienation. The mere psychologically impossibility of willing some action can never provide a satisfying excuse so long as the action *presents* itself as possible. Even if Juan could not in fact summon the willpower to donate to Oxfam, he surely views donation as an option that he might conceivably choose. And this could well be enough to give rise to powerful feelings of alienation when he does otherwise.

Reflecting on akratic actions in our own lives should be enough to convince us of this point. I have often found myself doing things that I believe to be wrong even knowing that I might be psychologically unable to will myself to do otherwise. Avoiding all animal products at every meal is one example among many. In these situations, the weakness of my will never seems to provide anything like a satisfying excuse. Indeed, reminding myself of my weak will is likely to make me feel worse rather than better! This would be true even if, like

[5] This argument would seem to contradict Railton's earlier admission that actions like Juan's are in fact wrong by even the sophisticated consequentialist's lights, but dropping that admission would, if anything, strengthen the case for the indirect response.

Juan, I had had a very good reason to put myself in a position where I would be psychologically unable to do the right thing.

But perhaps no such ought-implies-can argument is needed. Perhaps—and I suspect that this is Railton's own view—reflecting on the ultimate desirability of the character traits that necessitate his acting wrongly should just straightforwardly be enough to clear Juan's conscience. "I ought to give this money to Oxfam," thinks Juan, "but it is good to have the character of a loving husband, and it is precisely that character trait that is making me spend my money on a trip to visit my wife." He stops there, satisfied, and buys the ticket without any feelings of doubt or alienation. This sort of picture has substantial prima facie plausibility, but I believe that it ultimately implies a confused notion of what it is like to act for a reason.

Here is what it's like to act for reason: You see that you can make a change in the world, the change seems to be worth making, and you make it. It might look like I'm making a bold claim here, but I don't mean to be. From Aristotle[6] to Anscombe[7] to Davidson[8] to Dancy,[9] and even perhaps to Hume,[10] I think that just

---

[6] "Every craft and every line of inquiry, and likewise every action and decision, seems to seek some good" (*Nicomachean Ethics* p. 1).

[7] "Intentional actions, then, are the ones to which the question ' Why ? ' is given application, in a special sense [...]; positively, the answer may (a) simply mention past history, (b) give an interpretation of the action, or (c) mention something future. In cases (b) and (c) the answer is already characterized as a reason for acting [...] and in case (a) it is an answer to that question if the ideas of good or harm are involved in its meaning as an answer" (*Intention* p. 24).

[8] "A reason rationalizes an action only if it leads us to see something the agent saw, or thought he saw, in his action-some feature, consequence, or aspect of the action the

about every theory of motivation that I've come across can, with some finagling, be made to fit this general model.[11]

If Juan were to take himself to be acting for a reason in the way that I describe, he would not find much comfort in the knowledge that the character traits which determine his action are good ones to have, because his character does not directly provide him with his reasons for action; he would, after all, still be acting for a bad (or at least insufficiently good) reason! And I believe that that is exactly what is happening in Juan's case; he is acting for a reason that he judges to be a bad one, or at least an insufficiently good one. There is no impossibility in acting for a reason that one sees as bad (or insufficiently good). I

---

agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agreeable" ("Actions, Reasons, and Causes" p. 685).

[9] "[N]ormally there will be, for each action, the reasons in the light of which the agent did that action, which we can think of as what persuaded him to do it" (*Practical Reality* p. 1).

[10] "'Tis obvious, that when we have the prospect of pain or pleasure from any object, we feel a consequent emotion of aversion or propensity, and are carry'd to avoid or embrace what will give us this uneasiness or satisfaction" (*A Treatise of Human Nature* 2.2.3.3, p. 266). Of course, Hume thinks that it is passion, and not reason, that ultimately moves us. But I intend the word "seems" in its broadest possible sense, so that the fact that the prospect of a pain or pleasure excites aversion or propensity might count as a *way* of seeming worth realizing or avoiding. If this is unconvincing, see footnote 11 below.

[11] Perhaps I am being too bold here. Perhaps some passions-first Humeans would reject even this cautious formulation, even with its (intentionally broad) appeal to "seemings." If so, I can think of two responses. First: Isn't there still something different and strange about making weighty decisions based on the desirability of the *character traits* that are pushing one around—something that one ought to be able to account for in one's theory of action? Second, in case the first doesn't work: If a theory can't explain how one could act for a reason in my intentionally broad sense, well then so much the worse for that theory.

do it all the time, and I suspect that many other people do too. But it is inherently alienating.

Still, Railton seems to think that Juan should be immediately placated by the knowledge that his action stems from a desirable character trait. I do not see how this can be true if Juan takes himself to be acting for a reason; isn't there something disquieting in the thought that the reasons on which one acts are no good? For Juan to be immediately placated, his psychology must be very strange indeed. Rather than acting for a reason—even a bad one—he must be doing something else entirely. Perhaps he looks inside himself, picks out the psychological and characterological traits that he judges to be good, and follows their commands blindly wherever they lead. Perhaps he doesn't even bother to look inside, and unreflectively follows the nudgings of his drives. His consequentialist beliefs are thus utterly abstract to him, and except when he wheels them out to mollify any potential guilt, they have nothing whatsoever to do with his practical life. He believes that consequentialism describes the reasons we have to act, but he never acts *for* those reasons, and is apparently unbothered by this fact! For Juan, the moral facts and practical life are entirely divorced. Juan's is a bizarre sort of human agency—one that might not deserve the name "agency" at all.

I do not of course mean to suggest that we can never take the source or nature of some inclination as a reason for action. If I believe that jealousy is

unhealthy and tends to mislead, I might take the fact that jealousy inclines me towards some action as a reason not to perform the action, either because I hope to reform my jealous nature or because I suspect that the action is likely to end up having been a bad one for reasons that I do not currently see. (I might have this suspicion, for example, because I know that when I am jealous I tend to jump to incorrect conclusions about other people's motives.) There is nothing wrong, alienating, or irrational about this form of reasoning. But it is not what is happening in Juan's case. As the case is described, he is not taking his character as a loving husband to provide (or even indicate the presence of) a *reason* to visit his wife. He is not trying to reinforce his character by visiting his wife, and he does not believe that his character is, in this case, a good guide to right action. Juan treats his character not as a reason but as a tool with which to convince himself that acting wrongly and for a bad reason is unproblematic. It is not and cannot be up to the task.

My aim in this section has been to show that the indirect strategy cannot make consequentialism non-alienating, but as usual, my arguments apply equally well to any indirect approach to ECR. Though I am not sure that ECR, by itself, implies that consequentialism is true, I do believe that it implies some form of impartialism and that it pushes us in a consequentialist direction. Juan could have been a non-consequentialist ECR theorist and my argument that his indirect strategy is necessarily alienating would apply straightforwardly. In fact, the

arguments would apply even if ECR implied an ethical theory entirely unlike Juan's consequentialism. This is because the indirect strategy fails in the consequentialist case not because of any particular features of consequentialism but because the strategy attempts to allow us to feel justified or excused in taking actions that are are contrary to what we believe we have decisive reason to do. As I have argued, this is an impossible task.

Before moving on to the direct and hybrid responses, I want to emphasize that all I take myself to have shown in this section is that the indirect approach, whether to consequentialism or ECR, is alienating. This claim does not imply that we should avoid indirect approaches. For one thing, as I'll continue to argue, *all* available responses to ECR are necessarily alienating. But even if the indirect approach turned out to be the *most* alienating alternative (which I suspect that it may be), that wouldn't mean that we should not take it. Alienation can be painful and disconcerting, but it is not the worst thing that can happen to a person. I would much rather live a rich, involved, and alienated life than a stunted one in good faith—to say nothing of my effect on the people around me.

Thus we should reject the indirect approach as a cure-all, but we should keep it in mind when we turn our attention to the difficult practical problem of what to do in the face of ECR or any of our other alienating convictions.

Scarre's Direct Response

Indirect approaches recommend that we should have no worries about acting in ways that are not warranted by the actual reasons at play so long as those actions are necessitated by or in line with a desirable sort of character, personality, motivational profile, etc. The reasoning is that attempting to always act and care in the most warranted way leads to alienation, stunted character and, ultimately, bad action down the line. I have argued that acting contrary to what we know that the reasons recommend (likewise caring or valuing things which we know do not to matter) is inevitably alienating. In the face of this gloomy conclusion, we may stop and wonder whether Railton et al are too quick in recommending an indirect strategy—is it really so hard to do and care in the right way for the right reasons?

The most optimistic direct strategy that I have seen is Scarre's.[12] On his view, it is a mistake to believe that utilitarian ends are too distant to care about and promote without alienation. He believes that the mistake rests in the tendency of philosophers to conceive of "impersonal utility" as something above and beyond and, most importantly, disconnected from the individual moments of happiness and suffering that it comprises. If that were what utility were like, it might be impossible to genuinely care about; but of course it isn't like that. All of those individual moments of happiness and suffering are constitutive parts of

[12] Given in his *Utilitarianism*, ch. 8.

impersonal utility, and they are paradigmatically things of a sort that we can care about and promote or prevent in good faith.

So far as all that goes, Scarre is undoubtedly right; everything that goes into impersonal utility is, in principle, something that anyone, including the utilitarian, should be able to care about in good faith. This simple point is a welcome corrective to views that cast anything approaching utilitarian practical reasoning as psychologically impossible and seem to hold that the best thing to do in the face of the truth of utilitarianism is to plug one's ears and never think about the matter again.[13] But he goes much too far, I think, in declaring a way out of the alienation problem.

On Scarre's reckoning, maximizing impersonal utility will require one to sacrifice their ordinary ends rarely enough that direct utilitarianism is available to all but the "morally lazy."[14] I see two problems with this line of argument.

First, and most obvious, is the apparent fact that promoting impersonal utility and promoting one's ordinary ends come apart more frequently than Scarre seems to suppose. It is true enough that we (often) know what's good for ourselves and our friends better than we do for strangers and that we (often) are in a position to provide those things more effectively or with less effort than we

[13] I've had the privilege of looking at an unpublished paper by my colleague Andrew Ingram which argues that approaches like Railton's undermine the revolutionary weight that a utilitarian ethics should carry in a world containing so much suffering and inequality. I agree with his view wholeheartedly.
[14] *Utilitarianism* p. 203

could provide them to strangers. In a world of equal wealth but diverging taste, everyone would be better off buying their own sandwiches. But when my ten dollars could feed another person for a week, it's not so obvious that I'd be maximizing utility by spending it on a roast pork Italian. Of course, as I noted earlier, there's good reason to believe that constant self-sacrifice is suboptimal in the long term for almost everyone, so maybe I really should buy myself a sandwich every now and then so as to stay happy and productive. If this is so, then utilitarianism seems like it should at least be *less* alienating than is often supposed. This brings me to my second line of argument against Scarre's direct strategy.

Often, promoting one's own ends coincides with promoting interpersonal utility only because of human weakness in one form or another; decision fatigue, hunger, stress, tiredness, and so on all sap our willpower and render us less able to act well. Recall Railton's example of Juan. As we originally imagined the case, Juan's decision to visit his wife really was wrong by consequentialist lights, but we can just as easily imagine that it was the right thing to do. Perhaps this missed trip would be the nail in the coffin of Juan's relationship, the resulting divorce would send him spiraling into alcoholism, and so on. In this case, taking the trip would actually be the best thing he could do; giving the money to Oxfam would, by consequentialist lights, be a well-intentioned mistake. Since it is right

by consequentialist lights for Juan to visit his wife, perhaps he can do so without experiencing any guilt or alienation. But I do not think that he can.

For Juan to visit his wife as a good faith consequentialist, he must visit her for the reason that by doing so he is protecting their relationship and his mental health so that he can best promote impersonal utility in the long run. I argued against Railton's indirect view that for Juan to act in good faith he would have to somehow take the fact that some generally-good-consequence-producing feature of his psychology or character nudged him in the direction of some action as a reason to do that action, and that this is nothing like what it is like to act for a reason, especially for a consequentialist. In this new case, Juan is doing better in at least one respect, because at least when he visits his wife in order to be better able to promote impersonal utility in the future he is acting for what he takes to be a good reason. But it is an indirect, alien sort of reason in comparison to an ordinary everyday reason like "it would be good to see my wife." It is not just, as Williams would put it, that Juan's concern for impersonal utility is "one thought too many;"[15] it's that this new thought *replaces* the old one and forces Juan to act on an alien reason.

Scarre may be right that utilitarians can sometimes—even often—promote the goods of themselves and people close to them in good faith as parts of impersonal utility. But they cannot always do so. Often, utilitarians will have

[15] "Persons, Character and Morality" p. 18

to do things that one would ordinarily do out of partial care or self-interest only as *indirect means* to impersonal utility. But this is very different from what it is ordinarily like to act out of partial care or self-interest, and it seems likely that it would be difficult or impossible to act in this way consistently without slowly eroding one's relationships and mental health, thus undermining impersonal utility in the long run.

On top of these two objections is a more fundamental one. Scarre's arguments purport to show that the "impersonal" part of consequentialism should not bother us. As such, to the extent that they are effective, they are most effective as defenses of consequentialisms like Moore's that only come drastically apart from commonsense morality because they are impartial in that they focus on the impersonal good rather than on some more particular set of ends. For the hedonistic utilitarian or the ECR theorist, common sense needs to be revised not just by making it impartial but by eliminating entire swaths of purported sources of value—lives as wholes, desert, etc. Perhaps I can value my own happiness or the happiness of my friend as a part of impersonal utility. But if there is no value in desert (contra Moore but pace the hedonist utilitarian or ECR theorist) then we could not reward or punish people who have helped or wronged us as a part of "impersonal desert" as a hypothetical Moorean-Scarreian might argue.

Thus, though Scarre's sophisticated direct view effectively narrows the effective scope of the traditional arguments against naive consequentialism, it

can't allow us to completely avoid the dilemma that motivates Railton's indirect strategy, especially if we are hedonist utilitarians or ECR theorists; either we act in good faith and undermine our own aims or we move to an indirect response and face the alienation such a response entails.

Siderits's Hybrid Approach – Ironic Engagement

Siderits intends his hybrid approach, which he calls "ironic engagement," to combine the best features of the direct and indirect approaches while avoiding their difficulties. Unlike Railton and Scarre, Siderits is concerned primarily with his own brand of Reductionism about personal identity rather than with consequentialism. By "Reductionism," which I will follow him in capitalizing when discussing his view, Siderits intends something more than I have meant in this dissertation by "reductionism." On Siderits's Buddhist Reductionism, persons are "conventionally" real but "ultimately" unreal.[16] The claim that persons are "ultimately" unreal is, as I understand it, similar to Parfit's claim that there is no deep further fact about personal identity, which I take to be the central claim of reductionism. To be "conventionally real" in Siderits's sense is just, I think, to be accounted for in a conventional conceptual framework or ontology. Importantly, as I understand Siderits's position, he is not so concerned with what our conventions *are* but about what they *should* be.

[16] *PIBP* p. 22

So, on my reading, Reductionism is the combination of a metaphysical view—reductionism—with the practical claim that we should keep persons in our conceptual framework and person-regarding attitudes in use. If we accept the metaphysical view but reject the practical claim, then we are in Siderits's terms Eliminativists.

It is tempting to identify Siderits's Reductionism with Parfit's Moderate Claim and his Eliminativism with Parfit's Extreme Claim, as Siderits himself comes close to doing.[17] I believe that this is a mistake. For Siderits, person-regarding attitudes are not justified because they are "close to" the ultimate truth or anything like that. Person-regarding attitudes are justified only indirectly, generally on the basis of the utility of adopting person-regarding attitudes. For example, he writes:

> I identify with and care about my future states because my having learned to do so better insures that there will be fewer pains among them. I identify with my present preferences and projects because these should be

[17] At a minimum, he thinks that Reductionism *implies* the Moderate Claim: "If the Extreme Claim were true, there would turn out to be no middle ground between Non-Reductionism and Eliminativism: either persons are ultimately real, or else all our person-regarding attitudes are rationally unjusitifiable. Thus a Reductionist must deny the Extreme Claim and hold instead a Moderate Claim, to the effect that if Reductionism is true, then mitigated forms of the four features [of personhood—interest in one's own survival, egoistic concern for one's future states, holding persons responsible for their past deeds, and compensation for one's past burdens—]may be grounded in facts about the impersonal entities and events that persons just consist in" (*PIPB* p. 72). His "four features" are borrowed from Marya Schechtman.

> seen as resulting from a process of self-revision that likewise better promotes maximization. I identify with my past pains because doing so facilitates appropriation of my present properties, which is necessary if self-revision is to be ongoing.[18]

There is nothing here with which the ECR theorist needs to disagree. On Parfit's original explicit usage, the Extreme Claim says that we have no reasons for partial care for our own futures. On my extended reading of the claim, it says that nothing else matters in *any* of the ways that the deep further fact of identity is ordinarily taken to matter. As I understand both of these versions of the Extreme Claim, they are views about standard, direct, object-given reasons and values.[19] There is no reason I can see that the ECR theorist could not adopt (or, better, try to cause or allow themselves to adopt or maintain) person-regarding attitudes for exactly the reasons that Siderits says we should. Thus my ECR theorist is *not* committed to full Sideritsian Eliminativism, and can happily be a Reductionist. Since the ECR theorist can be a Reductionist, they can apply the

---

[18] *PIBP* p. 78

[19] I don't think I depart from Parfit here. At the time of *Reasons and Persons*, he does not (so far as I'm aware) have the language of standard, non-standard, object-given, and state-given reasons. But I suspect that his analysis of the ECR theorist would be exactly parallel to his analysis of the sophisticated egoists and consequentialists he discusses the first part of the book. These sophisticated egoists and consequentialists, like Railton's Juan, nourish certain sorts of character for egoist (or consequentialist) reasons. Just as we could nourish certain sorts of character for these reasons, we could nourish certain person-regarding attitudes.

ironic engagement strategy without revision.

Siderits explains his strategy by way of an analogy to civic pride. He imagines himself as an "urbanist"—a reductionist about cities—who recognizes the utility of adopting and maintaining city-regarding attitudes and civic pride while also denying the ultimate reality of cities and recognizing that he might just as well have been born elsewhere. Since his reasons for feeling pride in his city have absolutely nothing to do with the city itself, he imagines himself feeling alienated; just like in the personal identity case, the worry is that "having a life is not the sort of thing one can choose as a means to further some separate end."[20]

The "ironic engagement" that Siderits recommends to his imagined sophisticated urbanist looks something like Pascalian habituation:

> I should reflect on what it is about this place that I particularly enjoy and appreciate, and begin to dwell on these features. Then I should share the fruits of these reflections with my neighbors, some of whom will no doubt respond with their own suggestions of valuable features to add to my list [...] All of this is, I think, perfectly consistent with my urbanism. And it is hard to see how the feeling I come to have in the end is not the genuine article.[21]

[20] *PIBP* p. 133
[21] *PIBP* p. 137

In spite of successfully cultivating these feelings, Siderits imagines himself holding on to his knowledge of the truth of urbanism. This knowledge allows him to take an "ironic distance" from his feelings of civic pride, but apparently does not induce alienation.

Extending the example to the case of personal identity is straightforward enough: Rather than cultivating civic pride by reflecting on the best aspects of our hometown, we cultivate care for personal projects, friendships, and so on by working at them diligently and reflecting frequently on what we value about them. Eventually, "the activity begins to take on a life of its own, so that what was initially valued only extrinsically now has intrinsic value for me."[22]

Siderits anticipates what seems like the correct objection to make at this point: The fact that feelings of civic pride and a belief in urbanism can coexist does not mean that they are comfortably compatible:[23] "Pride," says his imagined interlocutor, "involves the sense that one is somehow ennobled through one's relation to the thing in which one takes pride. How can the urbanist rationally maintain that they derive value from their relation to a fiction?" He responds that "to take pride in something is to be disposed to do such things as praising it, defending it against its detractors, seeking to correct its flaws, and the like. This

[22] *PIBP* p. 137

[23] He says "logically compatible," which I think may mean something like "rationally compatible," though I'm not sure (*PIBP* p. 137). Regardless of how the "logically" should be read, what ought to matter is whether or not a belief in urbanism can exist alongside civic pride without engendering alienation, which is why I say "comfortably" instead.

is what feeling proud—feeling ennobled by one's relation to the object—ultimately amounts to."[24]

This response seems insufficient. Pride, if it does not include a cognitive component as a constitutive part, at least requires an appropriate cognitive *accompaniment* if we are to be able to believe it warranted. It may be possible to feel genuine civic pride if one does not believe, as by hypothesis the urbanist does not, that one is "ennobled through one's connection" to one's city, just as it may be possible to feel genuine fear in response to something—a movie, a flight, a clown—that one does not believe is actually dangerous. But without those beliefs, or with contrary beliefs, reflection will show us that our pride or fear is unwarranted and irrational.[25]

That is, even if Siderits is right and all that pride—and likewise friendship, love, self-concern, feelings of responsibility, first-personal anticipation, and so on—amount to are cognitively neutral dispositions, that is not enough to prove his point. The mere coexistence—even the stable coexistence—of an attitude together with beliefs that render that attitude unwarranted by one's own lights is not enough to free a person from alienation. It is in large part *because* a person believes their attitude to be unwarranted by their own lights that they experience alienation!

---

[24] *PIBP* p. 137

[25] The scary movie case might be different from the civic pride, flying, and clown cases, since we might want to deny that our reaction to the scary movie is really fear or that if it is fear it is unwarranted. But no matter.

It might be objected that I am requiring Siderits to succeed on my terms rather than his own. Again, Siderits characterizes the alienation objection as the claim that "having a life is not the sort of thing one can choose as a means to further some separate end,"[26] and if someone comes to have the life, cares, and projects they do through cultivation and habituation and do not attempt any extrinsic or instrumental justification, it might be that they are not alienated in precisely this sense. But while it is true that the ironic engagement strategy, unlike Railton's indirect strategy, does not force us to have or justify the cares that we do by appeal to their indirect consequences, it seems to do so only by imagining that we do not need to have or justify them for any reason at all!

Imagine an urbanist who asks themselves what reason they have for taking pride in their city. If the urbanist is a Railtonian, they might say that they are proud of their city because having such pride will have good consequences. As we have seen, this is the wrong kind of answer. The Sideritsian urbanist will not make this mistake. If they do not simply remain silent, they will say that they are proud of their city because of its parks, its sandwiches, its barbecue, its culture. This answer is in one way an improvement, because these at least *look* like reasons of the right kind. But because they are an urbanist,[27] they will recognize that they are not, in the end, reasons at all. In one way, it is better to have an attitude for which one knows one cannot claim warrant than to have an

[26] *PIBP* p. 133
[27] Or an "Extreme Claim urbanist," perhaps.

attitude which one mistakenly believes can somehow be indirectly warranted because of its consequences. But neither situation is good, and both, given honest reflection, have the potential to alienate.

It might also be objected that I am, as Siderits cautions against, committing the genetic fallacy,[28] arguing that the fact that the ironically engaged reductionist has the cares and projects they do because of habituations renders these cares and projects irrational or alien. But the genesis of the reductionist's cares and projects is not the reason that their cares and projects are irrational and alien; the two have nothing to do with one another. In fact, in my view, knowledge that one came to have certain cares and projects because of cultivation and habituation should not by itself even be enough to induce the mild "ironic distance" that Siderits suggests it will.[29] This is because the genesis of cares and projects provides them with a *causal* explanation, but not a *justifying* explanation. For the ECR theorist, there is no explanation that can do the work of justifying their person-regarding cares and projects in the appropriate way.

An analogy: I was chronically ill for the majority of my teenage years. During that time, because I lacked the energy to do much else, I watched a lot of movies, and I ended up watching some of my favorites probably dozens of times. I still love most of these movies, even if I haven't seen them for over a decade. I know that the *cause* of my extreme fondness is probably that I spent so much

[28] *PIBP* p. 137
[29] *PIBP* p. 137

time enjoying them when I was younger. But that has nothing to do with whether my love is *justified*. My justification depends on the *reasons for which* I love these movies, if I have any—things like characters or or shots or scenes or melodies or performances. Of course, as I have grown, my tastes have changed, and I might decide upon a re-watch that the various features that I once thought made some movie good in fact do the opposite. In *this* case, I would decide that my love of the movie was unwarranted, and if I kept loving the movie in spite of myself I might do so with "ironic distance."

It is the same in the case of personal identity. It is not the fact that I have developed my cares and projects through years of habituation that makes my cares and projects unwarranted. It is the much more straightforward fact that the objects of my care do not warrant care and my projects do not warrant being pursued.

Ironic Engagement and Kayfabe

In spite of my criticisms of Siderits's ironic engagement strategy, I view it as a practical improvement on the indirect and direct strategies that it is meant to replace. Unlike the indirect strategist, the ironic engagement strategist understands that merely recognizing the utility of a belief or the desirability of a motivation is not (or should not be) enough to allow a person to form the belief or act on the motivation in good faith. Unlike the direct strategist, the ironic

engagement strategist recognizes that it is not always desirable or even possible to bear the full import of the normative facts in mind. But like the indirect and direct strategists, the ironic engagement strategist believes that there is a way out of alienation. In this they are mistaken.

As a practical matter, the biggest advance that Siderits makes over Railton and Scarre is that he recognizes that maintaining person-regarding attitudes and projects alongside a belief in Reductionism is an ongoing Pascalian project. This is undoubtedly so. But he goes wrong in thinking that it is the sort of project that a single, unified, unalienated agent can undertake. Maintaining person-regarding attitudes and believing in Reductionism—or at least ECR—isn't like walking and chewing gum. It's like chewing sugary gum while looking at your insulin injector. It's like walking on a broken hip. A more complete strategy needs to acknowledge this necessary fracturing and say something about how to deal with it. My own kayfabe strategy, outlined at the end of Chapter Three, is an attempt to do just that.

Works Cited

Anscombe, G.E.M. *Intention*. 2nd ed., Cambridge, Harvard UP, 1957/1963.

Aristotle. *Nicomachean Ethics*. Trans. Terence Irwin, 2nd ed., Indianapolis, Hackett, 1999.

Barthes, Roland. "The World of Wrestling." *Mythologies*. Trans. Annette Lavers, New York, The Noonday Press, 1972, pp. 13-23.

Bratman, Michael. "Practical Reasoning and Weakness of the Will." *Nous*, vol. 13, no. 2, 1979, pp. 153-171.

Butler, Joseph. "Of Personal Identity." *The Words of Bishop Butler*. Ed. David White, Rochester, University of Rochester Press, 2006.

Carroll, Noël. *The Philosophy of Horror or Paradoxes of the Heart*. New York, Routledge, 1990.

Chisholm, Roderick M. "Firth and the Ethics of Belief." *Philosophy and Phenomenological Research* vol. 51, no. 1, 1991, pp. 119-128.

Conway, John H, and Richard K. Guy. *The Book of Numbers*. New York, Copernicus, 1996.

Dancy, Jonathan. *Practical Reality*. Oxford, Oxford UP, 2000.

D'Arms, Justin and Daniel Jacobson. "Sentiment and Value." *Ethics* vol. 110, no. 4, 2000, pp. 722-748.

Davidson, Donald. "Actions, Reasons, and Causes." *The Journal of Philosophy* vol.

60, no. 23, 1963, pp. 685-700.

— "How is Weakness of the Will Possible?" *Essays on Actions and Events* 2nd ed. Oxford, Oxford UP, 2001, pp. 21-42.

Feldman, Richard. "The Ethics of Belief." *Philosophy and Phenomenological Research* vol. 60, no. 3, 2000, pp. 667-695.

Hare, Richard Mervyn. *Moral Thinking: Its Levels, Method, and Point*. Oxford, Clarendon Press, 1981.

Hieronymi, Pamela. "The Wrong Kind of Reason." *The Journal of Philosophy* vol. 102, no. 9, 2005, pp. 437-457.

Hume, David. *A Treatise of Human Nature*. Ed. David Fate Norton and Mary J. Norton. Oxford, Oxford UP, 2000.

Kavka, Gregory. "The Toxin Puzzle." *Analysis* vol. 43, no. 1, 1983, pp. 33-36.

Korsgaard, Christine. *The Sources of Normativity*. Cambridge: Cambridge UP, 1996

Lewis, David. "How Many Lives has Schrödinger's Cat." *Australasian Journal of Philosophy*, vol. 82, no. 1, 2004, pp. 3-22.

— "Survival and Identity." *Philosophical Papers Vol. 1*. New York, Oxford UP, 1983, pp. 55-77.

Locke, John. *An Essay Concerning Human Understanding*. Ed. Peter Niddich. Oxford, Oxford UP, 1975.

Nagel, Thomas. "The Absurd." *The Journal of Philosophy*, vol. 68, no. 2, 1971, pp. 716-727

Nozick, Robert. "Newcomb's Problem and Two Principles of Choice." *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of his Sixty-fifth Birthday*. Ed. Nicholas Rescher et al. Dordrecht, D. Reidel, 1969, pp. 114-146.

Olson, Jonas. "Buck-Passing and the Wrong Kind of Reasons." *Philosophical Quarterly* vol. 54, no. 215, 2004, 295-300.

Parfit, Derek. "Experiences, Subjects, and Conceptual Schemes." *Philosophical Topics* vol. 26, no. 1-2, 1999, pp. 217-270.

— "Is Personal Identity What Matters?" The Mark Sanders Foundation, 2007.

— *On What Matters* vols. 1-2. New York, Oxford UP, 2011.

— "Personal Identity." *The Philosophical Review*, vol. 80, no. 1, 1971, pp. 3-27

— *Reasons and Persons*. New York, Oxford UP, 1987.

Pascal, Blaise. *Pensées*. *Pensées and Other Writings*. Trans. Honor Levi, ed. Anthony Levi. New York, Oxford UP, 1995.

Putnam, Hilary. "The Meaning of 'Meaning'." *Mind, Language and Reality*. Cambridge, Cambridge UP, 1975, pp. 215-271.

Rabinowicz, Wlodek and Toni Rønnow−Rasmussen. "The Strike of the Demon: On Fitting Pro-attitudes and Value." *Ethics* vol. 114, no. 3, 2004, pp. 391-423.

Railton, Peter. "Alienation, Consequentialism, and the Demands of Morality."

Philosophy and Public Affairs vol. 13, no. 2, 1984, pp. 134-171.

Raz, Joseph. "Reasons: Practical and Adaptive." *From Normativity to Responsibility*. New York: Oxford UP, 2011, pp.36-58.

Ross, William David. *The Right and the Good*. Ed. Philip Stratton-Lake. New York, Oxford UP, 2002.

Russell, Bertrand. *The Analysis of Mind*. London, George Allen & Unwin, 1921.

Scanlon, Thomas. *What We Owe to Each Other*. Cambridge, Belknap, 1998.

Scarre, Geoffrey. *Utilitarianism*. London, Routledge, 1996.

Schechtman, Marya. *The Constitution of Selves*. Ithica, Cornell, 1996.

Siderits, Mark. *Personal Identity and Buddhist Philosophy: Empty Persons* 2nd ed. Burlington, Ashgate, 2015.

Singer, Peter. "Famine, Affluence, and Morality." *Philosophy and Public Affairs*, vol. 1, no. 3, 1972, 229-243.

Stokes, Patrick. "Will it be me? Identity, concern and perspective." *Canadian Journal of Philosophy*, vol. 43, no. 2, 2013, pp. 206-226.

Strawson, Galen. "'The Secrets of All Hearts': Locke on Personal Identity." (manuscript)

Wachsberg, Milton. *Personal Identity, The Nature of Persons, and Ethical Theory.* Diss: Princeton, 1984

Walton, Kendall. "Fearing Fictions." *Journal of Philosophy*, vol. 75, no. 1, 1978, pp. 5-27

Williams, Bernard. "A Critique of Utilitarianism." *Utilitarianism For and Against*. New York, Cambridge UP, 1973.

— "Persons, Character, and Morality." *Moral Luck: Philosophical Papers 1973-1980*. Cambridge, Cambridge UP, 1981.

Wrenn, Chase B. "Why There Are No Epistemic Duties." *Dialogue*, vol. 46, no. 1, 2007, pp. 115–136.