Published In: *Oxford Studies in Normative Ethics Volume 9*. Edited by Mark Timmons, Oxford University Press (2019).

Well-Being, the Self, and Radical Change

Jennifer Hawkins

It is a well-known fact that people change over time. Sometimes the change is gradual. Sometimes it is swift. And some changes are deeper than others, in the sense that they alter more significant features of the person. In this essay I explore cases of "radical change." I define radical change as change where either (1) several of a person's core values change, or (2) some deeper feature of her psychology changes (which will typically also result in value changes). An example of a deeper psychological change might be a change in personality—e.g. a change in one of the big five personality traits discussed by psychologists. Or it might be some other even deeper feature of the individual that changes. As should be clear, radical change as I define it, is a degree concept. Within the category of radical change there are more or less radical changes.

I take it that most *really* radical change is bad for the person who undergoes it. Consider Phineas Gage, the nineteenth century railroad worker.[1] An explosion sent a tamping iron right through Gage's skull, yet miraculously he survived. However, he was dramatically changed. His fundamental personality traits were altered forever by the alterations to his brain, and his life went poorly from then on as a result. But there are also clearly cases in which radical change can be prudentially good. So what interests me is the question: What explains why some radical changes are prudentially good and others prudentially bad?[2] It is important to emphasize that I am discussing *prudential* change, the kind of change that

[1] The story of Phineas Gage is related in the first chapter of Antonio R. Damasio, *Descartes' Error: Emotion, Reason and the Human Brain.* (New York: Harper Collins Publishers, 1994). My comment simply relies on the reader's probable familiarity with Damasio's way of telling the story, which has been influential. However, I thank an anonymous reviewer for drawing my attention to the fact that the details of the extent to which Gage changed remain disputed. See e.g. Griggs, R. (2015), "Coverage of the Phineas Gage Story in Introductory Psychology Textbooks: Was Gage No Longer Gage?" *Teaching of Psychology*, 42 (3), 195-202.

[2] This is a metaphysical question, not to be confused with the epistemic question of how we can come to know which changes are good and which bad.

improves (or lowers) the welfare of the individual who changes. I set aside for now admittedly complicated questions about radical change and morality.[3]

One might suppose there is little point to my question because we so seldom have control over radical change. Certainly Gage could neither anticipate nor control what happened to him. However, some types of radical change do fall within our control, and in the future, with the development of new technologies, more forms of radical change may become possible. Even now, it is often true of us that we *could* change ourselves in particular ways *if* we chose to. But while many of us engage in smaller scale projects of self-change, few of us seek to *radically* change ourselves because it rarely occurs to us that a radical change would be best. However, that *may* simply be an unfortunate feature of our limited personal perspectives.

My own thinking about such questions has been deeply influenced by a particular picture of prudential facts I find attractive. I shall call this view the "future-based reasons view" or FBR. This is a very general metaphysical view about what makes it the case that such and such would be a good choice (prudentially) for someone. It has two components. First, whether a *choice* is prudentially good for an individual depends on the good (or bad) things that various possible futures hold for that individual. If choosing x now will lead to the greatest net future welfare for me, then I have the strongest prudential reason now to choose x. Second, these normative facts about future welfare are not in any way dependent on the subject's attitudes or desires either in the present or the past. What makes it the case that x is good for A in the future is not the fact that A now desires x, nor the fact that in the past A desired x for her future self. Nor is it the fact that a fully informed A would desire it for her lesser informed self. Instead, what determines facts about future good are either facts about A as she is in the future or facts about the world A will inhabit at that time, or most plausibly, the interaction between both. In other words,

[3] It is, of course, possible that a person might change in ways that make her more or less moral. Moreover, a morally good change might not be prudentially good, and a prudentially good change might not be morally good. But the question of how to resolve conflict between prudential reasons and moral reasons is a huge topic far beyond the scope of this paper. When I say a person has most prudential reason to do x, I leave it open whether she has most reason to do x all-things-considered.

the full explanation of what makes something good for a person at a certain time depends on facts about the person and the world *at that time*.

To be clear, what I am calling the FBR view is not a theory of welfare, but rather a framework upon which one might build a theory. Various contemporary approaches to welfare are compatible with it. Both hedonism and objective list theories are. But then so would be any subjective theory that focused on current appreciation of things (as opposed to the prospective pro-attitudes desire theories typically stress). But not all theories are compatible. As should be obvious, the view I am describing rules out desire-satisfaction theories of welfare including the informed desire versions.[4] It also rules out the idea, accepted by many theorists, that part of the welfare value of a life is determined by the overall shape of that life.[5] It is beyond the scope of this essay to defend these exclusions at any length, but the framework I favor is compatible with a variety of otherwise diverse theories such that what I have to say should be of interest to many theorists of welfare.

FBR has the virtue that it easily accommodates cases of good radical change. On this account, a radical change represents the best choice (and a subject has most reason to undergo it) if it leads to the possible future with the greatest net welfare value. But by itself, without qualification, FBR also appears to have the problem that it allows too many cases of radical change to count as good. Thus it seems that a

[4] Many theorists have defended some version of desire-satisfactionism, but the most popular versions among philosophers have been informed desire theories. One of the most developed such views is that of Peter Railton. See Railton, "Facts and Values," *Philosophical Topics*, 14: (1986): 5-31; and Railton, "Moral Realism," *The Philosophical Review*, 95: 2: (1986): It is beyond the scope of this essay to explain my exclusion of desire theories, but I have explained my opposition to them elsewhere. See Hawkins, "Well-Being, Time and Dementia," *Ethics*, 124: (2014): 507-542; and Hawkins, "Internalism and Prudential Value," *Oxford Studies in Metaethics*, vol. 14, forthcoming 2019). For good general overviews of the desire satisfaction approach, see Sumner (1996); Ben Bradley, *Well-Being*, (Malden, Ma: Polity, 2015); Guy Fletcher, *The Philosophy of Well-Being: An Introduction*, (New York: Routledge, 2016).

[5] Certain theorists reject additive views (or as some say "intra-life aggregation") because they take seriously the idea that the shape of a life matters prudentially and assume the two claims are incompatible. See e.g. Michael Slote, "Goods and Lives," in *Goods and Virtues*, 9-37. (Oxford: Oxford University Press, 1984), 9-37; David Velleman, "Well-Being and Time," *Pacific Philosophical Quarterly,* 72: (1991): 48-77; Joshua Glasgow, "The Shape of a Life and the Value of Loss and Gain," *Philosophical Studies,* 162: (2013): 665-682. Dale Dorsey, "The Significance of a Life's Shape," Ethics 125: (2015): 303-330, has shown that, depending on *why* you think shape matters, taking shape seriously *may* be compatible with aggregation. Still it is clear that my approach *is* incompatible with common ways of understanding the shape of a life thesis.

defender of any version of FBR will need to seriously consider questions about radical change and what explains the difference between the good cases and the bad.

My overall aims are really quite modest. My first is simply to articulate the issues vividly, and thereby enable philosophers to feel the force of a certain kind of problem that, as I see it, has all too often been overlooked. Second, I hope to convince philosophers that some of what seem like the more obvious ways to resolve the problem—attempts that invoke one or another notion of identity—fail. No doubt many, once convinced that the simple solutions are flawed, will think that the obvious next move is to adopt a more sophisticated and nuanced version of the same general strategy. Yet I hope to make it clear that any such move must try to answer a deeper, as yet unanswered question, about the relationship between the self and welfare.

§1.0 Good Radical Change

I begin by presenting a case where it seems intuitively plausible that radical change would be prudentially good.

Consider Sharon. Sharon is a creative, artistically talented painter. Unfortunately, however, she is also predisposed to unipolar depression. Beginning in young adulthood she experiences recurrent episodes of what psychiatrists call "major depression." When she is depressed she can't do her artwork, or much of anything else. But when she emerges from a depression she can return to her art. Importantly, the depressive episodes take a toll not only on her creative work, but also on her personal relationships.

Significantly, Sharon has a certain image of herself to which, over time, she has become deeply attached. It is the description under which she now *values herself*. She values art, of course. But she also values *being an artistic person*, which she equates not simply with painting but with a certain lifestyle and with a certain degree of social nonconformity. In the past, when struggling with her depressions, she has often comforted herself with the thought that, at least in certain people,

artistic creativity and depressive mental illness appear to be somehow linked.[6] As a result she has brought herself to a point where she thinks of her depression as the price she pays for artistic creativity. Not only is she resigned to her illness, but she has even become somewhat proud of it because it seems (to her) a mark of other qualities she values. When Sharon contemplates the lives of various people she knows, she judges their lives to be (in contrast with hers) lacking in depth and meaning.

Now let us suppose a medication exists that could really help Sharon by putting a halt to the extreme emotional cycles she currently experiences. However, Sharon is skeptical about trying it. As we have seen, she has already formed a self-narrative in which depression is partly explained as a special sign of giftedness. And because of this she is now reluctant to give it up. She is also afraid that the medication might change her in ways that she can't anticipate, but which she is pretty sure she (as she is now) wouldn't like. She is both worried that it will make her "shallow" as opposed to "deep" and more concretely, that it may interfere with her artistic development. She fears she won't be creative in the same way or to the same extent if she takes medication.

These are Sharon's fears. Now let's assume they are well-founded. Let's assume that if Sharon doesn't take the medication she will continue to suffer major depressions and this will continue to limit her creative work. It will continue to make her life difficult in a variety of ways, and it will rule out certain kinds of long-term meaningful relationships. If she does take the medication, however, it will have a variety of effects on her. On the one hand, it will lift the depressions completely. However, it will simultaneously alter her personality in subtle ways, and this, in turn, will lead to her abandoning her current work. Though she will always appreciate art, for a variety of reasons, the person she would become with

[6] It is well established that poets, fiction writers, visual artists and musicians are much more likely than ordinary people to suffer from either manic-depressive illness (bi-polar disorder) or unipolar depression. For example, one study found rates of depression 8 to 10 times higher among artists and writers than in the rest of the population. See Kay Redfield Jamison, Jamison, *Touched with Fire: Manic-Depressive Illness and the Artistic Temperament,* (New York: Simon & Schuster, 1996). The exact nature of the association, however, is not understood and remains a topic of dispute and continuing inquiry.

medication would not pursue an artistic career. However, she would find other projects that would be just as fulfilling for the person she would then be as artistic pursuits currently are for the person she is now. Moreover, once she is no longer living the chaotic life of depressive ups and downs, she will be able to form and maintain loving relationships and accomplish more in her new line of work than she would accomplish in the old.

I take no stand for the moment on whether the goods of her new life are intrinsic or instrumental. Hedonists are free to see them as instrumental, and to think the medicated life will have more overall pleasure or happiness. And objective list theorists are free to imagine that her life with medication will contain more objective goods. And so on for other types of theorists. I will stipulate only one thing, which is that Sharon will appreciate the life she thereby gains if she takes the meds. She will not in retrospect view the change as a mistake, but instead as an excellent choice.

The change is radical in my defined sense. It will alter many of her values, and will do so in part by altering certain aspects of her personality (this is why she will no longer pursue an artistic career). Even though she herself has no current desire to change, and would not *now* view the possible life with medication as in any sense good for her, it is plausible that it would be good *for Sharon* to undergo this change. I suspect most readers will agree with me. If that is so, then they will also agree that at least some cases of radial change are prudentially good and we should want a theory of welfare that can accommodate that.

§2.0 Good Radical Changes & Future Based Reasons

The future-based reasons view is able to handle certain cases of radical change quite well. But only certain cases. To understand why this is so, it is helpful to have a more developed understanding of the view.

The FBR view is first and foremost a way of thinking about the relationship between good prudential choice and prudential value itself. In ordinary life when we try to make a self-interested decision (or try to see what would be the best decision

for someone else), we take it that there are various things that could be done. Associated with the different things that could be done are different outcomes that might follow. And these outcomes could lead to others and so on and so forth. Indeed, summarizing this familiar way of thinking, we can say that associated with each practical choice a person makes there are various possible futures, possible ways that the life she has been living up until then might continue. The FBR assumes that these possible futures have a certain welfare value for the subject in virtue of the various good and bad things those futures contain for her. What that value is will depend, of course, on one's preferred account of prudential value and how much value various things are thought to have.

A hedonist, for example, will want to know of each possible future how much happiness it contains, on the one hand, and how much suffering, pain, or unhappiness it contains on the other. The overall net value (negative or positive) of that possible future will depend on whether there is more happiness than suffering or vice versa. Alternatively, an objective list theorist would want to know of each possible future what kind of objective goods (friendship, achievement, knowledge, etc.) that life contains as well as what kinds of objective bads. No matter the theory of prudential value, each possible life continuation will have a positive or negative net value depending on whether the positives (as construed by that theory) outweigh the negatives (as construed by that theory), or vice versa.

One implication of this approach is that possible futures can, at least in principle, be *ranked* from best to worst. The FBR thesis assumes, as do many philosophers, that practical reasons—in this case *prudential* practical reasons—stem from facts about prudential value. For any two possible futures, x and y of an agent A, if x has greater net welfare for A than y, then A has more reason to choose x. But A has *most* reason, i.e. the *strongest* prudential reason, to choose the best possible future or the highest ranked one, whichever one that is.

Of course, this is a theory of what reasons we have, and not a theory of what reasons we can easily know about or perceive. It may be that we often simply don't know about certain options, or don't think of them, and so fail to identify the best ones. That is an epistemic problem, and a serious one. But then it seems right to me

that it should turn out to be extremely difficult—perhaps even impossible—to know what the *best* path through life is. We do reasonably well when we are able to recognize the more obvious paths forward, assess their relative strengths and choose one of the better ones. Finally, it is because of these epistemic barriers that normative reasons and motivating reasons so frequently come apart.

It is important to see that not all good choices lead to what we would ordinarily call good outcomes. One kind of case where this is so involves bad lives. Sadly, some lives are so filled with suffering or other negatives that the negatives outweigh any positives the life might contain. I shall refer to all such lives where net value is negative as bad *simpliciter*. But of course, even within the category of lives that are bad *simpliciter*, there can lives that are better or worse than one another. Now suppose someone faces a grim choice between two lives each of which is bad *simpliciter* in virtue of containing great suffering. Still, one life has less suffering than the other. If these are the only options, then it still makes sense to choose the less bad life, even though the life chosen would not count as good as people ordinarily think of "good.". This is a case where the best choice does not lead to a good life.

This idea that there is a point at which the negatives begin to outweigh the positives (however construed) suggests for many a way to draw a line between a life worth living (or a continuation that is worth living) and one that is not. And some go further and suggest that when the only options available are lives not worth living, then death is preferable. Of course death is not a "possible future" but rather the absence of a future. Still, it is fairly easy to adjust the comparison of futures to make it a comparison of lives that are the same up to a certain point. If a life that ends at that point is better than any of the possible lives where it continues, then the FBR can allow that in such a case death would be better for the individual in question.[7] However, even when death is the better option, it is not always a real, practical

[7] This is just the familiar thesis (known as the deprivation view) that death is a comparative harm and that the proper unit of comparison is whole lives. For detailed introduction to such views and their problems see Steven Luper "Death", *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2016/entries/death/>.

option, in which case it still makes sense, insofar as there are choices, to choose the future that is least bad.

Let us now consider what the FBR says about radical change. As we saw at the outset, some radical changes are not chosen, but simply happen to us. But when choice is involved and we can either change radically or not, then the important question will be about the overall welfare value of various possible futures with the change and without it. In some cases FBR will not rank highly any possible futures involving radical change. But sometimes it will. It can certainly occur that the best possible future for an individual is one that includes a radical change. If that's so, the individual in question will have most prudential reason to undergo the change.

Sharon's case is like this. The best possible futures for her are those where she takes the meds. She will be much happier in the future with meds, and even non-hedonists typically grant significant prudential value to happiness. She will also have more success in her endeavors, even though those endeavors will be different in her new life than in her old. And she will have loving, lasting relationships. These three features will be absent from the non-med life, which will also contain much suffering from depression. If these are the only options, then it is clear that in welfare terms Sharon has most reason to take the medications.

Here FBR seems to give the right verdict. But now consider a very different type of case, where the FBR view also recommends change. Consider the case of Chloe who is offered the following opportunity, which is presented to her (by the eager scientist who developed it) as "the magnificent alteration."[8] The purpose of the procedure is to optimize future welfare prospects. However, the change required to do so is very radical. A person who undergoes this procedure has her mind wiped clean, thereby losing all her beliefs, memories, values, etc. When she

[8] This example was inspired by Jeff McMahan's example "The Cure." Jeff McMahan, *The Ethics of Killing* (Oxford: Oxford University Press, 2000), 77. In McMahan's case an individual must choose between certain death or continued life after going through an equally radical change. McMahan's purposes are different from mine. He aims to illustrate the fact that many people *care less* about a future self that is radically different from their current self. He goes on to use these facts about care to support his theory of time relative interests. However, even if he is right about most people's feelings, I doubt these feelings have the normative significance he assigns them. In an unpublished companion essay to the present one I argue against McMahan's theory of time-relative interests.

awakes she will not remember her past life. However, the changes go beyond alterations of mental content. The procedure will also alter her natural psychological dispositions in ways that "optimize" her personality, giving her just the right degree of (for example) extroversion, or conscientiousness. This will be the precise degree of extroversion (or conscientiousness) that most often leads to happiness (or success as defined by other theories of prudential value). Other traits will be similarly fine-tuned. After the procedure is over, she will have no real reason to try to re-create her old life because, given the deeper psychological changes she has undergone, that life will no longer fit her.

Chloe, understandably, is appalled by this description and wonders why this crazy man thinks she would want it. But his answer is that it creates for her a possible future with much greater overall well-being. As she is now, her life is going fairly well, but there are still some definite, long-standing problems. The person she is currently faces a lifetime welfare ceiling, one she could surpass if she changed. Currently, her values and her personality sometimes undermine her ability to achieve her goals. Studies of people who have previously undergone the magnificent alteration show that it works wonders. A fresh start with a different psychological profile will enable Chloe to have a life much higher in overall welfare value than any life open to her as she is.

The example is deliberately extreme. Yet it is not crazy to suppose that for some people really radical change would enable them to live much better lives from the standpoint of welfare. In such a case, the best possible prudential future is the one that includes radical change and so FBR says that the individual has the strongest prudential reason to undergo the change.[9] And yet, most people, including myself, have the intuition that it couldn't be prudentially *best* to undergo such a change. Not only does Chloe not have *most* reason to choose it, she may have *little* or *no* reason to given that her alternative, while not perfect, is still a perfectly good life

[9] McMahan says of his example, "The Cure," that "most of us would at least be skeptical of the wisdom of taking the treatment and many would be deeply opposed to it" (2000, 77). However, FBR will clearly say that one ought to take the cure, because the only options are radical change or death, and it is clear that the life after radical change is well worth living, i.e. it has net positive welfare value for the subject.

overall. How do we explain this in a way that preserves our original intuition about Sharon, namely, that sometimes radical change can be the best option, the one we have most reason to choose?

§3.0 Why Identity is Not the Solution: Part I. Numerical Identity

Our question then is why the magnificent alteration is not the prudentially best choice for Chloe. What explains this fact? When presented with this kind of case, most philosophers reach for a certain kind of answer, an answer framed in terms of *identity*. The basic gist of the response is that this is not a good option for Chloe, because the individual post-change *is no longer Chloe*. Now it *may* be that the answer lies somewhere in this area. But there are at least two ways of understanding or interpreting this idea that strike me as quite unpromising. I want first intend to explain what these are and why they fail. This will set the stage for my explanation of why I remain skeptical of the entire strategy.

There are at least two quite different things philosophers talk about using the language of identity. These are what I shall here call *numerical identity* and *character identity*.[10] I examine numerical identity first.

Questions about numerical identity are questions about the persistence conditions over time for a certain type of entity. It is true that if a change were so radical that it caused the individual who changes to go out of existence, we could not, in the cases that interest us, say that this was good *for* the individual. But is that what radical change does in a case like Chloe's? Does it destroy one of us? That depends on what you think *we* fundamentally are.

A theory of numerical identity for x's presupposes an understanding of what an x essentially is. Thus, closely allied with theories of numerical identity are theories of our essence, theories that answer the question of what *we* most

[10] Schechtman draws the same basic distinction in terms of two questions: the re-identification question and the characterization question. Marya Schechtman, *The Constitution of Selves* (Ithaca: Cornell University, 1996). DeGrazia draws the same distinction in terms of numerical identity and narrative identity, but since I want my argument to apply to a broader set of theories of self than just narrative views, I avoid his terminology and refer to "characterization identity." See David DeGrazia, *Human Identity and Bioethics* (Cambridge: Cambridge University Press, 2005).

fundamentally are. There are (at least) three popular answers to the question of our essence. To begin with, we are human beings, a type of animal. Animalist theories of identity over time explain our persistence conditions in terms of the persistence of a particular animal or organism.[11] We are also, however, sentient creatures in virtue of having a brain that persists through time and produces conscious experiences. Embodied mind accounts of identity explain our persistence conditions in terms of the continued existence of a brain capable of conscious experience.[12] And finally, we are persons in the philosophical sense of beings with certain complex psychological capacities such as higher-order desires, and awareness of ourselves as temporally extended beings.[13] On these views our persistence through time consists in the persistence through time of certain psychological connections, often described as overlapping chains of psychological connections.

For a long time the question about the persistence conditions of beings like you and me was referred to as "*the* problem of personal identity." But this description is misleading. It runs together two distinct questions. One is the question of what it takes for the same *person* to survive a change. The other is the question of whether we are essentially persons, such that the destruction of the person I now am would be the death of me, or whether I could survive as a living non-person. For a person essentialist an answer to the first question is also an answer to the second. But for other types of theorists—for example, animalists and embodied mind theorists—the questions are importantly distinct. For these theorists it is quite possible to accept *as a theory of what it takes for a person to continue* one of the familiar psychological accounts of psychological continuity, even while rejecting the idea that you and I are essentially persons. For these theorists

---

[11] See several examples of such a view see e.g. Eric T. Olson, *The Human Animal: Personal Identity Without Psychology*, (Oxford: Oxford University Press, 1997) and Peter van Inwagen, *Material Beings* (Ithaca: Cornell University Press, 1990).

[12] For several defenses of this kind of view see McMahan (2000); DeGrazia (2005).

[13] There are many such views, e.g. Derek Parfit, "Personal Identity," *Philosophical Review* 80: (1971):3-27 and *Reasons and Persons*, (Oxford: Oxford University Press, 1984); Sydney Shoemaker, "Self and Substance," in *Philosophical Perspectives*, (Vol. 11) J. Tomberlin, (ed.): (1997): 283-319; "Self, Body, and Coincidence" *Proceedings of the Aristotelian Society* (Supplementary Volume), 73: 287-306.

personhood is a phase of our life, but it does not necessarily set the boundaries of our existence.

Importantly, for a long time the two questions were routinely conflated, in part because so many theorists unreflectively assumed person essentialism. There are various diagnoses of what led to this. For example, Eric Olson contends that for a long time philosophers conflated issues of practical concern—ethical issues—with metaphysical ones. Psychological continuity theories may help us answer many of our important ethical concerns, but they are very poor accounts of what we are metaphysically.

I myself am inclined to think I could survive as a non-person (for example, if I ever develop dementia and live to the end stages of the disease) or as a different person (for example, if I went through a procedure like the magnificent alteration). It also seems plausible to me that I started life as a non-person (a fetus). On the view I favor "person" is just something I am during a certain phase of my life, much as I am a parent now. Saying this need not imply that persons are unimportant or that it is not typically better to be a person than to be a non-person. However, the important point for the purposes of this essay is that radical change of the sort that disrupts psychological continuity enough to potentially create a new person or a non-person, will only disrupt numerical identity *if* it turns out that we are *essentially* persons. So we can only hope to use numerical identity to explain the problem with Chloe's case if we are person essentialists.

I have so far not explained *why* I think we should reject person essentialism, but I will now. As I see it, the decisive considerations have to do with the familiar notion of anticipation. Anticipating our own future experiences is very different from knowing that someone else will experience certain things in the future. And when we look to intuitions about which future experiences we can anticipate having, these suggest that one of us would survive even radical changes of the sort described in Chloe's case.

Bernard Williams offered the following famous example that highlights the issue. He asked his reader to imagine being captured and then told the following: "A series of things are going to be done to you, but the upshot is that the contents of

your mind will be fully erased. Once the procedure is complete, the living individual who remains—an individual who will have the same brain as you but who will not think of himself as you nor have any memories of your life—will gradually be fed a series of false memories and other mental contents so that he comes to think of himself as a different person. After this is all complete, this individual will be tortured."[14]

Williams asked whether or not it makes sense to be concerned about what will happen to this later individual. Obviously you will be upset by the prospect of procedures that will erase your mind. But by hypothesis, you can't stop that. Should you fear the torture that the post-procedure individual will experience? Williams suggests (and I believe) that fear of the torture makes perfect sense. But this, in turn, suggests that numerical identity is something that *can* survive the destruction of all one's psychological features. It can survive quite radical change of the sort that would alter or even destroy the original person.

Admittedly, not all theorists will agree with me, as these issues remain deeply contested. Anyone committed to person essentialism will think that the destruction of the person is the death of me, and that there is no sense in which I can be said to experience that later pain. However, for those who accept my rejection of person essentialism, it should be clear that we can't explain the badness of radical change in Chloe's case *by appeal to numerical identity*. On the view I favor, Chloe is still around and has not changed her (numerical) identity even after such a radical alteration.

[14] Williams stops short of drawing the strong conclusion that I draw, namely the conclusion that numerical identity can survive the destruction of the person. His paper instead emphasizes the ways in which our intuitions can apparently go quite different ways depending on how a case is described. However, many philosophers (including myself) have subsequently been willing to use examples like this one, which comes from Williams, to draw stronger conclusions than Williams did in the essay where this first appeared. Bernard Williams, "The Self and the Future," *The Philosophical Review* 79: 2: (1970): 161-180.

Published In: *Oxford Studies in Normative Ethics Volume 9*. Edited by Mark Timmons, Oxford University Press (2019).

§4.0 Why Identity is Not the Solution: Part II. Character Identity

There is another very common use of 'identity' according to which our identity is, roughly, our *self*. An account of our identity in this sense is an answer to the question "Who am I?," a question Marya Schechtman labels "the characterization question."[15] There are actually two closely related concepts here. Sometimes philosophers are interested in what they call "the true self." An account of the true self is an account of which character traits, values, beliefs, etc. are truly definitive of a particular person. Sometimes, however, philosophers are more interested in an individual's own self-conception, the particular way that *she* answers *for herself* the characterization question. Either way, however, on most such views, character identity is primarily defined in psychological terms. *Who* I am depends on things like my personality, my values, my characteristic responses to things, and so on and so forth. Although interesting questions can arise about who someone really is if or when her self-conception and the facts about her psychology diverge, for the purposes of this project such issues are not really important. We are simply interested in how much change a self (or a self-conception) can undergo before we have a new self (or before a person's self-conception is so different as to be a new self-conception). The notion of self may seem more relevant to explaining the badness of certain radical changes.

Unfortunately, these views, though interesting and important in many respects, tend to be vague when it comes to the questions that interest me. How much change and what kind of change is compatible with being the same self or the same person in the characterization sense? What accounts for the difference between a case where a change occurs and what we have afterwards is the same person with some different qualities (call this "old-self-modified") and cases where so much has changed that now we have a new person entirely (call this "new-self-

[15] See Schectman (1996) and "Staying Alive: Personal Continuation and a Life Worth Living," in *Practical Identity and Narrative Agency*, (eds.) Kim Atkins and Catriona Mackenzie (New York: Routledge, 2008), 31-55.

formed")? It is presumably a matter of degree, and different theorists will answer differently. Obviously, because the Chloe case is so extreme, almost any view will see it as a case of new-self-formed. But recall that we want an account that allows us to draw a plausible line *between* a case like Chloe's and a case like Sharon's. We want to continue to be able to say that radical change is the best option for Sharon and that she has most prudential reason to choose it. We don't want to purchase the right answer for Chloe at the price of giving up our claim about Sharon. Thus everything turns here on whether theories of the self would see Sharon after she takes the meds as a new-self-formed or as an old-self-modified.

Consider then a cluster of popular theories of self united by the emphasis they place on a person's core values. Call these value theories of self. For example, many theorists, inspired by Harry Frankfurt have come to understand the self in terms of commitments (and values) that are made (or adopted) by a higher-order self (or perhaps the rational self).[16] Frankfurt actually says more about personhood than selfhood.[17] He gives us an account of what it takes for a person to remain a person (as opposed to a non-person). Roughly for Frankfurt personhood is tied to the *capacity* for developing stable higher-order attitudes. Lose that capacity and you will cease to be a person. However, many people interested in the self have also been inspired by Frankfurt's ideas, and it is clear why. There are materials here that seem well suited to answering the characterization question. Who I am at a given time may also be determined by the attitudes of the higher-order self, or the rational self. What I come to value in this reflective way may serve as an account of who I am, most fundamentally.

---

[16] Schechtman (1996, 2008) among others cites Frankfurt as an example of someone who offers a theory that can be read as an answer to the characterization question. Another example she cites is Christine Korsgaard, *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996); and *Self Constitution: Action, Identity and Integrity* (Oxford: Oxford University Press, 2009), though Schechtman writing in 2008 cites the version that, at the time, was available on Korsgaard's website.
[17] See e.g. Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68: 1: (1971): 5-20. For other Frankfurt essays key to the interpretation of theorists like Schechtman see "The Importance of What We Care About," *Synthese: An International Journal for Epistemology, Methodology, and Philosophy of Science*. 53: 2: (1982): 257- 272; "Identification and Wholeheartedness In: *Responsibility, Character and the Emotions: New Essays in Moral Psychology*, ed. Ferdinand Schoeman (New York: Cambridge University Press, 1987), 27-45; "Identification and Externality," In: *The Identities of Persons*, ed. A. Rorty (Berkeley: University of California Press, 1976), 239-52.

However, this still doesn't help us to answer the question of when a person (without ceasing to be a person) becomes different *enough* to count as a new or distinct person. Presumably I could maintain throughout life the capacity for higher-order thought (and so always remain a person), even while changing my mind about some of the things that matter. I doubt any advocate of such a view would deny that. But what they don't say is how much such change is compatible with remaining the same person in the characterization sense. Alas, there is no precise answer to be found. So it remains unclear whether this kind of approach would count Sharon's changes as sufficiently radical to make her a new person.

Still, somewhat worrisomely, it seems plausible that a philosopher working in this vein *might* say that post-medication she is a new-self. A few years out from her change (if she undergoes it) she will care very little about most of the things that seemed important to her before. People who knew her before will be struck by the depth of the change. Our goal, however, is to find a way to use the notion of identity to distinguish between Sharon and Chloe. So only if we are certain that such a view of self can accommodate the intuition that Sharon's change is best for her should we appeal to it.

Now consider briefly a very different approach to understanding the self and answering the characterization question. This is a type of view that has become quite popular in recent years: the narrative self view.[18] Though there are many variants of this approach, the basic idea is easy enough to grasp: it is that a person's self is constituted by a narrative. This narrative is the story she tells herself about herself and about what has happened to her and why. It is composed of much less than all the things that have happened in her life. Rather it incorporates those happenings that seem *important to her,* and that stand in special narrative relations to other events in her life. Nor are these events limited to just the ones that seem good to her. Rather, the narrative incorporates events both good and bad that need to be understood or that help to make other events intelligible. The themes and patterns that emerge help individuals create meaning out of what might otherwise

[18] Again, there are many examples, but Schechtman (1996) is a main defender. DeGrazia (2005) also accepts a narrative view of self.

seem like a random series of events. This self-constituting narrative allows a person to make sense of where she has been and where she is going.

As before we want to know whether the narrative-self view can handle Sharon. It is pretty clear that it will deem Chloe to be a new-self after the magnificent alteration. After all, there could be no more thorough disruption to a narrative. Yet presumably the narrative self view, like value based theories of self, can accommodate a certain degree of change as long as there is a way to make sense of the change *within* a single, coherent story. The story must, however incorporate what has happened before the change and after it. This is not a huge constraint, since stories can develop in so many ways. For any particular beginning there are presumably many coherent continuations.

And yet, even so, I don't know whether such a theory could be trusted to give the right verdict about Sharon. Sharon has a self narrative that she has developed over time. Indeed, as explained earlier, part of her resistance to taking the medications is that she worries it will change her in ways that, from the perspective of her current narrative, are incoherent. Suppose she takes the meds. For her to reap the benefits of the change, sooner or later she will need to acknowledge that the changes she thought would be bad are not really so. But this will require abandoning key elements of the old story. For example she may eventually have to admit that her earlier self was not nearly as "deep" as she took herself to be, nor were others so "shallow." I would thus not be surprised if it were difficult or even impossible to find a single coherent narrative that could incorporate all of this in the right way. Again, until we are sure the narrative self view can give the right verdict about Sharon, we should not appeal to it to draw the line between good and bad radical change.

§5.0 Persons, The Self and Welfare

I have described a problem that arises in one form or another for many theories of welfare, namely, the problem of radical change—the problem of capturing deeply held intuitions to the effect that certain kinds of extreme changes

could either *never* be prudentially best or could *only rarely* be so. I have also tried to explain why what might seem like the most obvious response to this problem will not straightforwardly work. In this last section I present considerations that suggest that the whole *general strategy* of appealing to characterization identity is misguided. I then sketch an alternative approach.

Ordinarily we tend to suppose that we have at least *a* reason to resist too much change to our selves, to *who* we are. Of course, this reason can be overridden if enough factors point a different way. But it would presumably take a lot to override it. Yet, what makes us think we have such a reason? There may be moral reasons in the vicinity. But even so, if that is all we can say, we have not done anything to block the conclusion about welfare. The intuition we started with, however, was precisely that radical change was bad *for Chloe*. What we would need instead is a *prudential* reason to remain as we are. But why think remaining the same has general prudential value?

If we set aside temporarily the cases we have been discussing, we can see that in *many* real life situations radical change is prudentially bad. Moreover, radical changes are risky because in real life it is so hard to know what the net effects of such a comprehensive change will actually be. These facts may explain our general presumption against radical change, but they don't support a reason not to change. And they don't help us with Chloe.

There are other reasons why we are usually strongly resistant to change. On most views of the self a person's values are a key component of who she is. But it is at least partly constitutive of genuinely valuing something that one takes oneself to be *correct* in doing so.[19] In other words, if I value a certain kind of project, I not only approve of it or see value in it, but I may feel that I would be mistaken not to approve of it *and also* mistaken not to have it in my life. While understandable, this seems to conflate the value something has in itself (in terms of which it may merit a

[19] For example, Jaworska writes, "one would always view the possibility of not valuing something one currently values as an impoverishment, loss, or mistake." Agnieszka Jaworska, "Respecting the Margins of Agency: Alzheimer's Patients and the Capacity to Value. *Philosophy and Public Affairs* 28: 2: (1999): 105-38.

positive evaluation) and its prudential value *for me*. Even if the thing has great value, it might not be a mistake to remove it from my life. Again, however, this explanation of our attitudes does nothing to justify them or show that we have reason to resist self change.

Perhaps one might argue that radical self-change just is, in itself, a brute kind of prudential bad, one that has significant negative weight and which gains more negative weight as change becomes greater. Of course, it would still be just one of many prudential goods and bads that would factor into an assessment of the net value of a possible future, but because of it is so very bad, really radical change would rarely (if ever) turn out to be prudentially best. This yields the right answer at the price of great mystery. What could motivate such a move?

If that general line of thought goes nowhere, then perhaps we should reconsider the resources available within the FBR. One important kind of consideration can be brought to bear to explain why *many* radical changes would not be prudentially best. Whenever a person undergoes a radical change there will be transition costs. The greater the change the greater the transition costs. For example, if I have changed dramatically, I may not enjoy or even have the aptitude for many of the things I once did. So I will need to find new projects to engage me. I also might need to find new friends or form new relationships. I might need to find a new career. Yet all of these things take time. While I am working to re-create my life there will be a period of time where I have lost the old goods and do not yet have the new ones. So there will be a period of very low perhaps even negative welfare value.

The thing to notice is that any possible future that involves radical change and yet has net positive welfare value is one where the good that follows the transition must be *so* good it can compensate for the bad and still tip the balance to the positive. If the transition costs are big enough then the goods that follow must be *tremendous*. Finally, if a possible future is *best* (as opposed to just net positive), then it will have to be good enough to cancel the transition costs and still come out with more positive value than other relevant possibilities, and there will typically be a number of these all of which involve less change. This is a steep requirement.

However, I don't deny that some theories will meet it. Yet which theories meet it and how often they recommend radical change will depend on the details of the theory of prudential value. Consider briefly a simple form of Benthamite hedonism. Pleasure is the only within-life intrinsic value, and it is construed as a simple feeling the value of which is exhausted by facts about intensity and duration. Transition costs are typically highest when we try to recreate post change complex goods like a network of friends or an engaging career, things that typically take years to establish. But if things like friendships and careers have value only instrumentally insofar as they lead to pleasure, then it may well be that the best strategy after a radical change is to look for simpler, more easily attainable sources of pleasure. Assuming there are some, as seems plausible, then there will be a fair number of cases where radical change is part of the best possible future for a person. So she will have most reason to select it. But those possible futures will look quite different from futures in which we work to establish new networks of friends. In short, simple hedonism would presumably recommend radical change far more often than other theories because it has ways to minimize transition costs. But its ways of minimizing transition costs will not strike many as really prudentially good, thus revealing another implausible consequence of simple hedonism.

Suppose, however, we assign prudential value directly to things like engagement in valued projects and the development and maintenance of good relationships. And suppose we allow that such things have much more value when (in the case of projects) they engage more of our faculties or when (in the case of relationships) our feelings are strongest and returned, and the relationship lasts a long time. On such a view the transition costs of radical change begin to look formidable.

Even so, I imagine there will be cases where radical change turns out to be worth it. Someone who is very young has years ahead to refashion a life, including time to develop and nurture new relationships, and so on and so forth. If that life, once created, would be dramatically better, then it might prudentially speaking be worth it for this person to start over. Even for someone who is not quite so young but who has no very good relationships or no valuable, engaging projects, it may

make sense to start over. Still, once we adopt a more sophisticated theory of value, transition costs can go a long way towards explaining why radical change is rarely the best option.

Such an approach can explain why Sharon's change is good. She is young, her depression is undermining her relationships and she has the potential to do more and relate better if she changes. As for Chloe, how we view the case may depend on details that were not initially provided. It was simply stipulated when the case was presented that it was one where the prudential value of the possible future with radical change was clearly much higher than any others. But is that even possible? It is of course logically possible, but is it going to turn up a real possibility given the world as it is and the facts of human psychological being what they are? Perhaps it would if her life to date were miserable, and she had no other possibilities that would lead to engaging work and deep relationships. Then it might be plausible that the greatest net value lies on the other side of the magnificent alteration, even given the huge transition costs. But it now seems like a very rare kind of case indeed.

Obviously I have not solved these problems entirely. It remains to be seen whether this approach can resolve all the problems. But it seems more promising to me than a focus on the importance of identity. For it seems to me that the importance of identity is not a brute prudential fact. Rather it matters, when it does, because of what *else* depends on it. Some of *my* prudential goods require that I be me. And since usually they are the best goods open to me, it remains important that I remain me.