This is an early draft of a paper forthcoming in *Philosophical Studies*. It is going to change quite a bit. Please cite the published version.

June 2008

The Ethics of Morphing

Sometimes we hurt people in some ways so as to benefit them in other ways. I do this all the time to my four year old daughter. She complains, and she has real grounds for complaint. Vaccinations *are* painful, broccoli *is* disgusting, she *will* have fun if she stays up past her bed-time. But I ignore her and mutter to myself, above the screaming and the raging and the aggrieved whimpering: 'on balance, the benefits to her outweigh the costs to her.' That consideration justifies what I am doing.

Can I legitimately aggregate costs and benefits across peoples' lives in the way that I legitimately aggregate costs and benefits within my daughter's life? Can I, for example, impose a cost upon one person so as to benefit someone else, and justify my behavior with the thought: 'on balance, the benefits to them outweigh the costs to them'?

Some very influential philosophers have claimed that I cannot. When Jack suffers a little so that Jill may prosper a lot, that is great for Jill and unfortunate for Jack. But there is no morally significant sense in which the benefit to her outweighs the cost to him. If I go about trying to aggregate costs and benefits across lives, under the impression that something morally important hangs on the result, then I am making a mistake.[1] So utilitarianism, the ethics of whole-sale aggregation, fails. When I consider what to do in a

Many thanks to Agustin Rayo, Liz Harman, Mark Heller, Steve Yablo, Sally Haslanger, Simon Keller, Joshua Schechter, Ted Sider and Michael Titelbaum, for most helpful discussions and comments.

[1] Different philosophers offer different diagnoses of the source of the mistake. For John Rawls and Robert Nozick it emerges from our failing to appreciate a metaphysical fact: the 'separateness of persons'. The notoriously enigmatic arguments to this effect are in Rawls (1971) sec. I.6. and Nozick (1974) 32-3. For Judy Thomson it emerges from our being lulled by the surface-syntax of phrases in which 'good' and 'better' figure into thinking that states of affairs can be better or worse simpliciter. See Thomson (2001) Part I and Thomson (2006).

case when people's interests conflict, I must think in terms of *rights*, *desert*, *fairness*… and so forth.

This is a very attractive idea. But sometimes very attractive ideas unravel when you inspect them closely. In this paper I will develop a line of thinking that tells in favor of inter-personal aggregation in a certain class of situations. It goes roughly like this: Identity across states of affairs is a slippery thing. Take a walk across physical space and you will find that the boundaries between people are, in all but exceptional cases (mothers and fetuses, conjoined twins), pretty clear. Take a walk along an appropriate path across the space of all possible states of affairs and you will find that the boundaries between people are less clear. You will find people blending smoothly into other people: Russell into Frege, Churchill into Hitler, Saint Francis of Assisi into Genghis Khan. Because identity across states of affairs is slippery in this way, if you care about making things better for particular people, and you are consistent, then, in certain situations, you must trade-off costs and benefits across lives.

I will begin by looking at one case in which there is a temptation to aggregate inter-personally, and then generalize.

1. MORPHING AND NON-IDENTITY CASES

1.1 One Non-Identity Case

> *Charlotte*
> Charlotte is wondering whether to conceive and bear a child now, when she is fourteen, immature, feckless, besotted with an unsteady boyfriend, or wait eleven years.

If Charlotte appealed to you for advice, you might be tempted to tell her that she ought to wait – “After all,” you might be tempted to say, “even if you have convinced yourself that this is what *you* want, consider your baby. If you go ahead now then your baby will, most probably, have a poor quality of life. If you wait then your baby will, most probably, have a higher quality of life. It will be better, on balance, if you wait.”

If the influential philosophers are right, then this looks like one instance of the inter-personal aggregative mistake. If Charlotte goes full-steam ahead then she will have one baby. If she waits then she will, most probably, have a different baby. So the states of affairs she is in a position to bring about are, most probably, something like these:

S_{James} In which Charlotte conceives and bears her only baby when she is fourteen, and baby James has a predictably poor quality of life.

S_{Jane} In which Charlotte waits until she is twenty five to have her only baby, and baby Jane has a predictably high quality of life.

The first state of affairs may be better for James than the second, the second better for Jane than the first[2], but there is no significant sense in which the second is simply better than the first (or ought to be preferred over the first, or ought to be brought about rather than the first), just in virtue of the fact that Jane’s quality of life in the second outweighs James’ quality of life in the first. If you are to convince Charlotte that she ought to wait

[2] Some writers (e.g. John Broome – see Broome (1999) sec. 14.3) would have it that we can’t even say this. Loosely: you can’t be better or worse off existing than not. More precisely: for any states of affairs A, B, and person P, A is better than B for P only if P exists in both A and B. Others (e.g. Josh Parsons – see Parsons (2002)) disagree.

by appeal to her child's quality of life, then you will have to do so in a much more circuitous way.[3]

But I say that, although James and Jane are different babies, it follows more or less directly from the fact that Jane's quality of life in S_{Jane} is higher than James' quality of life in S_{James}, that Charlotte should prefer that S_{Jane}, rather than S_{James}, come about. And I have an argument. I will call it the *morphing argument*.

It begins with the observation that parents have a moral obligation to care about their children in certain ways.[4] At the very least, parents ought to prefer that, other things being equal, any given child they have be better off. So, at the very least, we can say of Charlotte:

(Personal Dominance)

Charlotte ought to prefer a state of affairs in which she has a baby, over an all-other-things-equal state of affairs in which she has the same baby and he or she is worse off.

This is a weak principle. What about cases in which all other things are not equal? What if there is some cost associated with the baby's being better off? What if Charlotte

[3] Derek Parfit, who first drew attention to this sort of case in Parfit (1984) sec. 122, took the view that the girl should wait, and should do so because a state of affairs like S_{James} is, other things being appropriately equal, worse than a state of affairs like S_{Jane}. Many writers, wanting to accept Parfit's judgment but reject his justification for it, have come up with ingenious and (I think) fragile justifications of their own. Some (e.g. Woodward (1986), Benetar (1997) and Schiffrin (1999)) appeal to the idea that Charlotte would wrong James by bringing about S_{James}. Others (e.g. Freeman (1997) and Harris (1998)) appeal to the idea that parents have unusual impersonal duties.)

[4] Someone might object: 'We certainly have a moral obligation to care *for* our children – to act in certain ways on their behalf. But it's not true that we have a moral obligation to care *about* our children – to prefer that they be better off. The proper objects of moral assessment are things we do, not things we want.' This strikes me as implausible. When my child falls off a distant play-structure, and I run over to her, I want her not to have hurt herself. If I didn't, if I was completely indifferent about whether she had broken her nose, I would be morally deficient.

will be worse off? What if a sibling will be worse off? What if there will be unwholesome inequity between the baby and its peers? Ought she to prefer that her baby be better off in spite of the cost? – The principle says nothing about whether she ought to have preferences in such cases. And what if Charlotte is deciding which baby to bring into the world? Ought she to prefer that she have a better off baby rather than a different, worse off baby? – The principle says nothing about whether she ought to have preferences in such cases.

Weak though the principle may be, it is all we need for now. But to make progress with the morphing argument we need to state it more precisely. In particular, we need to be clear about what sorts of things *states of affairs*, the things that can be better or worse for people, are. And we need to be clear about what it is for a baby in one state affairs to be *the same* as a baby in another state of affairs.

There are two very different ways of thinking about sameness-across states of affairs. I do not want to presume in favor of either way of thinking here, but the appropriate way to run the morphing argument depends on which way of thinking we adopt. So, over the next two sections, I will provisionally adopt one way of thinking (*counterpart theory*) and run the morphing argument one way, and then provisionally adopt the other (*real identity theory*) and run the morphing argument another way.

1.2 Morphing Mk. 1: Counterpart Theory

Let's assume, first, that states of affairs are possible worlds, and let's adopt a counterpart-theoretic treatment of sameness-across-worlds: Ordinary entities, like people, chairs and slugs, exist in one world only. Entities that exist in distinct worlds may be qualitatively similar, and for the purposes of assessing de re modal claims (claims about how particular things could have been) it may be useful to consider appropriately similar entities in distinct worlds to be *counterparts*. This will allow us to say, for example, that it is true of any actual thing that it could have been thus-and-so if and only if it has a counterpart that is thus-and-so (so I could have been a superb tennis player, because I have an other-worldly counterpart who is a superb tennis player.) But an entity and its other-worldly counterpart are distinct things.[5]

Assuming all this, here is the natural way to state the dominance assumption precisely:

(Personal Dominance)

For worlds W_i, W_k, in which Charlotte-counterparts have counterpart babies B_i, B_k, if B_i is better off than B_k and all other things are appropriately equal, then Charlotte ought to prefer that W_i be actual.

When are babies in different possible worlds counterparts? How you answer this question will depend on your views about personal essence, framed in the language of counterpart theory. Prima-facie plausible doctrines in this area are *genetic essentialism* (babies are counterparts only if their genetic profile is sufficiently similar), *essentialism about origins* (babies are counterparts only if the conditions under which they come into being are sufficiently similar) and *psychological essentialism* (babies are counterparts only if their present and future psychologies are sufficiently similar).

[5] See Lewis (1968), Lewis (1971), and Lewis (1986) section 4.3, for evolving expositions of the idea.

On no plausible view about personal essence will it turn out that James and Jane are counterparts. We would clearly be blundering if we said to James "If Charlotte had just waited eleven years, and had a girl, with a different father, then you would have been a girl, born eleven years later, and you would have been better off." So *Personal Dominance* does not immediately imply that Charlotte ought to prefer that W_{Jane}, the world in which she conceives happy Jane, rather than W_{James}, the world in which she conceives miserable James, be actual.

But now assume (plausibly enough, again) that our essences are not, along any dimension, *perfectly fragile* – which is to say that there is no respect of similarity and difference such that any two counterparts are precisely the same in that respect. Essence would be perfectly fragile with respect to origin, for example, if two babies are counterparts only if they came into being in exactly the same way – so, if my parents had conceived a baby qualitatively just like me, but one millisecond after I was actually conceived, and that baby had gone on to live a life just like mine, then I would never have existed. We assume that essence is not fragile in this way.

It follows that we can construct a *morphing*[6] *sequence* of intermediary worlds W_1, W_2,…,W_n such that:

(Morphing)

James in W_{James} has a counterpart in W_1, who has a counterpart in W_2,…, who has a counterpart in W_n, who is a counterpart of Jane in W_{Jane}.

[6] I take it that the reader is familiar with 'morphing' animation software, invented in the 1980s and refined ever since. You input images of two objects (canonically: the president and a chimp) and it outputs an intermediary sequence of images. Play them in order and you see the one object smoothly transforming into the other.

As we move along this sequence so we encounter babies who are born increasingly later, increasingly less like James, increasingly more like Jane.

How many intermediary worlds will there need to be? What will the transitions need to be like? That depends on which views about personal essence are correct. So, for example, if *gender essentialism* is correct[7], if no determinately male baby is a counterpart of any determinately female baby, then there will need to be some gender-ambiguity at some world in the sequence. There will need to be a gradual transition, mid-way through the sequence, from determinately male babies to determinately female ones. But if gender essentialism is false then there can be a clean jump. No matter. So long as essence is not perfectly fragile along any dimension, we know that a morphing sequence of some kind can be constructed.

Now assume (plausibly enough, again) that well-being is fine-grained. Between a great life like Jane's, full of love and hope and long, lazy evenings playing on the lawn, and a lousy life like James', full of deprivation and insecurity and long, anxious evenings wondering if he is going to eat dinner, there are ever-so many intermediaries, enough to allow us to construct a morphing sequence with the following feature:

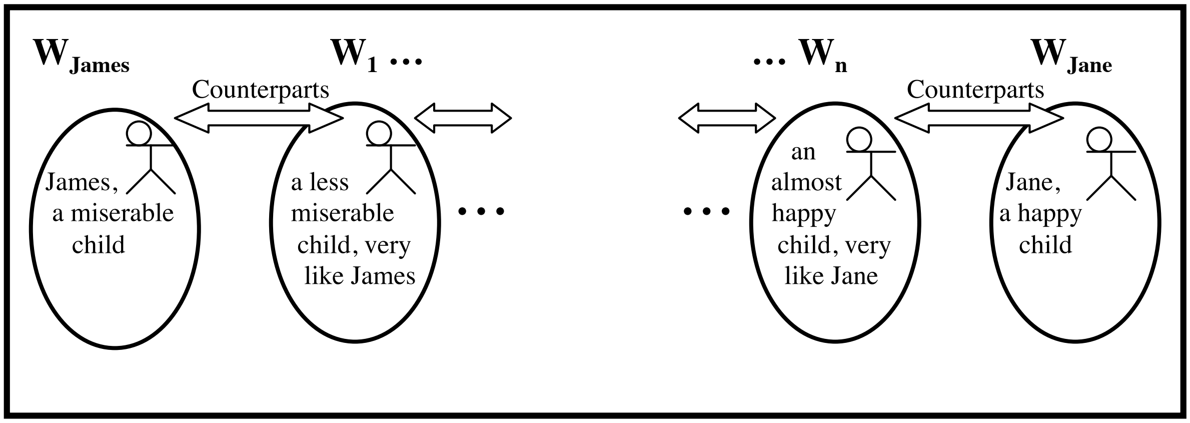
(Up-Slope Morphing)

James in W_{James} is worse off than his counterpart in W_1, who is worse off than his counterpart in W_2, …, who is worse off than her counterpart in W_n, who is worse off than her counterpart in W_{Jane}, Jane.

[7] This is an odd view, of course. We ordinarily think that people survive sex-change operations. If a woman can become a man, then why couldn't a baby girl have been born a baby boy?

Here is a picture of the sequence:

An Up-Slope Morphing Sequence



As we move along it so we encounter worlds that, by *Personal Dominance*, Charlotte ought to prefer over their predecessors, because in each world Charlotte's baby is better off than its counterpart in the predecessor world.

Finally, let us assume (plausibly enough, again) that practical rationality places certain constraints upon what states of affairs to prefer and bring about. In particular let us assume that if you are rational then, for worlds W_1, W_2, W_3,

(Transitivity)[8]

If you prefer that W_1 rather than W_2 be actual, and you prefer that W_2 rather than W_3 be actual, then you prefer that W_1 rather than W_3 be actual.

[8] This is a foundational assumption of much decision-theory. Some dispute it. I won't get into that debate here. Let it suffice to say that if the cost of being an anti-aggregationist is to have intransitive preferences then many will, quite understandably, find that cost too high to bear.

It follows from *Transitivity* that if Charlotte is decent (she has the preferences that she ought, morally, to have – in particular, her preferences conform to the *Personal Dominance* principle) and rational then she will prefer that W_{Jane}, rather than W_{James}, be actual. If she is decent and rational then she will prefer that Jane come into being rather than James, just because Jane will have the higher quality of life.[9]

1.3 Morphing Mk. II: Real Identity Across States of Affairs

What to make of this argument? It relies, of course, upon a particular way of thinking about sameness-across-states of affairs – where states of affairs are the things that we are in a position to bring about by acting one way or another, and the things whose relative merits for people have a significant bearing on what we ought to do. Entities that exist in one state of affairs do not exist in another. There is a surrogate for identity-across-states of affairs: the counterpart relation. But, while identity is transitive, the counterpart relation is not. That is what gets the argument going.

You can, reasonably enough, take issue with this way of thinking about sameness-across states of affairs. You can insist upon thinking of states of affairs as ways for things to be for particular people. Things can be different ways for one person, so one person may figure in more than one states of affairs. In short: there is *real identity* across states of affairs.[10]

[9] Note that the conclusion is not that it is morally permissible for Charlotte to prefer to have the later child. This is something that opponents of inter-personal aggregation readily accept – after all, we are free to care about all sorts of things that have no moral significance. The conclusion is rather that if Charlotte does not prefer to have the later child then she is either morally or rationally deficient. This is something that opponents of inter-personal aggregation do not accept.

[10] A side-observation: this view, combined with the view that identity is a primitive notion (What is it for this thing and that to be one and the same? Just for them to be one and the same. Nothing more informative can be said.) is often attributed to Saul Kripke, and often contrasted with David Lewis' counterpart theoretic treatment of sameness-across possible worlds. But Lewis may not have disagreed with Kripke

Given this different picture of sameness-across-states of affairs, here is the natural way to state the dominance assumption precisely:

(Personal Dominance)

For states of affairs S_1, S_2, if Charlotte has one baby in both, and her one baby in S_1 is her one baby in S_2, and that baby is better off in S_1, then, all other things being appropriately equal, she ought to prefer that S_1 come about.

When is Charlotte in a position to bring about states of affairs in which she has the very same baby? How you answer this question will depend on your views about personal essence. Let's consider a particular example, a variant on our original example: Charlotte can bring it about that she has a baby by pressing any one of a range of buttons. If she presses the first then she will have a miserable baby boy at the age of fourteen. If she presses the next then things will be, qualitatively speaking, just as they are in the first intermediary world of the up-slope morphing sequence we considered earlier: she will have a slightly less miserable child, slightly later. If she presses the next then things will be, qualitatively speaking, just as they are in the second intermediary world in the up-slope morphing sequence… And if she presses the last then she will have a happy baby girl at the age of twenty five.

On any plausible view about personal essence it will be true that:

very much at all. Lewis also thought that identity is a primitive notion, and he thought that, although actual things do not exist in non-actual possible worlds, non-actual possible worlds in which counterparts of actual things exist nonetheless *represent* actual things. Supposing that Humphrey lost the presidential election, a non-actual world in which a Humphrey counterpart wins represents Humphrey, our Humphrey, that very same person, winning. It represents a way for Humphrey to be. See Lewis (1986) p.194. Think of the totality of what is represented as a state of affairs and you have Kripke's view: states of affairs are ways for particular things to be.

(Essence is Somewhat Fragile)

If Charlotte actually presses the first button then, if she had pressed the last button, her actual child, James, would not have existed.

Now things look promising for opponents of inter-personal aggregation. Suppose that Charlotte actually presses the first button. *Personal Dominance* does not entail that she ought to prefer the state of affairs that would have come about if she had pressed the last button, call it S_{waits}, over the actual state of affairs, call it $S_{@}$. (Why? – Because her child in S_{waits} is not her child in $S_{@}$.) Nor is there any intermediary sequence of states of affairs $S_1,\ldots,S_n$ such that *Personal Dominance* entails that she ought to favor S_1 over $S_{@}$, S_2 over $S_1,\ldots$, and S_{waits} over S_n. (Why? – Because *Personal Dominance* only applies to a pair of states of affairs if Charlotte's baby in the one is her baby in the other. It cannot be that for some $S_1,\ldots,S_n$, her baby in $S_{@}$ is her baby in S_1, and her baby in S_1 is her baby in $S_2,\ldots$, and her baby in S_n is her baby in S_{waits}, because her baby in $S_{@}$ is not her baby in S_{waits}, and identity is a transitive relation.)

Problem solved, you might think. And perhaps, you might think, this is a way to make sense of the notoriously elusive 'separateness of persons' objection to inter-personal aggregation: if counterpart theorists were right, if there were no real identity across states of affairs, just more or less qualitative similarity, then perhaps it would make sense to aggregate inter-personally, but there is real identity across states of affairs. Utilitarians fail to appreciate this, and that is where they go wrong.

But this is way too quick. Even if we take it that there is real identity across states of affairs, there remain good reasons to think that Charlotte ought to aggregate inter-personally. For, on any plausible view about essence, it will also turn out to be true that:

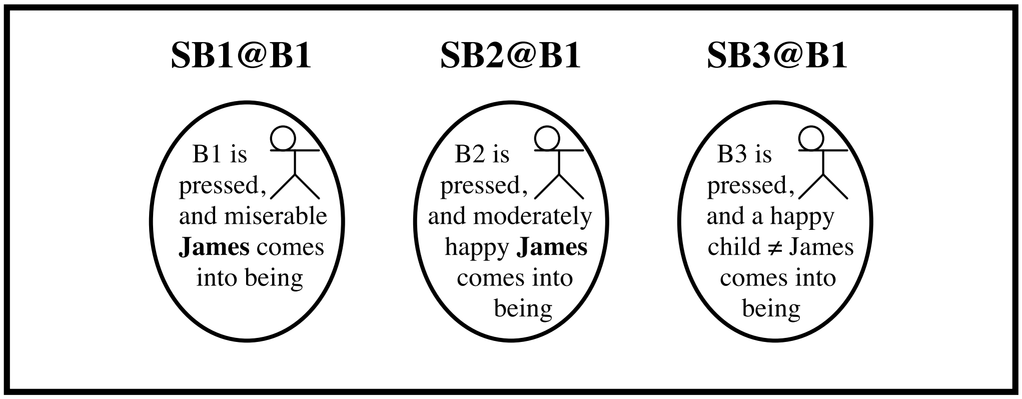
(Essence is not Perfectly Fragile)
Whatever button Charlotte actually presses, if she had pressed the previous or next one, then her actual child would still have existed.

All actual people could have been ever-so-slightly different along any given dimension.

To understand the implications of this principle, it may be helpful to consider a simplified version of Charlotte's case, in which she has three buttons to press. Pressing the middle one will bring a moderately happy, moderately male, moderately female baby into the world. And it may be helpful to introduce some notation – let SB*M*@B*N* be the state of affairs that, supposing Charlotte actually presses B*N*, would have come about if she had pressed B*M*.

Supposing that Charlotte actually presses B1, the actual state of affairs, SB1@B1, is one in which her actual child, call him *James*, is miserable. And, by *Essence is not Perfectly Fragile*, the state of affairs that would have come about if she had pressed B2, SB2@B1, is one in which that very same child, James*,* is moderately happy. And, by *Essence is Somewhat Fragile*, the state of affairs that would have come about if she had pressed B3, SB3@B1, is one in which some other child is moderately happy.

So, supposing that Charlotte actually presses B1, the states of affairs that she would have brought about by pressing each of the three buttons are:


SB1@B1
SB2@B1
SB3@B1
B1 is pressed, and miserable **James** comes into being
B2 is pressed, and moderately happy **James** comes into being
B3 is pressed, and a happy child ≠ James comes into being

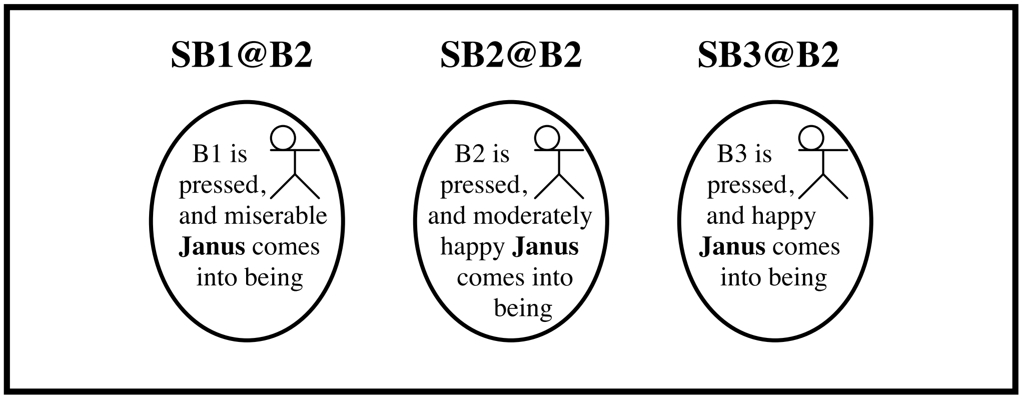
Personal Dominance tells us that she ought to have at least these preferences between these states of affairs:

SB1@B1 < SB2@B1 SB3@B1

Which is to say that if she actually presses B1, then she would have brought about a preferable state of affairs by pressing B2, but would not have brought about a preferable or less-preferable state of affairs by pressing B3.

But, supposing that Charlotte actually presses B2, the actual state of affairs, SB2@B2 is one in which her actual baby, call him/her *Janus*, is moderately happy. And, by *Essence is not Perfectly Fragile*, the state of affairs that would have come about if she had pressed B1, SB1@B2, is one in which that very same baby is miserable. And, by *Essence is not Perfectly Fragile*, the state of affairs that would have come about if she had pressed B3 is one in which that very same baby is happy.

So, supposing that she actually presses B2, the states of affairs that she would have brought about by pressing each of the three buttons are:


SB1@B2
B1 is pressed, and miserable **Janus** comes into being
SB2@B2
B2 is pressed, and moderately happy **Janus** comes into being
SB3@B2
B3 is pressed, and happy **Janus** comes into being

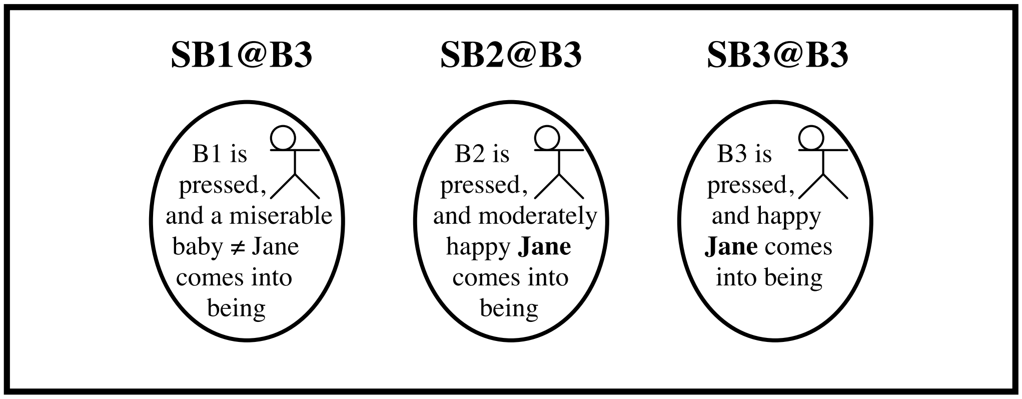
Personal Dominance tells us that she ought to have at least these preferences between these states of affairs:

$$\overbrace{\text{SB1@B2} \; < \; \text{SB2@B2} \; < \; \text{SB3@B2}}^{<}$$

Which is to say that, if she actually presses B2, then she would have brought about a preferable state of affairs by pressing B3, and would have brought about a less-preferable state of affairs by pressing B1.

And, supposing that Charlotte actually presses B3, the actual state of affairs, SB3@B3 is one in which her actual baby, call her *Jane*, is happy. And, by *Essence is not Perfectly Fragile*, the state of affairs that would have come about if she had pressed B2, SB2@B3, is one in which that very same baby is moderately happy. And, by *Essence is Somewhat Fragile*, the state of affairs that would have come about if she had pressed B1, SB1@B3, is one in which a different baby is miserable.

So, supposing that she actually presses B3, the states of affairs that she would have brought about by pressing each of the three buttons are these:


SB1@B3
SB2@B3
SB3@B3
B1 is pressed, and a miserable baby ≠ Jane comes into being
B2 is pressed, and moderately happy **Jane** comes into being
B3 is pressed, and happy **Jane** comes into being

Personal Dominance tells us that she ought to have at least these preferences between these states of affairs:

SB1@B3 SB2@B3 < SB3@B3

Which is to say that, if she actually presses B3, then she would have brought about a less-preferable state of affairs by pressing B2, but would not have brought about a preferable or less-preferable state of affairs by pressing B1.

So this is a curious situation in which whether or not Charlotte would bring about preferable states of affairs by doing one thing or another depends upon what she actually does, because what states of affairs she would bring about by doing one thing or another depends upon what she actually does. Call this a situation in which *counterfactuals are actuality-sensitive*.

A great deal has been written about what we ought, rationally, to do in situations in which counterfactuals are actuality-sensitive.[11] I will not propose a theory that tells us

[11] This is because a great deal has been written on Newcomb cases, and, on the standard interpretation of these cases, they give rise to situations in which counterfactuals are actuality-sensitive. Take the classic Newcomb case (in brief: I stand before two boxes, one transparent and one opaque. I can take either or both home with me. I see that the transparent box contains $100. What does the opaque box contain? I know this: Some time ago a fantastically accurate predictor predicted what I would do. If it predicted that I would take the opaque box only, then it put $1,000,000 in the opaque box. If it predicted that I would take both boxes then it put $0 in the opaque box.) On the standard interpretation, before playing the game it is right to

what rational people will do *whenever* counterfactuals are actuality sensitive. But I will propose and defend a partial theory, a theory that covers cases like Charlotte's. If she is rational then she will press B3.

To state the central principle I will need to introduce some terms. Say that option A is *pair-wise superior* to option B when all of the following hold:

(i) Supposing you actually take A, you would have brought about a less-preferable state of affairs by taking B.

(ii) Supposing you actually take B, you would have brought about a preferable state of affairs by taking A.

(iii) The state of affairs you will bring about, supposing you actually take B, is not preferable to the state of affairs you will bring about, supposing you actually take A.[12]

Now consider a procedure:

(Step 1) Choose an option.

(Step 2) If there are pair-wise superior options, choose one, otherwise keep the option you have.

(Step 3) Continue, until there are no pair-wise superior options.

think: 'I am confident that the fantastic predictor has correctly predicted the choice I will actually make, so if I actually two-box then I will end up with $100, and if I actually one-box then I will end up with $1,000,000. But I am equally confident that my present decision has no influence over what's in the boxes, so if I actually two-box then it will turn out to true that if I had one-boxed then I would have ended up with nothing, and if I actually one-box then it will turn out to be true that if I had two-boxed then I would have ended up with $1,000,100. What states of affairs I would bring about by doing one thing or another depends upon what I actually do.'

[12] You may wonder why we need this third condition. The idea is to make the notion of pair-wise superiority strong enough for it to be uncontroversial that, if option A is pair-wise superior to option B, and you know it, and you intend to take option B, and you are rational, then you will cease to intend to take option B. A causal decision-theorist will think that conditions (i) and (ii) alone suffice. An evidential decision theorist will not – for two-boxing is pair-wise superior to one-boxing in this weaker sense. I don't want to presume in favor of either theory here.

Say that option O is an *attractor* if, no matter how you apply this procedure (no matter which option you start off with, no matter which pair-wise superior options you choose along the way) you will always get to O.

Finally, say that option O is *stable* when both of the following hold:

(i) There is no option K such that, supposing you actually take O, you would bring about a preferable state of affairs by taking K.

(ii) There is no option K such that the state of affairs you will bring about, supposing you actually take K, is preferable to the state of affairs you will bring about, supposing you actually take O.[13]

Now here's the principle:

(Stable Attraction*)*

If you are rational, and an option is a stable attractor, then you will take it.

The motivating idea is that if an option is a stable attractor and you are rational then, no matter what your intentions are as you begin to think about what to do, you will end up with a settled intention to take it. Imagine yourself to be rational, surveying a range of options and trying to decide which to take. You form a tentative intention to take one of them. If there is another, pair-wise superior option then this intention is self-weakening. With the intention comes a belief that this is the one you will take. With the belief comes another belief, that you would bring about a preferable state of affairs by taking the other option. So your tentative intention fades, and is replaced by an intention to take the other option. But this new intention is not self-weakening. It brings with it a new belief, that

[13] Again, this condition is here so as to avoid presuming against evidential decision theory.

you will take the other option, but you still believe that you would bring about a less-preferable state of affairs by taking the original option, and you do not think that the state of affairs that will come about, supposing you actually do what you now intend to do, is less-preferable to the state of affairs that will come about supposing you actually do what you originally intended to do. It is not self-weakening unless, of course, there is yet another pair-wise superior option… And so you cycle through the options, arriving in time (as you must, because it is an attractor) at an intention to take the stable attractor. But this intention, at last, is stable. Supposing that you take the stable attractor, you wouldn't bring about preferable states of affairs by taking any other option, nor is the state of affairs that you suppose to be actual less-preferable to the state of affairs that will come about, supposing you take any other option. So you stick with it, being rational.

In Charlotte's case, the pushing-B3 option is a stable attractor. As she deliberates, no matter what her initial inclinations are, insofar as she is rational she will come to be inclined to push B3, and once she is so inclined there will be no considerations that tell against her seeing through on that inclination.

So, when Charlotte is in a position to push buttons B1, B2, B3, if she has the preferences that she ought, morally, to have (in particular: her preferences conform to *Personal Dominance*) and she is rational (in particular: she picks stable attractors) then she will press B3.

All very well, but what if she does not have the intermediary option? What if she has only two options – bringing miserable James into the world or happy Jane into the world? Well, I take it that practical rationality places a further constraint upon her. If you are rational then, for complete states of affairs S_1, S_2, S_3, (The Practical Insignificance of Irrelevant Alternatives)[14]

If, given the options of bringing about S_1, S_2, S_3, you will willingly bring about S_3, then given the options of bringing about S_1*, S_3* (complete states of affairs relevantly just like[15] S_1 and S_3, but in which the option of bringing about S_2 is not available to you), you will willingly bring about S_3*.

It follows[16] that, when Charlotte is in a position to press only buttons B1 or B3, if she has the preferences that she ought, morally to have, and she is rational, then she will press B3. She will bring Jane into existence rather than James, just because Jane would be better off than James would be.

1.4 Some Other Picture of Sameness-Across-States of Affairs?

I have considered two different ways of thinking about sameness-across-states of affairs, and argued that according to either, it follows from the fact that Charlotte's later

[14] A related but weaker principle (known variously as 'Basic Contraction Consistency' and 'the Chernoff Condition' and 'Principle α') is another mainstay of decision theory. It is a constraint on rational *preference* – if from S_1, S_2, S_3, you prefer S_1, then from S_1, S_3, you prefer S_1. It will become clear why I need the stronger principle.

[15] When are S_1* and S_3* 'relevantly just like' S_1 and S_3? When there are no differences that are evaluatively relevant by your lights, no differences that give you grounds for having different preferences between them. To see the idea, consider an example in which there *are* evaluatively relevant differences: Box A contains $10; Box B contains $15; Box C contains $20; you care about money, but you also care about keeping promises, and you have promised that, if you have the option of taking Box B, then you will take Box A. In this case it is not irrational for you to take Box A when you can take A, B or C, but to take Box C when you can only take A or C. Why? Because there is an evaluatively relevant difference between the state of affairs in which you have three options and take Box C and the state of affairs in which you have two options and take Box C – in the former you break a promise, but in the latter you do not. This difference gives you grounds for having different preferences in the two-option case and the three-option case, and for behaving differently in the two option-case and the three-option case. Thanks to Peter Graham for the form of this example.

[16] Note that the presence or absence of the B2-option is not evaluatively relevant by Charlotte's lights. It has no bearing on whether she prefers the state of affairs in which she presses B1 over the state of affairs in which she presses B3.

child would be better off that, if she is decent and rational, she will bring him into existence. Is there some yet different way of thinking about sameness-across states of affairs that blocks the argument? I think not. No matter about its details, any alternative to the counterpart-theoretic treatment of states of affairs will, on pain of grave implausibility, say that, supposing that things actually are the way things are in some world in the morphing sequence, if Charlotte had behaved as she does in the successor world in the morphing sequence, then she would have brought about a state of affairs in which her actual child exists, better off than he or she actually is. This will be enough to get the morphing argument going.

2. GENERALIZED MORPHING

2.1 Saving the Healthy

So what? You might wonder. Byzantine though the morphing argument may appear, its conclusion is a platitude. Of course, other things being equal, parents ought to create happier, healthier children.[17] These are not the difficult cases, in which inter-personal aggregation is especially problematic. The difficult cases are those in which we must decide how costs and benefits are to be distributed among existing people.

But the morphing argument can readily be applied to such cases. Consider:

> *Saving the Healthy*
>
> Billy and Ben are stranded on separate islands. You are a life-boat captain, committed to helping them. You can save one or the other by setting an appropriate course, but not both. Who should you save? Here’s one thing you

> know: Billy will have a higher quality of life, if you save him, than Ben will, if you save him.

You might, once again, be tempted to aggregate inter-personally, to think 'I ought to save Billy because, on balance, his interest in living is stronger.' And, once again, you would be right to do so. In this case, we may reasonably assume, your commitment to helping Billy and Ben extends far enough that you ought to favor one world over another if it is better for one of them and no better or worse for the other, and all other things are appropriately equal. And (thinking in the counterpart theoretic way) there are two possible worlds to consider:

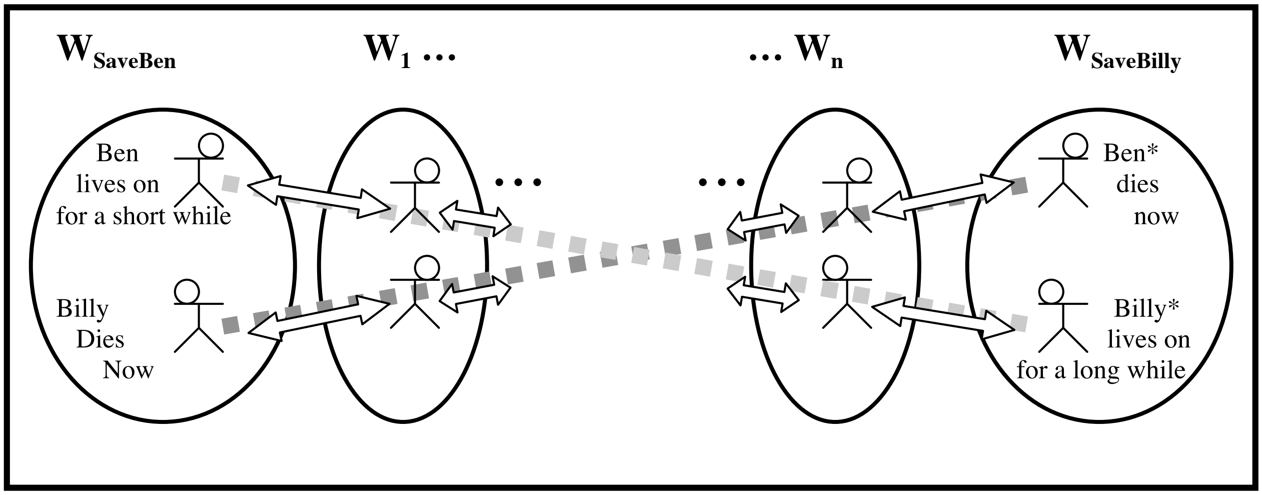
$W_{SaveBen}$ In which Ben gets saved and benefits little. Billy dies and loses a lot.

$W_{SaveBilly}$ In which Billy* gets saved and benefits a lot. Ben* dies and loses little.

Ben is the counterpart of Ben*, and Billy the counterpart of Billy*, so it might appear as if dominance considerations give you no obligation to favor either world – the first is better for Ben than the second, the second better for Billy than the first. But, so long as essence is not perfectly fragile, we can construct an *up-slope cross-morphing sequence* of intermediary worlds, through which there runs one chain of counterparts linking Billy to Ben*, and another chain of counterparts linking Billy* to Ben, and such that each person in the Ben-to-Billy* chain is better off than his predecessor, and each person in the Billy-to-Ben* chain is no better or worse off than his predecessor.

[17] Some philosophers would deny this – see Roberts (1998) for example. My own view is that the conclusion is over-determined – see ** (2007). We have independent reasons for thinking that we do wrong by choosing to conceive unhealthy children.

Up-Slope Cross-Morphing


W_SaveBen
W_1 …
… W_n
W_SaveBilly
Ben lives on for a short while
Billy Dies Now
Ben* dies now
Billy* lives on for a long while

It follows from either version of the morphing argument that if you are decent and rational then you will favor $W_{SaveBilly}$ over $W_{SaveBen}$.

2.2 Saving the Many

Now consider:

Saving the Many

Sam and Samantha are stranded on one island, Lonely is stranded on another. Their prospects, if saved, are all equally good. What should you do?

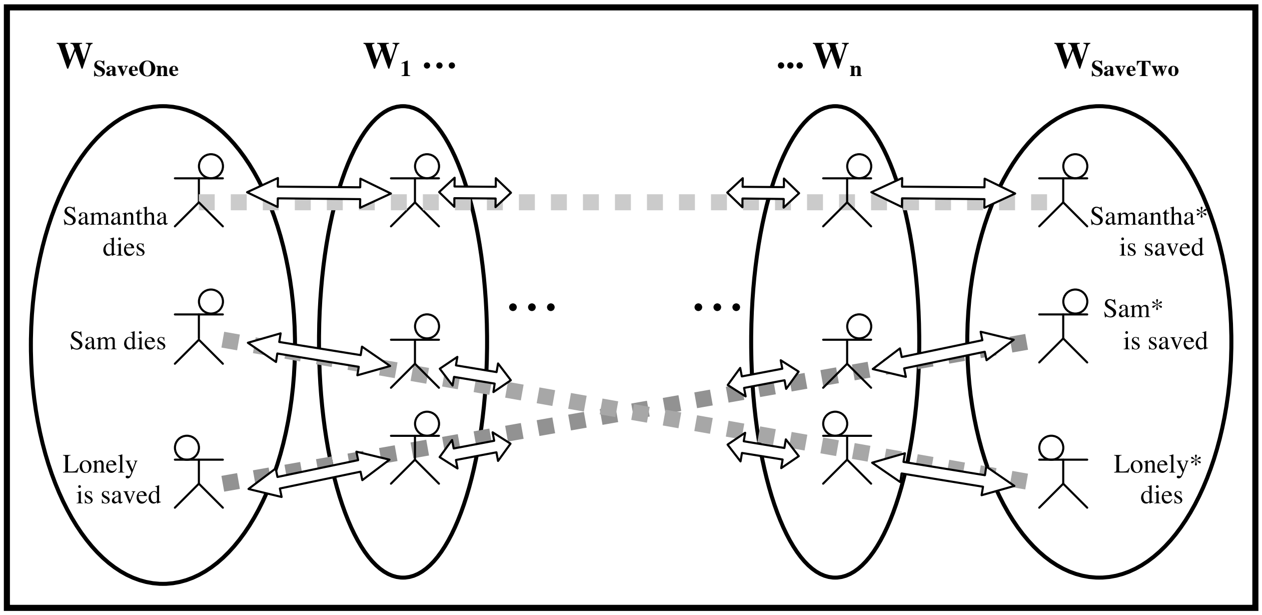
Here (thinking in the counterpart theoretic way) the two relevant worlds are:

$W_{SaveOne}$ In which Lonely is saved. Sam and Samantha die.

$W_{SaveTwo}$ In which Sam* and Samantha* are saved. Lonely* dies.

And we can construct an *up-slope selective cross-morphing sequence* of intermediary worlds, through which there runs one chain of counterparts linking Samantha to Samantha*, another linking Lonely to Sam*, and another linking Sam to Lonely*. Each person in the Samantha-to-Samantha* chain is better off than her predecessor. Each person in the Sam-to-Lonely* chain is no better or worse off than his predecessor. Each person in the Lonely-to-Sam* chain is no better or worse off than his predecessor.

Up-Slope Selective Cross Morphing



It follows from either version of the morphing argument that if you are decent and rational then you will favor $W_{SaveTwo}$ over $W_{SaveOne}$.

2.3 In General

I take it that you have grasped the general idea. Whenever there is a group of people whose interests I should be taking to heart, to the extent that I should favor one state of affairs over another when it is better for all of them, and whenever there is a pairing relation between members of the group in state of affairs S_1 and members of the group in state of affairs S_2, such that each member of the group in S_2 is better off than his pair in S_1, we can (thinking counterpart theoretically) construct a morphing sequence that links each member of the group in S_1 to his pair in S_2 via a chain of increasingly well off counterparts. It follows, given either version of the morphing argument, that if I am decent and rational then I will favor S_2 over S_1.

3. IMPLICATIONS OF THE MORPHING ARGUMENT

3.1 Utilitarianism and Unrestricted Inter-Personal Aggregation

Where does this leave us? – With some inter-personal aggregation, certainly, but with full-blooded utilitarianism? No. For one thing, there are many cases in which utilitarians would have us aggregate inter-personally, but in which the requisite pairing relation between the states of affairs we are in a position to bring about does not exist. Suppose, for example, that I can impose small costs upon the very well-off (by taxing them or what-not) and thereby bring big benefits to everybody else. In this case the very well-off in the state of affairs in which I do not tax them are better off than anybody in the state of affairs in which I do, so there is nobody to pair them off with. The morphing argument is silent. Or suppose, for example, that I can bring it about that one person suffers terrible harm or that a billion people suffer mild harm. In this case the one person in the state of affairs in which he suffers terribly is worse off than anybody in the state of affairs in which he does not, so again, there is nobody to pair him off with. The morphing argument is silent.[18]

For another thing, the morphing argument only applies to those actions of ours that affect a group of people such that we should favor one state of affairs over another when it is better for everybody in that group and all other things are appropriately equal. One might think (contra the spirit of utilitarianism) that we do not owe such benevolence to any old group of people. Parents owe it to their children. Life-boat captains owe it to people stranded in their vicinity. But everybody doesn't owe it to everybody.

Indeed, the morphing argument itself provides us with some grounds for thinking that there are substantive restrictions on how benevolent we should be. I will close by looking at two last kinds of case that illustrate this point. They also illustrate that the morphing argument has some rather amazing implications.

3.2 Infinite Worlds

Suppose that two worlds each contain a countable infinity of people. Suppose you can put numbers to the well-being of the people in either world. Suppose that, adding up the numbers, you get infinity either way. Might it be the case that we ought, nonetheless, to favor one world over the other? This question may seem obscure, but for act-consequentialists it is very important. For all act-consequentialists know, it may be that our world is infinitely extended in space or time. It may be that, over the course of all

[18] Can the argument be extended so as to cover cases like these? Alistair Norcross has suggested that we might be able to construct a morphing sequence linking one person, with (e.g.) a broken leg, to two people, each with (e.g.) a broken arm, in such a way that a plausible version of Personal Dominance would yield that we ought to favor each world over its successor. There would be indeterminacy about how many people were suffering, midway through this sequence. I find the suggestion intriguing, but it raises some complications. Indeterminacy about whether there is one person in Alice's vicinity or two is in significant ways unlike indeterminacy about whether Alice is bald or hirsute, male or female… and so forth.

world-history, there will be a countable infinity of people. It may be that we never have the power to affect this – whenever we face a decision problem, there will be infinitely people whatever we do. Act-consequentialists do not want to say that, if this is true, everything is morally permissible, so they need some way of ranking worlds in which countable infinities of people exist.

This is widely acknowledged to be a hard problem. What if W_1 is better than W_2 for infinitely many people, and W_2 better than W_1 for infinitely many people? If all the same people exist in both worlds and they are ordered in some natural way (by time of birth, for example) in both worlds, and, for example, W_1 is better than W_2 for the third, sixth, ninth… people born and W_2 better than W_1 for the first, second, fourth… people born, then one wants to say that W_2 is preferable to W_1. But what orderings should count as 'natural' for these purposes? And what about cases where the distribution of interests across the ordering is less clean? Various exotic ranking principles have been proposed.[19]

But I say that the problem is *really* hard, much harder than it has been acknowledged to be. For consider a first, tentative step that you might make towards ranking infinite worlds:

(Personal Dominance for Infinite Worlds)

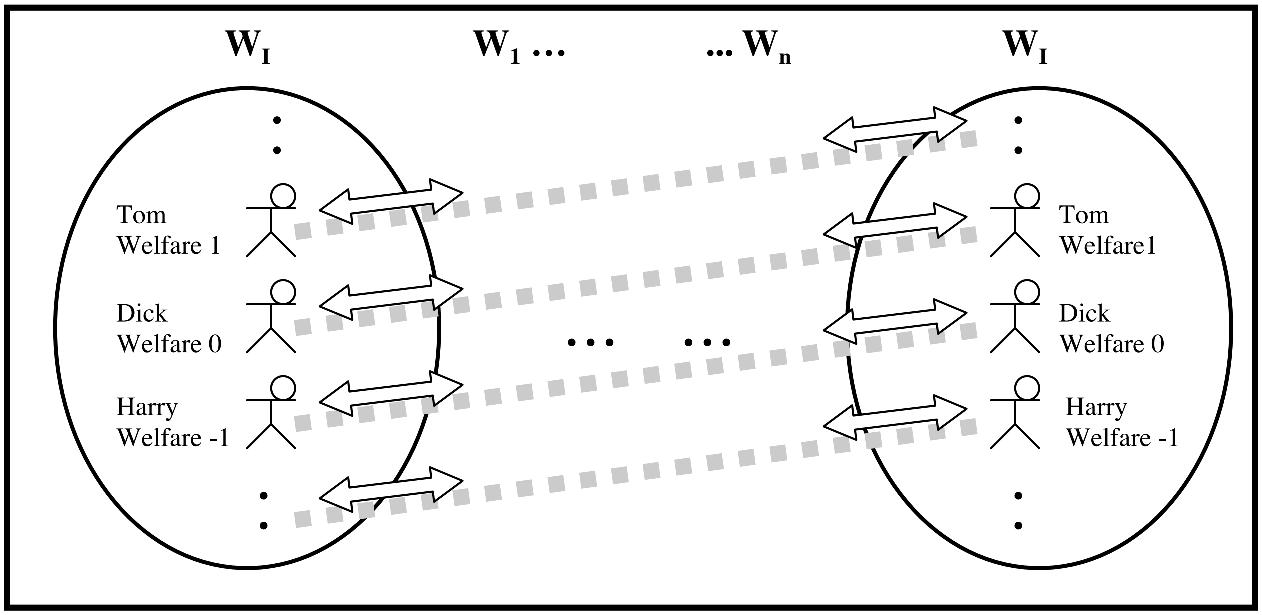
If all the same people exist in infinite worlds W_1 and W_2, and W_1 is better for all of them than W_2, then we ought to favor W_1 over W_2.[20]

[19] See Kagan and Vallentyne (1997), Hawkins and Montero (2000), and Mulgan (2002) for discussions of the problem.

[20] Shelly Kagan and Peter Vallentyne suggest, in the seminal article on this topic, that for any moral theory that identifies 'locations of local goodness', we start with a 'Basic Idea': 'if w1 and w2 have exactly the same locations, and if, relative to any finite set of locations, w1 is better than w2, then w1 is better than w2.' (Kagan and Vallentyne, p.9) If we take people to be 'locations' (a natural view) then their Basic Idea entails Personal Dominance.

The morphing argument shows that you have already gone astray. Given that, for some infinite worlds, the requisite pairing relation exists between those worlds and themselves, it would follow from *Personal Dominance for Infinite Worlds* that we ought to favor worlds over themselves. To see this, consider an up-slope morphing sequence from an infinite world, W_I, in which there exist people with levels of well-being …-1,0,1,..., to itself:

Up-Slope Morphing From an Infinite World to Itself



If *Personal Dominance for Infinite Worlds* is true then, by the morphing argument, we ought to favor W_I over W_I. But that's absurd.

What is the moral of this story? First, that act-consequentialists who acknowledge that the world may be infinite without acknowledging that moral nihilism may be true have a serious problem to solve.[21] Second, that rational constraints on our preferences (that they be transitive and irreflexive) and the slippery nature of identity across states of affairs together restrict how benevolent it is possible to be. Sometimes it is irrational to prefer that things be better for everyone.

3.3 Killing

Finally, consider:

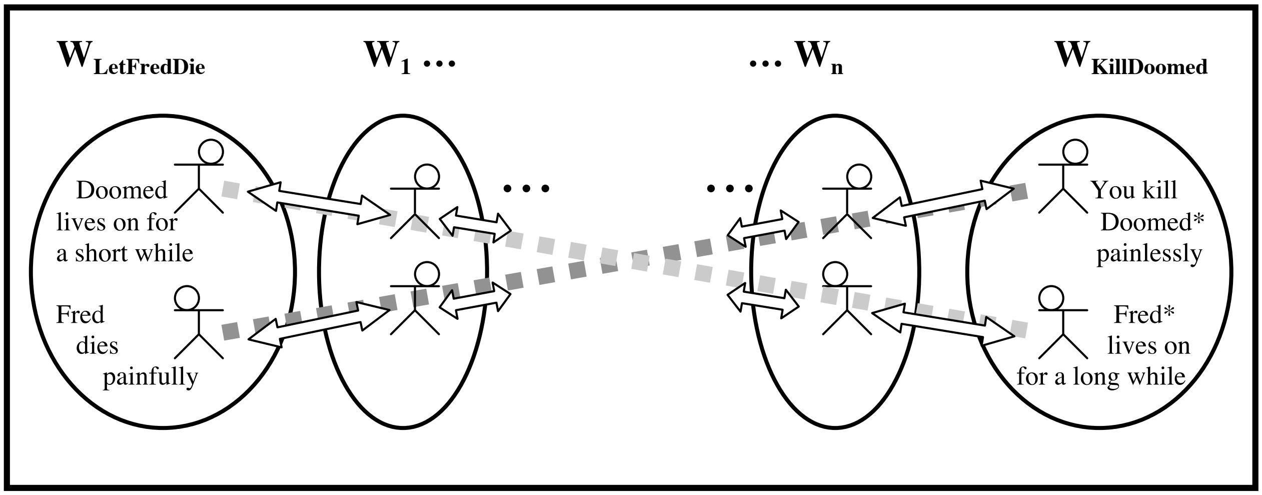
> *Killing the Sick to Save the Healthy*
> You are a doctor with two patients, Doomed and Fred. Doomed has untreatable lung cancer. His prospects are poor. Fred has an eminently treatable kidney condition, but the treatment is a transplant. You can save Fred by painlessly killing Doomed and transplanting his kidney. Should you do so?
>
> *Killing One to Save Two Others*
> You are a doctor with three patients. You can save two of them by killing one of them and transplanting his kidneys. Should you do so?

It may seem plausible to think that your responsibilities as a doctor extend to your favoring one state of affairs over another if it is much better for one of your patients and worse for none of them. It would follow from the morphing argument that if you are

[21] Notice that similar arguments can be run against *Spatio-Temporal Dominance for Infinite Worlds*, the claim that if there is more well-being at every spatio-temporal location in W_1 than W_2 then we ought to favor W_1 over W_2. Supposing that there is some flexibility in what counts as the 'same location' (as there will have to be for the principle to have any bite), we can construct up-slope morphing sequences from a world to itself, linking distinct spatio-temporal locations with a chain of pair-wise counterpart spatio-temporal locations.

decent and rational then you will kill the sick to save the healthy, and kill one to save two others. But there's a complication. Consider the relevant morphing sequence in the Killing the Sick to Save the Healthy case:

Up-Slope Cross-Morphing in the Killing-the-Sick Case


$W_{LetFredDie}$
W_1 …
… W_n
$W_{KillDoomed}$
Doomed lives on for a short while
Fred dies painfully
• • •
• • •
You kill Doomed* painlessly
Fred* lives on for a long while

Each of your patients in each world is better off than his counterpart in the predecessor world, but notice that in $W_{KillDoomed}$ you kill someone, while in $W_{LetFredDie}$ you do not. Suppose (this will simplify the point) that there is no indeterminacy about whether you kill at any world in the sequence. So there is a first world in which you kill, W_k, and in all previous worlds you do not. The person on the Fred-to-Doomed* chain in W_k is better off than his counterpart in W_{k-1}, because he dies less painfully, but you kill the person on the Fred-to-Doomed* chain in W_k and do not kill his counterpart in W_{k-1}. Perhaps, you might say, this gives you, as a Doctor, license to refrain from favoring W_k over W_{k-1}. Doctors are not obliged to practice mercy-killing. It is permissible to prefer that your patient be allowed to die a painful death rather than be killed, painlessly, by you. So the morphing argument does not apply.

Fair enough. But there is another important moral here: the claim that people of a certain kind (doctors, parents, army officers…) ought not to kill a person for somebody else's benefit is only as good as the claim that people of that kind ought not to kill a person for his own benefit. More generally (because the argument could be run for any kind of harm): if you think that there are situations in which we ought to impose local harms on people so as to benefit them in the longer term, then you must think that there area corresponding situations in which we ought to impose local harms on some people so as to benefit other people in the longer term.

This is a challenge to moral common-sense, which has it that there is a big difference between harming someone for his or her own sake and harming someone for someone else's sake. And it arises not from familiar, highly controversial consequentialist assumptions (that what matters, in both cases, is whether the outcome of our harming is worse simpliciter than the outcome of our not harming) but from seemingly innocuous assumptions about practical rationality and the metaphysics of identity across states of affairs.

3.4 Summing Up

If you are benevolent (in the sense that you want particular people to be better off) and rational (in the sense that your preferences are transitive and insensitive to irrelevant alternatives) then you will care very much about what happens, but surprisingly little about *to whom* it happens.

References

Benetar, David (1997): "Why it is Better Never to Come into Existence" *American Philosophical Quarterly* 34.

Broome, John (1999): *Ethics out of Economics*, Cambridge: Cambridge University Press

Hare, Caspar (2007): "Voices From Another World: Why We Cannot Help But Respect the Interests of People Who Do Not, and Will Never, Exist." *Ethics* 117 (3)

Harris, John (1998): "Rights and Reproductive Choice", in *The Future of Human Reproduction: Ethics, Choice and Regulation*, ed. John Harris and Soren Holm, Oxford: Clarendon

Hawkins, Joel and Montero, Barbara (2000): "Utilitarianism in Infinite Worlds," *Utilitas* 12(1), 91-96.

Hirose, Iwao (2001): "Saving the Greater Number Without Combining Claims", *Analysis* 61.4

Kamm, Francis (1993): *Morality, Mortality Volume 1: Death and Whom to Save From It*, Oxford: Oxford University Press

Kagan, Shelly and Vallentyne, Peter (1997): "Infinite Value and Finitely Additive Value Theory", *Journal of Philosophy* 94(1), 5-26.

Lewis, David (1968): "Counterpart Theory and Quantified Modal Logic", *The Journal of Philosophy*, 65: 113-26, reprinted, with postscripts, in his (1983) *Philosophical Papers Vol. 1*, Oxford: Oxford University Press

Lewis, David, (1971): "Counterparts of Persons and Their Bodies", *The Journal of Philosophy,* 68: 203-11, reprinted in his (1983) *Philosophical Papers Vol.1*, Oxford: Oxford University Press

Lewis, David (1986): *On the Plurality of Worlds*, Oxford: Blackwell

Mulgan, Tim (2002): "Transcending the Infinite Utility Debate", *Australasian Journal of Philosophy* 73(3), 401-404.

Nozick, Robert (1974): *Anarchy, State and Utopia*, Basic Books

Parfit, Derek (1984): *Reasons and Persons*, Oxford: Oxford University Press

Parsons, Josh (2002): "Axiological Actualism", *Australasian Journal of Philosophy* 80

Rawls, John (1971): *A Theory of Justice*, Cambridge, MA: Harvard University Press

Roberts, Melinda (1998): *Child versus Childmaker: Future Persons and Present Duties in Ethics and the Law*, Lanham, MD: Rowman and Littlefield.

Schiffrin, Seana (1999), "Wrongful Life, Procreative Responsibility, and the Significance of Harm," *Legal Theory* 5

Thomson, Judith (2001): *Goodness and Advice*, Princeton, NJ, Princeton University Press

Thomson, Judith (2006): "The Legacy of Principia", in *Metaethics After Moore*, Terry Horgan and Mark Timmons ed., Oxford: Oxford University Press

Woodward, James (1986): "The Non-Identity Problem," *Ethics* 96