Consequentialism in Transition

Christian Coons
Bowling Green State University

(Draft, do not cite without permission)

Abstract

Traditional consequentialism tells us that what we should do is determined by the intrinsic value of the states of affairs our actions would produce. Call this "the Standard Account." I argue that we must reject the Standard Account. The Standard Account, for example, is responsible for forcing utilitarians to choose either Parfit's "Repugnant Conclusion" or the intransitivity of the *better than* relation.

But the Standard Account is not an essential feature of consequentialism. Instead, consequentialists should assess actions in light of the intrinsic value of the *state transitions*, and not the *states of affairs* our acts produce. Evoking the intrinsic value of transitions is a nice way to address ubiquitous transitivity problems and make sense of the otherwise puzzling phenomenon that the "value" of a state seems to depend on the types of states that precede it. Talking transitions also better captures how we typically evaluate outcomes. We often speak about good or better changes, turns for the worse, and these assessments are never a simple function of the intrinsic features of the states we transition to.

I argue that our framework needs to be replaced to reflect the progress we've made in substantive axiological theory. Employing the value of state transitions can allow us to say everything we ever wanted to say, say it more simply, perspicuously, with greater explanatory power, while avoiding paradoxes that beset the conventional approach.

Suppose that at some particular time only two courses of action are available. Performing either act will change the world. The first act would produce a state of the world W_1, while the second would produce a distinct state of the world, W_2. Finally, suppose that the choice between realizing W_1 or W_2 has no further immediate or remote consequences. Traditional consequentialism tells us that what we should do is determined by the relative intrinsic values of W_1 and W_2. Call this "the Standard Account."

On its face, the Standard Account is simple, elegant, and powerful. After all, despite our non-consequentialist intuitions about particular cases, it remains difficult to see what could be said for or against any action independently of its effects. But the effects of our actions are not merely realizations of states; our acts also realize new *state transitions*. For example, in the case given above we don't merely realize W_1 or W_2, we also make it the case that the world *transitions* from its actual state to W_1 or W_2. A transition, in this sense, is a shift from one state of the world (or obtaining state of affairs) to another. For any set of available changes to the actual world, there is a corresponding set of available state transitions. When we realize a state of the world, W_n, we also realize a state transition from the actual world to W_n. If we symbolize the actual world as *"W_a"* we may symbolize such a transition as: W_a» W_n.

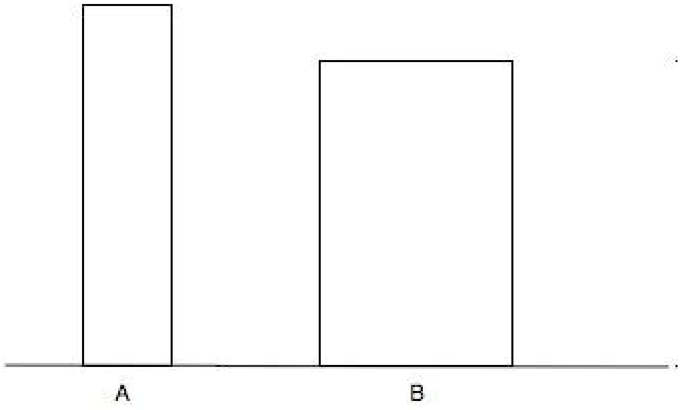
The Standard Account (hereafter "SA") ranks prospective actions by evaluating the states of affairs that would result. But we can imagine a view that ranks actions according to the value of the state transitions that would result – let us call this the "Transitional Account." On the Transitional Account, the question to ask would not be whether W_1 is better than W_2, but instead, whether [W_a» W_1] $_{is}$ better than [W_a» W_2].

In this paper, I sketch and advocate the Transitional Account (hereafter, "TA"). I will not provide a full theory of aggregation. My aim is not to provide a complete axiology for consequentialism, but rather to insist that state transitions and not states of affairs are the bearers of the type of value (i.e. intrinsic goodness/desirability) that most consequentialists are, or should be, concerned with. Thus, this paper should also be of broader interest to any non-consequentialist interested in what types of objects have value, and as we will see, of interest to anyone who wants to avoid the Repugnant Conclusion without giving up the transitivity of the *better than* relation.

I first explain why TA can make sense of the otherwise puzzling phenomenon that the "value" of a state seems to depend on the types of states that precede it. I then argue that TA better captures the manner in which most professional philosophers and the folk alike evaluate outcomes. Next, I explain why using TA finally allows us to avoid the alleged forced choice between the transitivity of the *better than* relation and accepting Parfit's "Repugnant Conclusion." I devote the rest of the paper to addressing some initial objections to, and puzzles regarding, TA. Ultimately, I hope to have at least hinted that citing the value of state transitions will allow us to say everything we ever wanted to say, and say it more simply, perspicuously, and with greater explanatory power, while avoiding the paradoxes that beset the conventional approach.

State transitions as bearers of intrinsic value

Notice that according to SA, the only non-normative knowledge required to rank prospective actions is information about the available acts/omissions and their respective effects. On TA, an additional bit of information is required; you need to know which world we are transitioning from, i.e. we require information about the actual world at the time of choice. As it turns out, almost every substantive theory of the good, (and arguably all *plausible* substantive theories of the good) cannot employ the Standard Account. This is because the putative value of a state typically depends on the states that precede it in time. For example, consider two possible states *A* and *B*:



A contains a population that is very well off, and *B* contains all the same people plus a few more. However, everyone in *B* enjoys a slightly lower quality of life than those in state *A*. Which state is better? Unless you are the very rare (and endangered) impersonal average or aggregate utilitarian, then it should be unclear to you which is better because you have no knowledge of the relation of these states to the actual world. Consider the case where the actual world (at the time of choice) is *A*, the agents in *A* are deciding whether to make a change to *B* or maintain their current state. In that case, most theorists would be hard pressed to deny that that *A* is better than *B*. After all, sustaining *A* rather than moving to *B* spares the current population from a loss of aggregate and average well-being, and an axiology that yields that *B* is better than *A* seems directly susceptible to Parfit's Repugnant conclusion. But now consider the reverse possibility that B is the actual world, and agents in *B* are considering whether to move to *A* or maintain the status quo. In that case, *B* seems much better than *A*; indeed, *A* would be tragic – a large portion of the population is annihilated in moving from *B* to *A*. Consequently, the relative value of *A* and *B* seems to crucially depend on whether we begin in *A*, *B*, or some other state.

However, notice that if we assess types of state transitions, rather than states of affairs, the variability disappears – the value of the transition would be intrinsic. Roughly, on plausible toy axiology for transitions, the transitions [*A*»*A*] and [*B*»*B*] are neutral, the transition [*A*»*B*] is somewhat bad, and the transition [*B*»*A*] is very bad. No matter which world obtains at the time of choice, the ranking {[*A*»*A*] = [*B*»*B*] > [*A*»*B*] > [*B*»*A*]} remains stable and transitive. Later, I'll explain why these features give TA a number of important and perhaps compelling advantages.

Intuitions that the apparent value of states often depends on "where you're coming from" are difficult to combine with the SA. And notice that these types of intuitions are ubiquitous. Consider the state that is, perhaps, the best candidate for intrinsic value – billions and billions of people and animals in pure bliss. Given a choice between making all the people and animals that currently exist blissful, and making a distinct same-sized (but currently non-existent) population blissful, it certainly seems better if we make the current population blissful. Indeed, it seems plausible that it would be better to maintain the currently existing world (including its current welfare distributions), than to realize a world full of blissful beings if doing so would require annihilating the current population. After all, replacing the current population with blissful beings is arguably not worth doing for anyone's sake – it would not benefit any existing person. In any case, it surely makes *some* difference whether the current population is replaced or not, and if that is so, then the value of realizing the blissful world depends on its relation to the actual world at the time of choice.

But if we allow the "value" of a state to vary according to the states that precede it in time, we can already conclude that the value of the state is not intrinsic; a state's value cannot be intrinsic and yet depend on the antecedent instantiation of independent states of affairs.[1] So, strictly speaking, axiologies that accommodate our common intuitions cannot employ the SA to rank actions. However, we can slightly modify the SA by omitting "intrinsic" and allowing that the value of a prospective state may vary according the actual history of the world.

[1] In this paper I use the expressions "intrinsic value" and "extrinsic value" in the traditional way. Specifically, a state is intrinsically valuable if and only if it is valuable in virtue of its intrinsic features (see Moore 1951, 260; Feldman 1997, 136-39 and Bradley 2002).

However, by allowing the "values" of states to vary, we then require an explanation as to why the putative "values" vary. Surely, they do not vary without some principled explanation. If they did, it would seem to render instantiations of value objectionably arbitrary, thereby undermining normative authority and the practical role anything worthy of being called "value" must have. Some theorists try to explain why the values of states vary by claiming that the variability depends on which states are accessible to the actual world at the time of choice. For example, the value of *A* may vary depending on whether we have the alternative option of realizing states *C* or *D* (etc.). But this explanation could not explain the variability of *A* and *B* above because in both cases (i.e., the move from *A* to *B* and *B* to *A*) *A* and *B* may be the only available alternatives.

Substantive axiological principles explain variability by appealing to non-intrinsic but evaluatively relevant features of states. For example, one might maintain that any state *S* is better than an alternative state *S'* if everyone does better in *S* (than in *S'*). Whether a given state meets this demand depends on the features of the actual world at the time of evaluation – we need to know who exists in the actual world and who exists in *S* and *S'* to see if everyone, in fact, does better. As we'll see, the same is true of most other axiological principles. I will argue that these principles, though often facially plausible, are better expressed as depending on evaluations of state transitions; appealing to an additional or residual "value" of states is obscure, superfluous, less explanatorily powerful, and ultimately leads to apparent paradox.

When states have variable, conditional, and non-intrinsic value, we should ask "in virtue of what?" Even if we develop plausible axiological principles that allow us to predict when and how state "values" vary, the metaphysical question remains: on what value is the conditioned and variable value of a state depend? Of course, one might say "none", but that

that would, again, seem to make the instantiation of value objectionably mysterious and arbitrary. Presumably, there is some normative explanation for the putative (and variable) value of states and the axiological principles that express and predict the variation. Such an explanation would require citing other evaluative or deontic facts that are metaphysically prior to and determine the values of states. But the consequentialist should avoid giving an explanation in terms of deontic categories such as *reasons*, or *oughts* – for the price would be reasons and/or obligations that are not a function of the value of consequences. A deontic explanation wouldn't be up to the task anyway; facts about the putative values of states vary even when there are no deontic facts to explain the variation. Deontic claims express a normative relation between an object, typically an action, and an agent or agents. They can be expressed by *ought, obligation,* and *reason* claims; consequently the truth of any such claims appears to require the presence of an agent. Evaluative facts do not work like that; a world of non-rational animals may be made better or worse even though no agent present, and hence even if no one ought, or has reason, to do anything. Because evaluative claims can be true even when no one ought, has reason, or is required to do anything, if there is any metaphysical priority between the evaluative and deontic, the evaluative is prior. But if no state-types have a fixed and invariable value, and an *evaluative* explanation is required, we must posit that some other object-type is a bearer of value. Supposing that *transitions* are the bearers of intrinsic goodness is the best bet. Let me now explain how many contemporary theories of the good are best cast as theories about the value of state transitions.

Consider, for example, Fred Feldman's influential proposal ("Desert-Adjusted Hedonism" or "Justicism") that the value of a state of pleasure is depends not only on its hedonic level but also by the recipient's desert level (Feldman 1997). On this view, it is impossible to evaluate a prospective state without looking at the states that would precede it in time. If some one deserves more or less pleasure, it is in virtue of something one has actually done. Consequently, someone's getting what one deserves is not a feature of any state of affairs. Instead it is a feature of some state transitions.

The same is true of popular "variable value" principles where upon the relative value of a state depends on the size of the actual population at the time of choice, and whether a change in welfare levels involves the "addition" of a new person (Hurka 1983, Ng 1989, Sider 1991). Application of such views, unlike SA and like TA, requires assessing a state relative to features of the actual world; specifically the value supervenes on facts about the size of the actual population and whether the bearers of well-being are new additions or not. These are not intrinsic features of any state.

TA is also best suited to capture and explain "person-affecting" axiological principles (views of this type are defended in Narveson 1973, 1978 and Roberts 1998, 2004). This approach is narrowly characterized by what Larry Temkin calls "the slogan" – an outcome can only be better (or worse) than another if it is better (or worse) for someone (Temkin 1993). More broadly, person-affecting axiology evaluates states differently depending on the whether the people in prospective states of affairs *presently* exist (Narveson 1973; Heyd 1988) or will *actually* exist (Warren 1978; Parsons 2002). Again, application of these principles requires knowledge of the actual world. These principles evaluate features of available state *transitions*, not states of affairs.

It is worth noting that most non-utilitarian putative goods, such as, *achievement*, *desire* or *preference satisfaction*, *personal development*, *learning*, *self-expression*, *emancipation*, and *autonomy*, also require realizing types of state transitions, and not merely types of states of affairs. States can manifest these values only when preceded by other states; so achievement, preference satisfaction and personal development (etc.) are again features of state transitions, not states of affairs. On these views, prospective states should have no value apart from the transitions that they would realize.

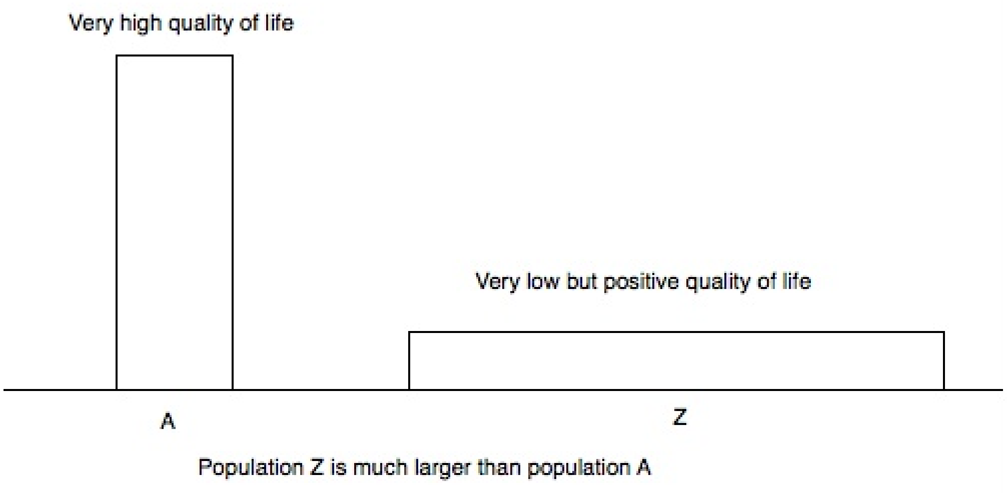
In general, our axiologies are primarily concerned with good or better changes and turns for the worse. For the vast majority of theories of the good, these assessments are never a mere function of the intrinsic features of the states we transition to. The SA does not do justice to the way we typically assess outcomes, and we would be best served by making our commitment to TA explicit. Doing so would better reflect our substantive evaluations, and avoid the obscure idea that the value of a state is somehow relativized to the actual world, while providing the metaphysical and explanatory underpinning for our axiological principles.

At this point, some readers may suspect that state transitions and not states of affairs are the appropriate relata in the *better than* relation but also insist that this merely a new way of casting obvious and old news. The Standard Account, they might claim, is an inaccurate characterization of modern act-consequentialist positions; consequentialists are already aware that they subscribe to TA, or something like it. But this diagnosis is unlikely. If it were true, they would have seen that we have a promising approach to doing what is both necessary and allegedly impossible – avoid Parfit's "Repugnant Conclusion" while preserving the transitivity of the *better than* relation.

The persistent problem

In *Reasons and Persons*, Parfit formulates the Repugnant Conclusion as follows: "For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better even though its members have lives that are barely worth living" (Parfit 1984). Finding an otherwise plausible axiology that can avoid the Repugnant Conclusion has proved to be perhaps the most vexing and difficult problem contemporary value theory faces. As Jesper Ryberg, Torbjörn Tännsjö, and Gustav Arrhenius recently put it, the question as to how the Repugnant Conclusion should be dealt with has turned into "one of the cardinal challenges of modern ethics" (Ryberg, Tännsjö, Arrhenius, 2008).

To illustrate, consider states *A* and *Z*.[2]

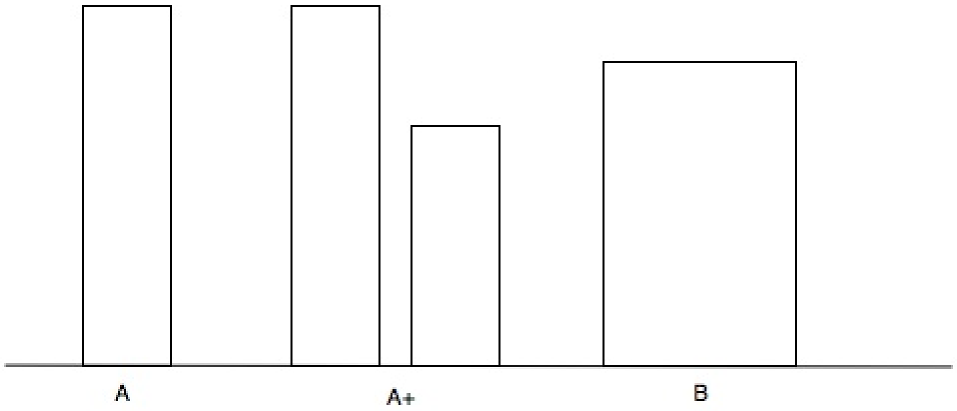

Very high quality of life
Very low but positive quality of life
A
Z
Population Z is much larger than population A

Quality of life is much lower in *Z*, but because *Z* contains many more people, aggregate well-being is higher in *Z*. By stipulation, the people in *A* lead very good lives and the people in *Z* have lives that are barely worth living.

The problem is that while many axiological principles and our conventional wisdom yield that *Z* is worse than *A*, other seemingly indispensible and innocuous steps of reasoning

[2] I should note that my diagrams are lifted from Ryberg, Tännsjö, and Arrhenius (2008)

yield that *Z* is better nevertheless than *A*. The most important version of this type of argument is Parfit's "Mere Addition Paradox" (Parft 1984, 419). The paradox is generated by first considering three possible states of the world: *A*, *A+,* and *B*.



Again, width indicates the size of the population, height indicates levels of personal well-being, and 0 along the x-axis represents the cut-off for a "life worth living." I've already used populations *A* and *B* in examples earlier, but have now added *A+*, which by stipulation, is merely a world with the same population and welfare distribution of as *A*, plus extra-people faring slightly less well. Parfit's paradox begins with a vigorous defense of the claim that *A+* is either *better than,* or *not worse* than *A,* provided that the populations are isolated from one another. In short, the idea is that the mere addition of extra worthwhile lives cannot make an outcome worse. Parfit then asks us to consider state *B* with the same people as *A+*, all leading lives worth living and at an average welfare above the average in *A+*, but lower than the average in *A*. Parfit concludes that *B* must be better than *A+* since it is better in regard to both average welfare and equality. But, if *A+* is at least *not worse* than *A* and *B* is better than *A+*, then *B* is also better than *A*. The conclusion that *B* is better than *A* already seems to be an uncomfortable result, but it gets worse. Using similar reasoning (applied to states *B*, *B*+ and *C* …etc.), we end up with the conclusion that *Z*, the state of a very large population having lives barely worth living is better than *A*.

Variations on the Mere Addition Paradox have appeared since Parfit's original formulation. Each serves to make the problem seem all the more intractable (see for example, Rachels 2004, Arrhenius 2000, and Tännsjö 2002). In each case, three populations are compared, the second appears no worse than the first, and the third better than the second, leading to the conclusion that the third is better than the first. Then, by reiteration we arrive at the Repugnant Conclusion –that *Z* is better than *A*. The problem has become so difficult that theorists insist we face the tough choice of either accepting the conclusion, or rejecting the transitivity of the *better than* relation which generates it. I don't think either option is acceptable. Rejecting transitivity appears to challenge the very concept consequentialists rely on; what is worse, it threatens to make value practically irrelevant –on a consequentialist framework *no action* would be right in the mere addition type cases (for these types of worries Broome 2004, Temkin 1987, and, Ryberg, Tännsjö, Arrhenius 2008). The solution, I suggest, requires making explicit the fact that state transitions, and not states of affairs, are the relata in the *better than* relation.[3]

TA, intransitivity, and the Mere Addition Paradox

There is a trivial sense in which TA allows us to avoid the conclusion that *Z* is better than *A* – after all TA cannot and does not evaluate states in isolation, it would be a kind of category error to claim *Z* is better than *A*. This result, I think, is actually a vindication of common

[3] I am sympathetic to the view that goodness is a function of being better than or preferable to other objects; see (Broome 1999, chapter 10). My position is meant to be consistent with that sort of view. I'm not committed to the view that some state transitions are simply good or bad independently of their being better or worse than others.

sense. After all, if the actual world were *Z*, then moving to *A* would be bad (most everyone would be annihilated!). Whereas if the actual world were *A*, then a move to *Z* really would be rather bad. Positing a shifting and relativized value for these respective states is not the best way to understand this phenomenon. That is the route that commits us to intransitivity! Rather, we can neatly capture what we want to say by claiming that the transition [*Z»A*] is bad, and the shift from [*A»Z*] is bad as well. Notice that we now know that the value of these transitions cannot be a simple function of the value of the states we shift to, because that would yield the ridiculous conclusion that both *A* and *Z* are bad.

That trivial response aside, the crucial question looms in the background. Is the sum of the state transitions from *A* to *Z* better than remaining at *A*? Were we in *A*, can we avoid the repugnant conclusion that it would be better to move to *Z* through transitions *A+,* *B*, *B*+, *C*….etc.? Although we cannot fully answer this question without a complete axiology for state transitions, let me suggest that a plausible transitional axiology ought to be able to do the trick, without making the better than relation intransitive.

First, consider the first transition [*A* »*A*+]; we can call this transition "*T1".* In my experience people's intuitions about whether this is a good transition are mixed. Some claim the transition is neutral, others claim it is slightly negative, and finally some that claim it is somewhat positive. However, I don't think anyone can plausibly maintain (without ulterior motives) that *T1* would be either *very* positive or *very* negative. Thus, let us assume the worst-plausible-case-scenario for avoiding the Repugnant Conclusion – that *T1* is slightly positive. Now let us turn to the second transition, [*A*+»*B*], call it "*T2".* Again, without ulterior motives and all else equal, it is hard to deny that *T2* is good – it constitutes an improvement in equality and average well-being. Now it initially looks like we've run head-long (again) into a version of the Repugnant Conclusion; if the move from *A* to *B* involves the two positive transitions *T1* and *T2* then presumably things go best if we move to *B*, and then *B+, C*…and so on down the line to *Z*. However, this is a mistake. On *TA*, actions are evaluated according the sum of the value of the transitions that would be realized. And when we realize *T2* (i.e. move from *A+* to *B*) a consequence is realizing the further transition, [*A*»*B*], let's call this transition "*T3"*. Analogously, suppose I start a trip in California, stop for gas in Kansas, and next travel on to Ohio. By traveling from Kansas to Ohio, I make it the case that I traveled from California to Ohio. Our unmotivated intuitions about *T3* are rather clear. In *T3* we make everyone worse off for no one's sake; peoples' well-being is sacrificed for mere mere addition. And, as we move along the alphabet, the transitions from *A* (*A* to *C*, *A* to *D*…*A* to *Z*) become increasingly bad, repugnant even. We arrive at the following results:

1. The transition, *T1*, [*A* »*A+*] is at best slightly positive.
2. The transition, *T2*: [*A+*» *B*] is positive.
3. The transition, *T3*: [*A* » *B*] is negative.
4. The sequence [*A* » *A+*»*B*] realizes *T1*, *T2* and *T3*.
5. If *T3* is more negative than [*T1 + T2*] is positive, the consequences of initiating this sequence are worse than that of staying at *A*.

Thus, any transitional axiology that predicts that *T3* is of greater disvalue than the sum of *T1* and *T2* is of positive value, avoids the conclusion that it would be better if we moved to *B*, even via *A+*. And, *a fortiori,* such an axiology would predict that it would be bad if we started in *A* and moved to *Z* (either directly, or via *A+* through *Y+*).

While it is not obvious that *T3* is of greater *dis*value than the sum of *T1* and *T2* is of *positive* value, it is certainly plausible. Therefore, a plausible axiology for state transitions will avoid the Repugnant Conclusion. And of course, the *better than* relation between these transitions would be transitive – *T1* is not as good as *T2*, and *T3* is worse than both. We'd have the following stable ordering: *T2 > T1 > T3*. Transitional theorists cannot be turned into money pumps; that point is especially obvious when one considers how they would likely evaluate any "return" transition to A from *A*+ through *Z*. Such transitions would be especially bad because they involve the annihilation of huge portions of the population.

This is not a deontic solution to the Mere Addition Paradox. For all I've said, the transitions could have been produced by natural (non-agential) forces, and effect only populations of non-rational animals without rights. Sure, *some* transitional axiology might have deontic elements, but that it not required for the solution-type I suggest. The theory does not evaluate choices directly. Like any good old-fashioned consequentialist theory, it evaluates choices by the effects that they have, and hence the states that they would realize. However, it is not states themselves that bear the values to be summed. Rather the values to be summed are values of the *state transitions* that are realized in virtue of our realizing new states. The beauty of this "solution" is that is not an *ad hoc* move to avoid paradox. Instead, it is the product of a theory that more faithfully captures our evaluation of consequences and provides the best metaphysical explanation for the types of axiological principles contemporary theorists defend.

One especially nice feature of this approach is that it always implies that it *is* better to go from *A*+ to *B*, *B*+ to *C*, *C*+ to *D* (etc.) provided that we did not reach *A*+, *B+, C*+ (etc.) by mere addition. For example, if A+ appeared, *ex nihilo*, we really should move to *B*. This implication seems desirable, but it is not captured by many other proposals.

The solution is a little surprising in that it *seems* to have the implication that moving from *A* to *A*+ would be fine, perhaps even slightly good, but once we're in *A*+ we'd better not move towards *B*. I think most people suspect that "Theory X" would somehow block the Mere Addition Paradox at the first step, not the second. But I do not think the type of view I suggest really has the implication in question. When we assess acts by their consequences, we don't just look towards their immediate consequences, but also their probable relation to future consequences. In the mere addition type cases, adding people yields a state where the best available acts will often involve "sacrificing" the well-being of the original population for no one's sake. Thus, doing what maximizes expected value may require not adding anyone in the first place. Interestingly, however, the toy axiology I've suggested does predict that it would be good to "merely add" (e.g. go from *A* to *A+)* if the populations remain causally distinct (e.g. if *B* is inaccessible to *A*+). Thus, amazingly (and perhaps suspiciously), evaluating state transitions allows us to capture every intuition that motivates the Mere Addition Paradox without being susceptible to its repugnant conclusion.

Objections and Unique Puzzles

Perhaps TA's most disconcerting feature is that it makes what was *is* or *has been actual* relevant to what it would be best to *make actual.* I've tried to mitigate this worry by earlier illustrating that it already is a feature of most axiologies. But nevertheless, it especially off-putting when we notice that every time we realize a new state, we also realize new a state transition whose "antecedent" is a state that obtained in 1000 BC – are we really supposed to assess those transitions too?! I'm tempted to say "Yeah, why not?" After all, it is unlikely that these special "distant" transitions will have features that are evaluatively relevant. For example, I suggested that perhaps *A* to *B* is negative, because it involves the loss of well-being for no one's sake. That can be a feature only of state transitions with over-lapping populations, no such relation holds in transitions which involve an antecedent in the distant past. In any case, employing TA does *not* require evaluating state transitions that include states of the distant past. Just as some axiologies discount future effects, TA theorists are welcome to discount or dismiss transitions (either asymptotically or at some critical temporal distance) according to how far the antecedent state is from the present. The simplest form of discounting/dismissing will be to discount any state transitions whose antecedent precedes the actual world at the time of evaluation. I find the idea of such "time-relativized" value perplexing; but it would have the ability to predict that we made things worse by going to A to A+ provided that we were going to end up in B, even though it remains best to move to B from A+ now that we are in A+.

Others might object that that it is misleading to call state transitions consequences of our acts. "Sure", the objector might say, "our actions can produce new state transitions, but these are not *causal* consequences of my actions. Any transitions I realize are non-causally realized only in virtue of the *states* my actions produce – so why not evaluate only the causal effects, or at least only those effects that are not conceptually tied to any past event? In other words, the evaluatively relevant consequences must be the sort of thing that could *conceivably* be realized *ex nihilo*."

I understand the complaint, but I'd like to know why only these types of consequences should count, and not others. Below, I will suggest that insisting on this standard for relevant consequences leads to implausibility. Furthermore, almost all substantive axiologies violate this constraint. No state can, by itself, constitute an *improvement*, *development*, *increase* or *progression*, or manifest *learning*, *getting what one wants or prefers*, or t*he performance of an act* without the prior realization of earlier states of affairs. Again these are features of transition types, not states. Because almost any axiology takes one or more of these features-types to be evaluatively relevant, most axiologies do not merely examine the states our acts produce, instead they evaluate those states relative to the actual world or its past –they are implicitly evaluating state transitions.

Of course, we can imagine someone insisting on the SA by saying something like this: "Contemporary axiologies be damned! A *real* consequentialist is truly forward-looking; he needs no knowledge of actual world to evaluate prospective states. What is more, some of these more traditional views can block the Repugnant Conclusion at the first step. On impersonal average utilitarianism, for example, *A* is preferable to *A+,* and so the argument never gets started."

I cannot accept such views for the simple reason that any theory that looks blindly only at available states cannot take into account the difference between effects on the currently existing population and prospective populations. Such axiologies, for a traditional consequentialist, entail that (if there is a button that makes the choice possible) that we morally must choose the annihilation of everyone currently in existence if they are to be replaced with a new population like ours, plus the addition of one extra person of above average well-being or well-being above the "critical level." Thank *goodness* that we've, so far, been successful at hiding this button from the more traditional consequentialists. While I'm merely insisting on an intuition, and not giving a proper argument, I confess I cannot do much better here.[4] At any rate, these views have very few adherents because they tend to affirm the Repugnant Conclusion directly or suffer from even more unsettling implications. Average utilitarianism, for example, features some special repugnant implications of its own, (Parfit 1984 chapter 19), including some that are very similar to the actual Repugnant Conclusion (see Sikora 1975; Anglin 1977). Furthermore, it is this sort of view that makes consequentialism especially susceptible to the charge that it objectionably treats individuals as mere means to the production of good states of affairs. By potentially obligating happy

[4] See (or ask for) my manuscript "A Defense of the Dependence Thesis" pages 10-23 for an extended defense of the view that such "mere additions" or "mere replacements" cannot be good.

folks to replace themselves with a slightly happier population, the theory arguably commits us to sacrificing ourselves to a produce states that are worth realizing for no one or nothing.

The primary objection I have heard is so far is that TA may solve the problem only by "getting things explanatorily backwards." For example, transitions to worlds full of 10 billion blissful people tend to be good *just because* states of bliss are good. And the transition from my being sad to my being happy is good because of how sadness feels and because of how happiness feels. TA, it might seem, cannot account for these facts.[5]

But TA can account for these facts. The value of the transition will be a function of the direction of transition and the intrinsic features of the relevant states. It just won't be a function of the independent "value" of the states (for example, their difference in value). The fact that there is more pleasure or less pain in the "consequent" state can (and presumably will) be relevant to the evaluation of the transition. The only thing the transitional theorist is barred from saying is that the value of the transition is determined by the differential value of the constituent states. And we would not want to say that anyway because it is incompatible with the apparent fact that a transition and its converse can be *both* of negative or positive value (Consider [*Z»A*] and [*A»Z*]).

However, the objector might now ask "how could the intrinsic features of the states matter, if those features don't constitute something of value (or disvalue)? Why care about pleasure or pain if they are not themselves valuable?" In response, we should remember that combinations of intrinsic features may be of a positive value that is not inherited from the value of their parts. For example, if we thought that pleasure is a particular combination of attitude and phenomenology, then this same style of questioning would lead us to conclude that it is not the combination that is good, but rather the phenomenology and/or the

[5] This was Stuart Rachel's worry when I presented the proposal to him in correspondence.

attitude. Thinking along these lines ultimately commits us to the value of sub-atomic particles, or at least the value of the "elements" of qualitative experience or propositional attitudes –that is absurd. Instead, we may happily maintain that states of pleasure are crucially value-relevant, without maintaining that such states are themselves good.

And even if feeling pleasure gives us an immediate insight that more of it would be good, this is an evaluation of a transition-type. But such an insight does not require the further cognitive step that more of it would be good *because* pleasure is itself good – a notion that is a bit obscure, and is in some sense practically irrelevant. Furthermore, there's no such "insight" anyway. Experience of pleasure does not provide the insight that it is good; rather (if anything) it provides the insight that such states are *good for* the subject of the experience. The well-being of or "good for" a person is not conceptually a species of the good. Being good for someone is not a normative property, although on just about every conception of the good it is a normatively *relevant* property. Notice that the mere fact that a state is *good for* something does not, by itself, support or entail that that state is *good.* A state may be good for mold, a virus, a corporation, or the devil, but the realization of that state does not thereby improve the world. Likewise, some states may be good for us, but the fact that they are does not by itself entail that they are good, or that there is even a defeasible reason to realize these states.[6]

I've tried to argue that the idea of an intrinsically valuable state of affairs plays no indispensible role in our axiological thinking. I've also suggested that no such states exist. However, something like the notion of an intrinsically valuable state of affairs can be expressed using the TA. Perhaps there are certain states that when placed in the "consequent" position of a state transitions always yield a transition (regardless of the

[6] For a further defense of this claim see Darwall (2002): 6, and Velleman (1999).

antecedent) of positive value. Similarly, we might try to identify which states if realized "from the void" (or ex nihilo) would yield transitions of value. The "consequent" states of such transitions might be worthy of the old moniker. The relationship is not symmetric, however. There is no plausible no theory for evaluating states yields a palatable theory for evaluating transitions.

Conclusion

Of course, many important questions remain. I did not defend a substantive axiology for state transitions. I merely tried to codify our intuitions, apply them to the transitional model, and show that it avoids the relevant paradoxes. Of course, some paradoxes may be unique to TA. In addition, some broadly "person affecting" principles are implicit in my toy axiology, and our everyday intuitions. But it remains unclear whether any such principles can adequately capture our intuitions about possible future populations. One also suspects that axiologies for TA which avoid the Repugnant Conclusion might entail that *adding people and then making them happy*, is better than *adding happy people.* I will not attempt to assuage those concerns here.

As I've insisted, changes can be better or worse. Whether they're good or bad does not depend only on where we end up, but also where we've come from. The current situation, I think, is uglier than most of us are willing to admit. Whether one is a consequentialist or not, we tend to think that, all else being equal, effects on individual welfare matter. And yet argument after argument seems to suggest that we can't even provide a coherent articulation. I worry that the simplicity, tone, and optimism of this paper reveals the hubris of a superficial and ignorant novice. Nevertheless, I hold consequentialism in esteem, and I think that many of the "paradoxes" uncovered in its formulation of our axiologies are not difficulties the consequentialist must shoulder. Perhaps it is not our intuitions that failed us, but the theoretical apparatus we employ to express them. We may be guilty of trying to strap the new engine-type (axiologies for which what *is* or *was* actual is relevant) onto an old frame (the Standard Model) that does not fit. The result is the apparent intransitivity of the better than relation. Our framework needs to be rebuilt to reflect the progress we've made in substantive axiological theory; for most axiologies the framework required may be the transitional account. It appears that employing the value of state transitions will allow us to say everything we ever wanted to say, say it more simply, perspicuously, with greater explanatory power, while avoiding paradoxes that beset the conventional approach. That would be a change for the better.

Works Cited

Anglin, (1977), "The Repugnant Conclusion," *Canadian Journal of Philosophy*, VII, No. 4, 745-754.

Arrhenius (2000) *Future Generations: A Challenge for Moral Theory*, FD-Diss., Uppsala University, Dept. of Philosopy, Uppsala: University Printers.

Bradley (2002) "Is Intrinsic Value Conditional?" *Philosophical Studies,* 107: 23-44.

Broome, John (1999) *Ethics Out of Economics*, Cambridge University Press.

Feldman (1997) *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy*, Cambridge: Cambridge University Press

Heyd (1988) "Procreation and Value: Can Ethics Deal With Futurity Problems?", *Philosophia* (Israel), No. 18, pp. 151-170, July.

Hurka (1983) "Value and Population Size", *Ethics*, 93, 496-507.

Moore (1951) *Philosophical Studies*, The Humanities Press.

Narveson (1973) “Moral Problems of Population.”, in M.D. Bayles, ed., *Ethics and Population*, Cambridge, Mass.; Schenkman Publishing Company Inc., 59-80

(1978) “Future People and Us.”, in R.I. Sikora and B. Barry, eds., *Obligations to Future Generations*, Philadelphia: Temple University Press, 38-60.

Ng (1989) “What Should We Do About Future Generations? Impossibility of Parfit's Theory X”, *Economics and Philosophy*, 5, 135-253.

Parfit (1976) “On Doing the Best for Our Children.”, in M.D. Bayles, ed., *Ethics and Population*, Cambridge, Mass.; Schenkman Publishing Company Inc., 100-115.

(1982) “Future Generations: Further Problems”, *Philosophy and Public Affairs*, 11, 113-172.

(1984) *Reasons and Persons*, Oxford: Clarendon Press.

Parsons (2002) “Axiological Actualism”, *Australasian Journal of Philosophy*, Vol. 80, No. 2, June, 137-147.

Rachels (2004) “Repugnance or Intransitivity: A Repugnant but Forced Choice”, in J. Ryberg and T. Tännsjö (eds.).

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

(2004) “Person-Based Consequentialism and the Procreation Obligation”, in J. Ryberg and T. Tännsjö (eds.) 2004.

Ryberg, Tännsjö, and Arrhenius (2008) "The Repugnant Conclusion", *The Stanford Encyclopedia of Philosophy*, Fall Edition, Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/fall2008/entries/repugnant-conclusion/>.

Sider (1991) “Might Theory X be a Theory of Diminishing Marginal Value?”, *Analysis*, pp. 265-71.

Sikora (1975) “Utilitarianism: The Classical Principle and the Average Principle,” *Canadian Journal of Philosophy* 5, 409-419.

Tännsjö (2002) “Why We Ought to Accept the Repugnant Conclusion”, *Utilitas*, 14, 2002, 339-359, reprinted in J. Ryberg and T. Tännsjö (eds.) 2004.

Temkin (1987) “Intransitivity and the Mere Addition Paradox”, *Philosophy and Public Affairs*, 16, 138-187.

(1993a) *Inequality*, New York: Oxford University Press.

(1993b) “Harmful Goods, Harmless Bads.”, in R.G. Frey and C.W. Morris (eds.), *Value, Welfare, and Morality*, Cambridge: Cambridge University Press, 291-324.

(1996) “A Continuum Argument for Intransitivity”, *Philosophy and Public Affairs*, 25, 175-210.

Velleman (1999) “A Right of Self-Termination?” *Ethics*, 109:606-628.

Warren (1978) “Do Potential People Have Moral Rights?”, pp. 14-30 in R. I.