PHILOSOPHICAL PERSPECTIVES

Philosophical Perspectives, 28, Ethics, 2014

UTILITARIANISM, DECISION THEORY AND ETERNITY

Frank Arntzenius
Oxford University

1. Introduction

There are two closely related problems concerning the possibility of 'infinite worlds', i.e. worlds whose histories contain infinitely many beings whose lives have non-zero utility.

The first problem is that on standard accounts of utility there are only three possible values for the total utility of infinite worlds, namely: positive infinity, negative infinity, and ill-defined. And that means that all we can say about infinite worlds is that worlds with positive infinite utility have more utility than worlds with negative infinite utility. Other than that there simply are no comparative utility facts, let alone numerical utility facts, concerning such worlds. This means that standard utilitarianism is nigh useless in classifying the merits of infinite worlds.

The second problem is that on standard accounts of expected utilities, as soon as one has non-zero credence (degree of belief, epistemic probability) that the world is infinite, the expected utilities of each of the possible acts in a given decision situation will either be positive infinity, negative infinity or ill-defined, and, more importantly, that these expected utilities will be independent of the acts in question. This seems disastrous. For even if one is not a utilitarian, i.e. even if one thinks that utilitarianism is not the correct moral theory, one should still find it extremely worrying that standard decision theory is useless as soon as one has a non-zero credence that the world is infinite.

These problems cannot easily be dismissed on the grounds that it is reasonable to have infinitesimal credence, or even zero credence, that our world is infinite. Here is an argument why not. The physics jury is still out on the question as to whether space-time is finite or infinite in extent, spatially and/or temporally. It would thus seem incautious to have an infinitesimal credence in space-time being infinite in extent, either spatially or temporally, let alone zero credence. But if space-time is infinite in extent then it is plausible that it contains infinitely many beings whose lives have some utility. For, if certain conditions lead to life, and if these conditions have independent chances of occurring in each of infinitely many disjoint space-time regions, and if each of these chances is greater than some non-zero number r, then the chance is 1 that there is life in infinitely many of these regions: if you play infinitely often, you will hit the jackpot infinitely often, no matter how unlikely the jackpot is on each play. Now, I don't mean to suggest that I have just given an inescapable argument against having infinitesimal, or zero, credence in our world being infinite. No, I made various assumptions in my argument that one could reasonably question. I merely want to claim that, prima facie, it seems implausible to dismiss the possibility that our world is infinite out of hand. And if that is right then we have very good reason to tackle the problem of the utility of infinite worlds.

I will start on this project by getting more precise about what the problem is with infinite worlds. Then I will evaluate solutions proposed by Peter Vallentyne and Shelly Kagan (Vallentyne and Kagan 1997) and by Nick Bostrom (2011), which have in common that they are trying to solve the problem in terms of utilities, rather than in terms of the expected utilities. I will end by proposing my own solution which is in terms of expected utilities.

2. The Problem with Infinite Worlds in More Detail

Suppose a world contains a (countable) infinity of people P_i whose lives have non-zero utilities $U(P_i)$. There are six types of such worlds.

Type 1. Each of the utilities $U(P_i)$ is finite, positive and greater than some non-zero number r. Then the sum of these utilities must 'diverge to positive infinity'. For the infinite sum $\Sigma U(P_i)$ is greater than the sum $r + r + r + r \ldots\ldots$, and this is greater than any finite number. I will simply say that this sum equals 'positive infinity', but the mathematically scrupulous should bear in mind that I am not claiming that 'positive infinity' is a number, rather, when I say 'the sum equals positive infinity' I just mean that the sum diverges to positive infinity. On standard accounts of total utility the total utility is just the sum of the individual utilities. So on standard accounts of total utility the total utility of such worlds equals positive infinity.

Type 2. Of course, similarly, there are worlds whose total utility equals negative infinity.

Type 3. Each of the utilities is positive and finite or each is negative and finite, and their sum converges to a finite number. For instance, suppose that one can order the people in such a way that their utilities are 1, 1/2, 1/4, 1/8, 1/16, ... Their sum converges to 2, since the partial sums get closer and closer to 2. So the total utility of this world equals 2. Worlds such as this world are not a problem from the point of view of total utility comparisons. However, presumably, our world is not likely to be such a world if our world in fact contains infinitely many lives with non-zero utility. It would take a rather unlikely systematic diminution of utility values to make sure that infinitely many of them add up to a finite number.

Type 4: There are infinitely many positive, finite-valued individual utilities, and infinitely many negative, finite-valued individual utilities. Generically, in such a case, what the sum of utilities converges to, or diverges to, depends on how one orders the utilities in question. For instance, suppose all lives either have utility 1 or utility −1, and that there are infinitely many of each of these two values. Then one can order them in the following three ways (as well as many other ways):

(a) $1, -1, 1, -1, 1, -1, 1, -1, \dots.$
(b) $1, 1, -1, 1, 1, -1, 1, 1, -1, \dots.$
(c) $-1, -1, 1, -1, -1, -1, 1, -1, -1, 1, \dots.$

The first of these sums does not converge to any value: its partial sums keep flipping between 1 and 0. The second of these sums diverges to positive infinity (it keeps adding 1 every 3 steps), and the last of these sums diverges to negative infinity (it keeps subtracting 1 every 3 steps). That is to say, if we do not make use of some preferred ordering, the total utility is ill-defined. If we do make use of a preferred ordering, the total utility is either positive infinity, or negative infinity, or ill-defined, depending on what the preferred ordering is.

Type 5. I said 'generically' in the above case because there are cases in which there are infinitely many positive utilities and infinitely many negative utilities and nonetheless what the sum converges to, or diverges to, is independent of the order in which one sums the utilities. This is so if and only if either the set of all the positive utilities converges to a finite (positive) number, or the set of all the negative utilities converges to a finite (negative) number. But, again, it does not seem plausible that our world is such a world if in fact our world contains infinitely many positive utility lives and infinitely many negative utility lives.

Type 6. At least one of the individual terms is 'positive infinity'. If all the other terms sum to a finite number, or to positive infinity, then the total is positive infinity. If this is not so, the total is ill-defined. The same goes, mutatis mutandis, if at least one of the individual terms is negative infinity. I will set aside such worlds (and hence also to the issue as to what sense one can make of an individual utility being 'positive infinity' without analysing it in terms of some infinite sum of finite utilities.) I will do so partly because it seems implausible that we live in a world which cannot be partitioned into countably many finite utility parts, and, more importantly, because there are enough problems concerning worlds which can be so partitioned.

I should also note that the existence of positive and negative utilities brings up another problem. I have so far assumed that the individual terms in the sum correspond to the utilities of the lives of people. But one might take the individual terms to be something different, say, the utilities during stretches of time. Now, the fact that the sum converges, or diverges, or is ill-defined, given one way of dividing the world up into individual finite utilities does not mean that it need do the same on another way of dividing the world up. For instance suppose that when we divide the world up into one-year stretches of time we get the sequence of utilities 1,−1, 1,−1, 1,−1,. . . .. Then it follows that when we divide the world up into two−year stretches of time we get sequence of utilities 0, 0, 0, 0, 0, 0,. . . . And these sequences have different divergence/convergence/ill-definedness features.

The upshot of the above is a precisification of the claim that on standard accounts of total utility the utility of infinite worlds is either positive infinity, negative infinity or ill-defined. The conclusion remains that there is next to nothing to say about the relative utilitarian merits of infinite worlds.

One might query an assumption that I made in the above, namely the standard assumption that utilities are additive across the lives of people. (For a defence of the view that utilities are additive across lives see Broome 1991 and Broome 2004.) If utilities are not additive, then it does not follow from the fact that the sum of utilities of all lives in a given world diverges to infinity that the total utility of that world is not finite. In order to accommodate the possibility that utilities are not additive one could re-define the notion of an 'infinite world' as follows: a world is infinite if and only if there exists a sequence of finite-sized parts P_i of that world such that every finite-sized part of the world is part of an element of the sequence P_1, P_1VP_2, $P_1VP_2VP_3$, ... ('V' stands for mereological fusion, or for set theoretic union, whichever one prefers), and the sequence of utilities $U(P_1)$, $U(P_1VP_2)$, $U(P_1VP_2VP_3)$, ... either diverges to (positive or negative) infinity, or neither diverges to infinity nor converges to a finite number. However, accommodating the possibility that utilities are not additive would make this paper even harder to read than it already is, and would not add anything of significance, so I will throughout this paper make the assumption that utilities are additive.

Finally, one might claim that the problem of infinite worlds is a pseudo-problem since one need only be concerned with an evaluation of the utility of the part of the world that one can influence, and since this generically will only be a finite part of the world, there is no real problem with infinite worlds. Indeed, I think that this is a very reasonable view to take. However, since it is not immediately clear how to make such a view precise I will regard this view as an attempted solution to the problem of infinite worlds, rather than as a demonstration that there is no such problem, and I will postpone discussion of such a solution until later on in this paper.

Now let me precisify the problem in the case of expected utilities. The expected utility of the world conditional upon an act A is $\Sigma_i Pr(W_i/A)U(W_i)$. So, as soon as one has non-zero conditional credence $Pr(W_k/A)$ in a world W_k such that $U(W_k)$ is infinite or ill-defined, the sum $\Sigma_i Pr(W_i/A)U(W_i)$ will be infinite or ill-defined. To be precise: if the sum $\Sigma_i Pr(W_i/A)U(W_i)$ contains a term $Pr(W_k/A)U(W_k)$ that is ill-defined, because $Pr(W_k/A)$ is non-zero and $U(W_k)$ is ill-defined, or if it contains two terms $Pr(W_k/A)U(W_k)$ and $Pr(W_l/A)U(W_l)$ that have opposite signs of infinity, then the sum $\Sigma_i Pr(W_i/A)U(W_i)$ will be ill-defined. Otherwise it will equal positive infinity if all the infinite terms are positive infinity, or negative infinity if all the infinite terms are negative infinity.

Moreover, suppose that, in some decision situation, one has a non-zero conditional credence $Pr(W_k/A)$ in some world W_k with infinite or ill-defined utility $U(W_k)$. As we have just seen it follows that the expected utility of act A will be infinite or ill-defined. Unfortunately, if that is so, then, generically the expected utilities of all the other acts that are possible in that situation will be the same as the expected utility of act A. Suppose e.g. that for some possible act A in some situation one has non-zero conditional credences $Pr(W_k/A)$ in each of some set of possible worlds $\{W_k\}$ whose utilities $U(W_k)$ all equal positive infinity and one also has non-zero conditional credences $Pr(W_l/A)$ in each of some other set of possible worlds $\{W_l\}$ whose utilities $U(W_l)$ all equal negative infinity. Then, as we have seen, the expected utility of act A is ill-defined. Now, the expected utility of another act B could only be different from that of A if conditional upon act B one has credence 0 either in all the worlds $\{W_k\}$ or in all the worlds $\{W_l\}$, or in both. Assuming that each life in the relevant worlds has finite utility this means that act B would have to affect the utility of infinitely many lives in such a way that it 'removes' *with probability 1* either all the worlds which have positive infinite utility, or all the worlds which have negative infinite utility. But it seems unlikely that one will ever be in a situation such that one can affect the utility of infinitely many lives with such certainty.

All in all it should be clear that utilitarianism has a problem comparing the merits of infinite worlds, and that decision theory has a problem comparing the merits of actions when one has a non-zero credence in infinite worlds. Let's finally start tackling these problems.

3. Vallentyne and Kagan's Dominance-Based Approach

Let me start showing that the problem with the utility of infinite worlds does not depend on the technicalities of summing infinitely many quantities, since it can be seen to arise from the conflict between two prima facie plausible symmetry principles and a prima facie plausible dominance principle:

(1) Permutation of Times: If worlds W and W' contain exactly the same individuals p_i with exactly the same lives and exactly the same utilities $U(p_i)$, except that some of the people live at different times in the two worlds, then U(W) = U(W')

(2) Permutation of People: If worlds W and W' have exactly the same utilities at all the same times, but some of the persons having those utilities are different in W and W', then U(W) = U(W')

(3) Dominance: If W and W' have exactly the same people at exactly the same times and some of those people have higher utility in W than in W', and the rest of the people have the same utility, then U(W) > U(W')

It would seem that the first two principles have to be valid according to utilitarianism, no matter what sense one can make of the utility of infinite worlds, since it merely expresses neutrality with respect to people and locations. It would also seem that the third principle has to be correct according to utilitarianism: if there ever is a case in which the utility of one world is more than that of another, surely it is when we have such straightforward dominance. Unfortunately however, these principles are jointly inconsistent.

Consider the following three worlds:

W: persons a, b, c, d, e, f, ..., whose lives occur in the temporal order corresponding to the listing, with utilities 0, 1, 0, 1, 0, 1, ...
W': persons a, c, e, b, g, i, k, d, ... with utilities 0, 0, 0, 1, 0, 0, 0, 1, 0 . . . . . I.e. all persons have the same utilities as in W, but their lives occur in a temporal order such that utility 1 occurs only every 4th period of time.
W'': persons a, b, c, d, e, f, ... with utilities 0, 0, 0, 1, 0, 0, 0, 1, 0 . . .

By 1) U(W) = U(W').
By 2) U(W') = U(W'').
By 3) U(W) > U(W'').

This is inconsistent. Another way to put this: Utility comparisons cannot be invariant under infinite permutations (across times and across people) and also satisfy dominance.

To get out of this problem let's start by thinking of utilities as pebbles. We then see that we can redistribute the very same set of pebbles so that fewer people, or more people, (in the subset sense of 'fewer' and 'more') have a pebble. I.e. the very same total amount of utility (the very same pebbles) can be distributed in such a way so as to make fewer, or more, people happy (make fewer, or more, people have a pebble). A natural conclusion is that the permutation principles are not valid, at least not when it comes to infinite permutations. To say that permutation invariance does not hold amounts to saying that (some aspects of) the order in which utilities are distributed matters regarding the total utility of a world.

Peter Vallentyne and Shelly Kagan (Vallentyne and Kagan 1997) have developed a dominance based approach to the comparative utility of infinite worlds which makes use of the order in which utilities are distributed across locations in worlds. Their basic idea is as follows. Suppose that one has two temporally ordered histories W and W' of utilities such that it makes sense to speak of the same location in time in W and W'. They then say that the total utility of W is larger than that of W' if for any finite interval T_1 and any 'expansion T_1, T_2, T_3 . . . ..' of T_1, there is an integer n, such that for all m > n:

$$U(W \text{ during interval } T_m) > U(W' \text{ during interval } T_m)$$

An 'expansion of an interval T_1' is a sequence of intervals of time T_1, T_2, T_3 . . . . . , such that each interval T_j has finite length and no gaps (i.e. consists of all the times between two instants in time), such that for any n, T_n is a subset of T_{n+1}, and such that every instant whatsoever is part of some such interval. (This is slightly different from their principle SBI2, but for current purposes the differences do not matter.[1])

For example, consider the following two histories of utilities where each number indicates the utility during some fixed period of time, say, one year.

W: . . . . , 5, 1, 5, 1, 5, 1 . . . .
W': . . . . , 3, 2, 3, 2, 3, 2, . . .

No matter what (finite, gapless) interval T_1 we start with, there is an expansion T_1,T_2 T_3 . . . . . such that U(W in T_j) > U(W' in T_j) for all j > 1, and there is no expansion and integer n such that U(W in T_j) < U(W' in T_j) for all j > n. This is so because for any interval of time T that is longer than 3 years U(W in T) > U(W' in T).

In a minute I will discuss some problems with Vallentyne and Kagan's basic idea, and some possible modifications in view of these problems. But I will start by indicating what I take to be the main shortcoming of their approach. The main shortcoming of their approach is that it yields only a partial ordering of the utility of worlds, and does not yield a numerical value for the utility of a world. It is not the fact that it is only a *partial* ordering that worries me. It is that it merely is an *ordering* rather than an assignment of a numerical value. For this means that it is useless from the point of view of a utilitarian decision maker. There is no quantity in their theory which we can use to multiply with probabilities in order to arrive at expected utilities. Suppose e.g. that we have a choice between an action, A, which with probability 1 leads to history W_1, and another action, B, which with probability ½ leads to history W_2 and with probability ½ leads to history W_3, where Vallentyne and Kagan's theory implies $U(W_2) < U(W_1) < U(W_3)$. Vallentyne and Kagan's theory leaves us unable to apply standard decision theory in order to decide what to do, since we have no numerical utilities to multiply with the given probabilities. Indeed, since we surely almost never are sure of the consequences of our possible actions this means that their theory, as it stands, is nigh useless when it comes to decisions.

I will eventually return to this issue, but first I want to discuss some other problems with their idea, and discuss some modifications of their idea in view of these problems. I will do so partly for completeness's sake, but mainly because my own view bears a lot of similarity to Vallentyne and Kagan's view, which means that I will have to discuss these objections and modifications anyhow.

Before I begin let me make a precisification. Vallentyne and Kagan do not merely demand that U(W in T_j) > U(W' in T_j) for all j > n, they demand that there is some non-zero positive real number r such that U(W in T_j)-U(W' in T_j) > r for all j > n. The reason is simple. Consider

W: 1, ½, ¼, 1/8, 1/16, ...
W’: 2, 0, 0, 0, 0, ...

In this case it follows that there is an integer n such that U(W in T_j) < U(W’ in T_j) for all j > n, no matter how we choose our initial T_1 and corresponding expansion. But it is fairly plausible to say that U(W) = U(W’) on the grounds that 1+1/2+1/4+1/8+1/16+...... = 2 by the standard limit definition of a countable sum. I have no quibble with this precisification.

Now let me turn to problems and modifications. (Of the below problems and modifications Vallentyne and Kagan mention only the second and third problems and the first and second modifications.)

First problem. Consider the following two histories of utilities

W: ...., 0, 0, 0, 0, 1, 1, 1, 1,...
W’: ..., 0, 0, 0, 0, 0, 1, 1, 1,...

By Vallentyne and Kagan’s dominance principle U(W’) < U(W). However, it also seems intuitive that the two worlds have equal utility, since W’ differs from W merely by the addition of a period -e.g. a day of limbo during which there are no beings, or the addition of a period during which the beings have a total utility equal to 0- and surely that does not matter regarding total utility of a world. A bit later on I will discuss a modification of their theory which addresses this problem. For now I will leave you with conflicting intuitions.

Second problem. Consider the following two histories of utilities

W: .... , 0, 0, 0, 0, 3, 3, 3, 3,.....
W’: .... , 1, 1, 1, 1, 1, 1, 1, 1,....

It should come as no surprise that for any initial interval T_1 there is an expansion and an integer n such that for all j > n, U(W in T_j) > U(W’ in T_j): just expand ‘to the future (right)’ equally fast as ‘to the past (left)’ and the addition of 3’s will eventually outstrip any initial deficit. But, more surprisingly, it is also the case that for any initial interval T_1there is an expansion and an n such that for all j > n, U(W in T_j) < U(W’ in T_j): just expand ‘to the past (left)’ three times as fast as ‘to the future (right)’. No matter where one starts, one will be adding 4 utiles at each stage to the utility of W’ (one adds 3 locations on the left and one location on the right at each stage), while one will eventually only be adding 3 utiles at each stage to the utility of W (the 3 steps to the left will eventually all be 0’s and the 1 step to the right will eventually be 3’s), which will eventually outstrip any initial deficit. So, as it stands, the utilities of these worlds are indeterminate.

First modification. Expansions have to be ‘uniform’. In the case of time the notion of uniformity is as follows: you have to expand just as fast to the future as to the past. In the above case this means that one will eventually be adding 3 utiles at each stage to the utility of W, and one will always be adding 2 utiles to the utility of W'. So with the 'uniformity' demand we get the, arguably, correct verdict that U(W) > U(W').

Vallentyne and Kagan want their theory to apply to cases in which there is more than one relevant dimension, e.g. in an infinite 3-dimensional space, or an infinite 4-dimensional space-time. In that case Vallentyne and Kagan say that a 'uniform' expansion of an initial spatial region S_1 means that at each step in the sequence one 'adds a band of constant width to the previous region'. This, of course, presupposes that the space in question comes equipped with a metric, but that doesn't seem to be a too severe a restriction. Vallentyne and Kagan do not make precise what they mean by a 'band of constant width'. But I suggest that we can take it to mean the following. A band of width w around a region R consist of all the points that are within distance w of some point in R. (This will not really look like a band of constant width if the region in question has 'deep dents', but this does not matter.)

Third problem. Vallentyne and Kagan's dominance principle presupposes that one can make sense of sameness of temporal interval, or sameness of space-time region, across pairs of worlds, for one must compare the utilities in the *same* expanding sequence of regions in the two relevant worlds. But what if we want to compare the utilities of two worlds that do not have the same locations? For instance, if one has a Lewisian metaphysics, according to which one and the same object never occurs in two distinct possible worlds, there is at best a counterpart relation between locations in different possible worlds. And what the counterpart of a location in one world is in the other world can be vague and can be context dependent on a Lewisian metaphysics. Furthermore, if the material contents of the worlds are sufficiently dissimilar there might be no plausible counterpart relations between the locations in the two worlds at all. Vallentyne and Kagan, who are aware of this problem, suggest a second modification: when assessing the relative utility of two possible worlds one should consider *all* 'admissible' counterpart relations, where a counterpart relation is admissible iff it is isometric i.e. if it preserves distances. This suggestion of theirs gives rise to yet another problem.

Fourth problem. In the first place Vallentyne and Kagan's suggestion seems to be too liberal, since it allows *any* isometry between worlds to count as an admissible counterpart relation, no matter how the matter is distributed in the two worlds. But, e.g. in Newtonian worlds there are infinitely many such isometries: any rigid spatial rotation or translation or reflection is an isometry, and any rigid temporal stranslation or reflection is an isometry. In the second place it seems to be too restrictive since on their account there are no admissible counterpart relations between curved space-times as soon as there is no isometry between the two curved space-times, which is almost always the case. For instance consider two general relativistic worlds which, intuitively speaking, differ only in the length of the nail of my index finger on January 11th 2009. Because all mass-energy produces curvature these two worlds will not be exactly isometric, so there will be no admissible counterpart relations.

In view of this problem I suggest a third modification: When identifying regions across worlds use the kind of counterpart relation that we would use in most ordinary contexts, i.e. do not identify admissible counterpart relations with isometries, rather, allow the same counterpart relations which we would normally use, namely ones which are highly constrained by the matter distribution in the relevant worlds. For (in most contexts) we surely identify locations in distinct possible worlds in large part by the occupation pattern of the locations in the worlds (rather than just their metrical relations). Now, admittedly, such counterpart relations can be vague, or even absent. However, in most cases of interest this will not affect whether a Vallentyne-Kagan dominance relation holds between the two worlds or not. More precisely: while there will be a large amount of pairs of worlds whose relative utility will be indeterminate due to the absence of, or vagueness of, the relevant counterpart relations, this will typically not be the case when we are considering worlds that according to an agent's credences are likely consequences of different actions between which the agent is deciding. That is to say, in the context of my, yet to be detailed, solution in terms of expected probabilities, the third problem will rarely, if ever, lead to indeterminacy.

As a bonus, we can use this modification to tackle our first problem. Consider the two worlds

W: . . . . , 0, 0, 0, 0, 1, 1, 1, 1, . . .
W': . . . , 0, 0, 0, 0, 0, 1, 1, 1, . . .

What we can now see we should have said about this case is that the comparative utility of these worlds depends on which temporal interval pairs in the two worlds are each other's counterparts (if any), and that that in turn depends on the detailed history of the two worlds. Whether the first utility 1 period in world W occurs 'at the same time' as the last utility 0 period in W' depends on the evolution of the contents of these worlds. If the evolution of W up until its first 1 is similar to that of W' up until its first 1, and the development after their first 1's is similar then the first utility 1 periods are each other counterparts. On the other hand, if W's development up until it's first 1 is similar to that of W's up until it's last 0 and the development of W from the first 1 on is similar to that of W's development from its last 0, with the minor difference that there is one less being at that time, then the counterpart of W's first 1 period is W''s last 0 period. If the history of W' is similar to that of W with an additional day of utility 0 added in just before the first 1, then it is vague whether the counterpart of W's first 1 is W''s last 0 or its first 1. And then there is no fact about the comparative utility of W and W'. And there is a host of other possibilities. Thus, whether U(W) = U(W') or U(W) < U(W') or U(W) > U(W'), or whether none of these relations holds, depends on the detailed contents of the histories of W and W'.

W_1	W_2	W_3
-	-	-
-	-	-
….0,0,0,0,0,0,0,…..	….1,-1,-1,-1,-1,-1, -1,….	…. -1, 1, 1, 1, 1, 1,-1,….
….0,0,0,0,0,0,0,…..	….1, 1, -1,-1,-1, 1, 1,….	…. -1,-1, 1, 1, 1,-1,-1,….
….0,0,0,0,0,0,0,…..	….1, 1, 1, -1, 1, 1, 1,….	…. -1,-1,-1, 1,-1,-1,-1,….
….0,0,0,0,0,0,0,…..	….1, 1, 1, 1, 1, 1, 1,….	…. -1,-1,-1,-1,-1,-1,-1,….
….0,0,0,0,0,0,0,…..	….1, 1, 1, -1, 1, 1, 1,….	…. -1,-1,-1, 1,-1,-1,-1,….
….0,0,0,0,0,0,0,…..	….1, 1, -1,-1,-1, 1, 1,….	…. -1,-1, 1, 1, 1,-1,-1,….
…. 0,0,0,0,0,0,0,…..	….1, -1, -1,-1,-1,-1, -1,….	…. -1, 1, 1, 1, 1, 1,-1,….
-	-	-
-	-	-

Figure 1. Three worlds

The next three problems are somewhat technical, having to do with the structure of space-time. The executive summary is that these three problems can all be solved, so feel free to skip them if you are not interested in the details.

Fifth problem. Newtonian space-time does not come equipped with a metric over space and time. It includes a metric over space, and it includes a metric over time. But it does not include a metric over space-time. In other words: it does not make sense to ask whether a given period of time is longer or shorter than a given distance in space. So there is no unique sense to be made of the notion of a 'uniform' expansion of a space-time region in Newtonian space-time. However, a plausible response to this worry is that if for a pair of worlds W and W' dominance depends on the rates at which one expands in space and in time, then there is no fact of the matter as to the relative utility of those worlds. (We already knew we were only going to get a partial ordering anyhow.) Consider, for instance three worlds whose utilities are as depicted in figure 1, where the temporal dimension is vertical and the spatial dimension is horizontal. World W_1 is a 'neutral' world, word W_2 has an 'expanding bubble of unhappiness surrounded by happiness' and world W_3 has an 'expanding bubble of happiness surrounded by unhappiness'.

In this case whether one gets dominance depends on how fast one expands in the temporal and spatial directions. It follows that for any pair of such worlds it is not true that the utility of the one world is greater or smaller than the utility of the other. This, it seems to me, is intuitively quite satisfactory. So this is not really a problem.

Sixth problem. In Newtonian space-time there is a notion of identity of spatial location across time, but in so-called neo-Newtonian space-time and in relativistic space-times there is no notion of identity of spatial location across

W_1	W_2
-	-
-	-
....1,1,1,1,1,1,1,1,1,1,.....	....0,0,0,0,0,3,3,3,3,3,....
....1,1,1,1,1,1,1,1,1,1,.....	....0,0,0,0,0,3,3,3,3,3, ...
....1,1,1,1,1,1,1,1,1,1,.....	....0,0,0,0,0,3,3,3,3,3,....
....0,0,0,0,0,0,0,0,0,0,.....	....0,0,0,0,0,0,0,0,0,0....
....0,0,0,0,0,0,0,0,0,0,.....	....0,0,0,0,0,0,0,0,0,0,....
....0,0,0,0,0,0,0,0,0,0,...	....0,0,0,0,0,0,0,0,0,0,...
-	-
-	-

Figure 2. Two more worlds

time. Now consider the following two worlds, which are a neo-Newtonian space-time analogy of our first problem.

If we take a frame of reference in which the temporal axis is vertical, and if we expand spatially just as fast to the left as to the right (it doesn't matter how fast we expand in the temporal direction), then eventually the addition of 3's will swamp the addition of 1's (no matter where we start) so that we reach the conclusion that $U(W_2) > U(W_1)$. However, in neo-Newtonian space-time there is no preferred frame of reference, i.e. any tilted straight line running up-down is as good a temporal axis as any other. In particular there is a frame of reference in which the utility distribution looks as in figure 2.

But if in this frame we expand equally fast to the left and to the right, then the fact that we keep adding more and more 0's in W_2 as we go up in time will have as a consequence that the addition of 1's in W_1 eventually swamps the addition on 0's and 3's in W_2 no matter where we start.[2] The main point here is that, as it stands, it will be frame dependent whether $U(W_1) > U(W_2)$ or $U(W_1) < U(W_2)$. But this is simply incoherent, for in a neo-Newtonian space-time there is no fact of the matter as to which is the 'correct' frame. The conclusion is analogous to the one in the previous example: it is neither the case that $U(W_1) > U(W_2)$ nor that $U(W_1) < U(W_2)$.

One might balk at this claim of indeterminacy, on the grounds that in W_1 at each time (after the utilities are non-zero) there are infinitely many 1's, while in W_2 at each time (after the same time) there are infinitely many 3's, so that surely $U(W_1) < U(W_2)$. Now, in the first place this kind of loose talk, which seems to involve summing infinitely many numbers, is not to be trusted. In the second place, notice that in my second depiction of the situation, at each location in space there are infinitely many 1's in W_1 while in W_2 at each location in space there are infinitely many 0's and at most finitely many 3's, which strongly

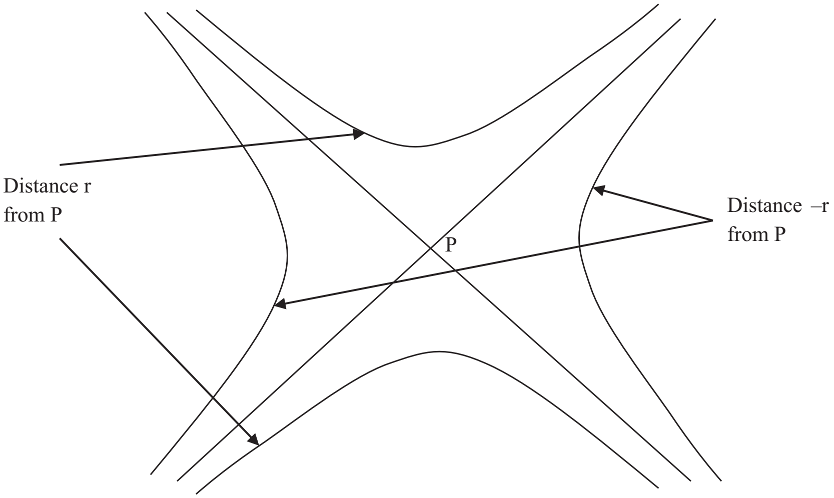

Figure 3. A double hyperbola

suggests that $U(W_1) > U(W_2)$. The conclusions of these two intuitive arguments are inconsistent, which again suggests that neither has any force.

What is clear is that in a neo-Newtonian world one's prescription for the set of allowable expansions should not be frame-dependent. The simplest way to enforce this is to demand that if some expansion is allowed then its Galilei-boost is also allowed, and then to say that U(W) > U(W') iff it is so according to all allowable expansions.

Seventh problem.

Relativity theory, while it does include a metric over space-time, and hence the ability to compare distances in space with distances in time, does not have a useful notion of all the points that are within a given distance from a given point in space-time. Let me be a bit more precise.

In Special Relativity temporal distances (time-like distances) have the opposite sign from spatial distances (space-like distances). E.g. if we take the convention that spatial distances are negative then temporal distances are positive. But there is still a sense in which a spatial distance can be said to be larger, smaller or equal to a temporal distance, namely if its absolute value is larger, smaller or equal to the temporal distance in question.

However, the set of all points that are a fixed distance away from a given point p in Minkowski space-time form a 'hyperboloid'. In a 2-dimensional Minkowski space-time this means that they form a 'double hyperbola'; in figure 3 I have indicated what such double hyperbolae look like. The problem now is that the (proper) volume of the region consisting of the points that are less than distance $|r|$ from a space-time point p is infinite no matter how small r is. (Even if one restricts attention to the points lying in the future lightcone of

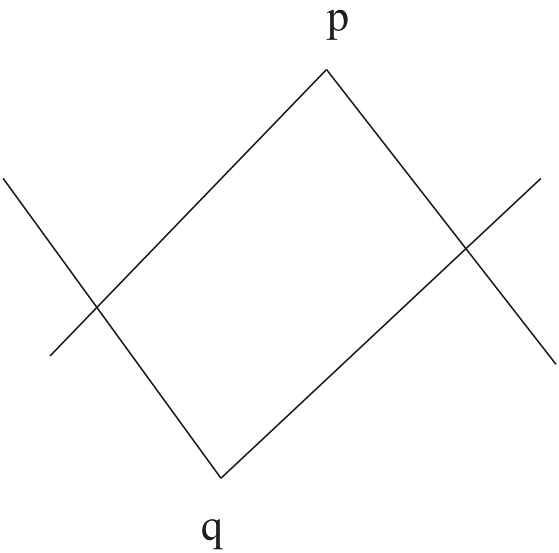

Figure 4. The double lightcone of p and q

p, i.e. the top quadrant in figure 3, this volume remains infinite.) So expanding by adding a 'band of fixed width' does not succeed in producing a finite volume region, nor, in many cases of interest, a region containing finite total utility.

What to do? Here is a natural solution, suggested to me by Cian Dorr (in conversation): one should consider expanding sequences of 'double lightcone' regions. One constructs a 'double lightcone' region as follows: one picks a spacetime point p, and a spacetime point q, and one looks at the region that consists of the overlap of the backward lightcone of p and the forward lightcone of q.[3] (Figure 4)

All in all, we have found just the one major problem with Vallentyne and Kagan's suggestion, namely that it does not yield quantities which we can multiply with probabilities, so that it becomes hard to see how one can judge actions as being utility maximising or not. And we have found a sequence of more minor problems which can be solved quite naturally, especially in the context of my — yet to be detailed — solution. Let us now look at an approach that tackles the major problem head-on.

4. The Hyperreals

Nick Bostrom (2011) has suggested that we can use the so-called 'hyperreal numbers' in order to represent the utilities of infinite worlds. Let me start by explaining what the hyperreal numbers are.

Suppose we want to have a set of numbers which includes infinite numbers of distinct magnitudes and infinitesimal numbers of distinct magnitudes, such that we still make sense of the addition and subtraction of such numbers and the multiplication and division of such numbers. One way to do this is by representing such numbers by means of infinite sequences of standard real numbers; we call those sequences the 'hyperreals'. Intuitively speaking, the sequence <1, 2, 3, 4, 5,....> represents an infinite hyperreal, since it 'grows to infinity', while the sequence <1, 1/2, 1/4, 1/8,....> represents an infinitesimal hyperreal, since it 'shrinks to 0'. We can also represent a standard real number, such as 5, as a hyperreal, namely as the infinite sequence <5, 5, 5, 5, 5....>. And we can represent the standard real number π as the hyperreal <π, π, π, π, ...>. Addition of hyperreals can then be defined 'pointwise'. For instance <1, 2, 3, 4, 5,....> + <5, 5, 5, 5, 5,...> = <6, 7, 8, 9, 10,....>. Multiplication can similarly be defined pointwise, e.g. <1, 2, 3, 4, 5,...> x < 3, 3, 3, 3, 3,...> = <3, 6, 9, 12, 15,...>. Immediately, however, a problem surfaces. For instance <0, 1, 0, 0, 1, 0, 0, 0, 1....> x < 1, 0, 1, 1, 0, 1, 1, 1, 0...> = <0, 0, 0, 0, 0,...>. Now <0, 0, 0, 0, 0,....> is our representation of the standard number 0. But <0, 1, 0, 0, 1, 0, 0, 0, 1, 0,...> is not equal to the standard number 0, i.e. <0, 0, 0, 0, 0,...>. Similarly <1, 0, 1, 1, 0, 1, 1, 1, 0....> is not equal to the standard number 0. So we have two non-zero numbers such that when we multiply the two we get 0. But according to arithmetic one cannot have two non-zero numbers that multiply to 0. This needs to be fixed.

The solution is to say that two sequences represent the same hyperreal if and only if the entries of the two sequences agree, i.e. have the same (standard real number) entry, on a 'large' set of locations in the two sequences. That is to say, each hyperreal corresponds to an equivalence class of infinite sequence of standard reals, where two sequences are in the same equivalence class iff their entries are identical at a large set of locations. In a minute I will explain how one defines the notion of a 'large' set of locations. For now let me just note that for any set S, either (exclusively) set S or its complement S^C is large, and that finite sets cannot be large. It immediately follows that either the sequence <0, 1, 0, 0, 1, 0, 0, 0, 1, 0,...> or the sequence <1, 0, 1, 1, 0, 1, 1, 1, 0....> agrees with the sequence <0, 0, 0, 0,.....> on a large set of locations, i.e. either <0, 1, 0, 0, 1, 0, 0, 0, 1, 0,...> or <1, 0, 1, 1, 0, 1, 1, 1, 0....> represents the standard number 0. So we no longer have the problem that two non-zero numbers can multiply to zero.

We can now also define one hyperreal to be greater than (resp. smaller than) another hyperreal iff it is greater (resp. smaller) at a large set of locations in the two sequences. One can show that this induces a linear ordering on the hyperreals: for any two hyperreals h_1 and h_2, we have either $h_1 < h_2$, or $h_2 < h_1$, or $h_1 = h_2$. We can also define multiplication and addition pointwise as we did before. Of course, for this to make sense it had better be the case that we always end up in the same equivalence class no matter which sequence in each of the two equivalence classes we take and multiply pointwise. But it is easy to see that that is indeed so.

To get a feel for the hyperreals let me give a few examples. In these examples, for brevity, I will denote hyperreals as infinite sequences, rather than that I will develop a special notation for equivalence classes of such sequences.

<1, 2, 3, 4, 5,... ..> is greater than the standard number 34.3, i.e. <34.3, 34.3, 34.3, 34.3, 34.3,....>, since its entries will be bigger from location 35 on. The set of all integers greater than 34 must be a large set since its complement is finite. Indeed, for the same reason <1, 2, 3, 4, 5,... > is greater than any standard number whatsoever. We therefore call <1, 2, 3, 4, 5... > infinite, or rather 'unlimited'. Similarly <1/2, 1/3, 1/4, 1/5, ...> is called 'infinitesimal' since it is smaller than any standard number. We can multiply an unlimited number by an infinitesimal number, e.g. <1, 1/2, 1/3, 1/4,....> x < 1, 2, 3, 4,....> = <1, 1, 1, 1,....> i.e. the standard number 1. So an infinite number, i.e. an unlimited number, times an infinitesimal, can equal a standard, finite, number. But it needn't. For instance <1, 1/2, 1/3, 1/4, 1/5....> x <1, 4, 9, 16, 25... > = <1, 2, 3, 4, 5,... > which is an unlimited number. Let's give the number <1, 2, 3, 4, 5,... > a name: α, and let's give <1,1/2,1/3,1/4,... > a name: ε. So α x$\varepsilon = 1$. Similarly, <1, 4, 9, 16, 25,... > = <1, 2, 3, 4, 5,... > x< 1, 2, 3, 4, 5,... > = α^2.

Now let me explain how one defines which sets of locations are 'large'. Each infinite sequence has a first entry, a second entry, a third entry, etc. So sets of locations correspond to sets of integers. The task we have thus is to define what it is for a set of integers to be large. As I indicated we want it to be the case that for any set S of integers either S, or its complement S^C, is large. And we want it to be the case that no finite set is large. We can make sure of this by means of a so-called 'non-principal ultrafilter' on the integers. A 'non-principal ultrafilter' F on the set of all integers, is a set of sets of integers such that

(1) If sets A and B are in F, then their intersection $A \cap B$ is in F
(2) If A is in F and $A \subseteq B$, then B is in F
(3) For any set A, either A is in F or its complement A^C is in F
(4) There is no set X such that F equals the set of all sets A such that $X \subseteq A$.

One can show that any non-principal ultrafilter contains every 'co-finite set', i.e. any set B whose whose complement B^C is finite. One can also show that every infinite set, and hence also the set of all integers, has a non-principal ultrafilter. (In order to prove this one needs the axiom of choice.) Since we know that there exist non-principal ultrafilters on the set of all integers, we can now say that a set of integers S is large according to a non-principal ultrafilter F iff S is in F. Of course, exactly which sets are counted as large will depend on our choice of non-principal ultrafilter F. In fact, one can show that for any infinite set whose complement is also infinite, there is an ultrafilter which contains it.

So which equivalence classes of sequences correspond to hyperreals will depend quite wildly on the non-principal ultrafilter in question.[4]

5. Bostrom's Hyperreal Utilities

Suppose that space-time is Newtonian, that there are no beings prior to some time t, which we may call $t = 0$, that there may, or may not, be beings at all times $t > 0$, but that the sum of the utilities of these beings during any finite stretch of time is finite. Then we can divide the history of the world after $t = 0$ into equal length stretches of time, namely from $t = 0$ to $t = 1$, from $t = 1$ to $t = 2$, from $t = 2$ to $t = 3$, and so on. The history of the utilities during each of these periods will then form an infinite sequence $<U_1, U_2, U_3, U_4, \ldots..>$, where U_n is a standard real number which corresponds to utility of all beings between $t = n-1$ and $t = n$. Assuming that utilities are additive, the history of the total utilities TU_n of the universe up until time $t = n$ will form an infinite sequence: $<TU_1, TU_2, TU_3, \ldots..>$ where $TU_n = U_1 + U_2 + U_3 + \ldots..U_n$.

We can then associate a hyperreal number TU with any entire history of such a world, namely the infinite sequence of standard real numbers $<TU_1, TU_2, TU_3, TU_4, \ldots..,>$. The suggestion I want to examine is that this hyperreal number TU correctly represents the total utility of the history of such a world.

Since the collection of all hyperreal numbers are linearly ordered this means that we now have a linear ordering of the total utility of all such worlds. Moreover, we can do arithmetic with the hyperreals, i.e. multiply and add and subtract and divide. So we can make sense of expected utilities, where the relevant probabilities can be hyperreals (between 0 and 1), or just standard reals (since the hyperreals include the standard reals).

In order to gain some confidence that what we are doing is not crazy, let me consider a few examples. Suppose the sum of the utilities U_n during each period n always equals 1, i.e. suppose that the history of utilities U_n is $<1, 1, 1, 1, 1, 1, \ldots.>$. Then the history of the total utilities TU_n is $<1, 1+1, 1+1+1, 1+1+1+1, \ldots..> = <1, 2, 3, 4, 5, \ldots..>$. So the total utility TU of the entire history of the world is equal to the hyperreal $<1, 2, 3, 4, 5,..>$, which is the unlimited hyperreal which we previously called α. If the history of the utilities U_n were $<2, 1, 1, 1, 1, \ldots.>$ then TU would be $<2, 3, 4, 5, 6, \ldots.>$ which equals $\alpha+1$. If the history of utilities were $<1, 1, 2, 1, 1, 1, 1, \ldots>$ then TU would be $<1, 2, 4, 5, 6, \ldots>$. Since this agrees with $<2, 3, 4, 5, 6, \ldots>$ in all but two locations, it follows that this TU also equals $\alpha + 1$. More generally if one starts with a given history of utilities U_n and modifies these utilities by finite amounts in finitely many places, the resulting total utility TU will differ from the original total utility exactly by the sum of the changes that one made. If one starts with a given history of utilities U_n and modifies it by finite amounts in infinitely many places, the difference in the TU's will be exactly equal to the hyperreal number that corresponds to the sequence of the infinitely many changes that one made (where the hyperreal that represents the sequence of changes has zeroes at the locations that one does not make any changes). If one doubles each utility U_n in some history H_1 to obtain a history H_2, then $TU(H_2) = 2xTU(H_1)$, since multiplication of hyperreals is defined pointwise. So the expected utility of history H_1 obtaining with probability 1 equals the expected utility of history H_2 obtaining with probability ½ and history <0, 0, 0, 0, 0,.....> obtaining with probability ½. That is to say, one can sum and multiply probabilities and hyperreal utilities in the usual way in order to obtain intuitively correct expected utilities.

As our final example let us consider the history of utilities <1, –1/2, –1/4, –1/8, ...>. The TU of this history is <1, 1/2, 1/4, 1/8,....>, which is an infinitesimal. One might worry about the intuitive correctness of this verdict. After all, one standardly says that the infinite sum 1–1/2–1/4–1/8..... = 0, on the grounds that this is what the sequence of partial sums converges to. Accordingly, one might maintain that the total utility of such a world should be 0 rather than infinitesimal. However, it seems to me that one could reasonably argue that the history of utilities <1, –1/2, –1/4, –1/8,....> must be better than the history of utilities <0,0,0,0,0,....> since at *every* moment t>0 the total utility TU_t of the first history up until that moment is strictly greater than that of the second history. In any case, whatever one thinks about this case, it hardly amounts to a devastating objection to the hyperreal approach. One might, for instance, decide to ignore infinitesimal differences in utilities. Our problem, after all, was to deal with infinite utilities, and other than considerations of mathematical uniformity there do not seem to be obvious reasons to want to make use of infinitesimal differences in utilities.

It might also be helpful to compare the hyperreal approach with an approach that Bostrom (2011) calls the 'value-density approach'. The value density approach defines the total value TV of the type of histories that we have been considering as the limit of the average utility per unit time as time goes to infinity: $TV = Lim_{n\to\infty}TU_n/n$. For instance, for a history with TU = <1, 2, 3, 4, 5,....> we have $TV = Lim_{n\to\infty}n/n = 1$. For a world with TU = <3, 4, 5, 6, ...> we have $TV = Lim_{n\to\infty}(n+2)/n = 1$. So worlds with distinct TU's can have identical TV. For a history with <2, 4, 6, 8,...> we have $TV = Lim_{n\to\infty}2n/n = 2$. For a history with TU = <1, 4, 9, 16,...> we have $TV = Lim_{n\to\infty}n^2/n = \infty$. For a world with TU = <1, 8, 27, 64,...> we have $TV = Lim_{n\to\infty}n^3/n = \infty$. For a somewhat bizarre world with TU = <1, –2,3,–4, 5, –6...> the TV is undefined.[5] This is so because $Lim_{n\to\infty}TU_n/n$ is undefined, since it keeps flipping between 1 and –1. It's TU is well–defined, since it is just the hyperreal <1, –2, 3,–4, 5, –6...>, or, more precisely, the equivalence class of all sequences that have the same entries as <1, –2, 3, –4, 5, –6...> on a large set of locations.

So it seems that the hyperreal approach is clearly better than the value density approach. For whenever the TV of two distinct histories is well–defined and one is greater than the other the same is true of the TU. But TU's are much more fine–grained than TV's: as we have seen there are many histories that have the same TV but distinct TU's, whether TV is finite or infinite, and our intuitions, my intuitions at least, line up with the TU's. Moreover, there is only one 'size of infinity' for TV's, represented by the symbol ∞, while there are many distinct sizes of infinite TU.

So far, so good. But there are problems with the hyperreal approach. In this case I will discuss my main objection last.

6. Hyperreal utilities depend on an ordering

Total utility TU, as I have defined it, relies on the temporal order of the utilities. But when we discussed Vallentyne-Kagan dominance we already found reasons to reject infinite permutation invariance, and to rely on some preferred ordering, so this is not an additional price to pay compared to the Vallentyne-Kagan approach. Of course this means that the hyperreal approach faces the same problem as the Vallentyne-Kagan approach concerning the question as to what the preferred ordering is. In fact this problem is more severe in the case of the hyperreal approach. For it had better give a unique result, rather than allow for a variety of such expanding sequences, since the numerical value of the corresponding hyperreal utility may very well depend on the expanding sequence in question, while Vallentyne and Kagan merely needed a set of allowable expansions for their theory to make sense since their theory only induces a partial ordering of utilities. Consider, for example, a world with the following hyperreal utility <1,–1, 1,–1, 1,–1, 1,–1, ...>. Suppose we now divide this world up into periods that are twice as long. Then we get the hyperreal utility <0, 0, 0, 0, 0, 0, ...>. The latter hyperreal utility, of course, is equal to the ordinary real number 0. However the former hyperreal is either equal to the ordinary real number 1, namely if the even numbered locations are a large set of locations according to the ultrafilter in question, or equal to the ordinary real number –1, namely if the even numbered locations are a large set according to the ultrafilter in question. I will discuss the dependence on the choice of ultrafilter in a later section. For now the thing to note is that the hyperreal utility of a world depends on how one partitions the world into periods of time, which is a problem which does not occur on the Vallentyne–Kagan approach, since it considers each of a set of 'allowable' expansions, and only yields a determinate utility comparison if dominance holds for all allowable expansions.

7. Hyperreal utilities are not countably additive

There is a sense in which countable additivity fails for the hyperreals. I will not discuss the sense in which countable additivity fails for hyperreals since it is rather technical and does not matter much for our purposes. (For details see, e.g., Benci, Horsten and Wenmackers 2011). But there is an associated worry that might seem to matter for our purposes: the existence of countable Dutch books. Here is an example of how this comes out in the case of hyperreal utilities. Suppose that the status quo is that world history W_1 is as follows: it contains exactly one being, Harry, whose life is 1 year long and has a utility of 1. Harry then gets the following offer from God: postpone his life by 1 year, but add a year to his life so that his life will have a total utility of 2. I.e. Harry has the choice between the following two histories of total utility:

$TU(W_1) = 1, 1, 1, 1, 1, 1, \ldots.$
$TU(W_2) = 0, 1, 2, 2, 2, 2 \ldots$

According to our hyperreal calculus, $U(W_2) = 2U(W_1)$, so Harry accepts the deal.

Then God comes in and offers Harry another deal: postpone his life one more year, and add another year to his life. I.e. Harry can now switch to

$TU(W_3) = 0, 0, 1, 2, 3, 3, 3, \ldots.$

Again, this is a good deal, so Harry accepts.

God then keeps making Harry offers like that. Each offer looks good to Harry so he keeps accepting them. To his chagrin Harry end up with the following world

$TU(W) = 0, 0, 0, 0, 0, 0, \ldots..$

That is to say, there exists a sequence of countably many offers, each one of which is good, no matter what other offers you accept or decline, such that accepting all of them is bad: a countable Dutch book. One might think that this is reason to reject a numerical notion of utility which is not countably additive. However, whether utilities are real or hyperreal or what-not, it is just a plain fact of life that there are countable sets of offers each of which is good no matter what other offers you take, such that the accepting all of them is bad. Failure of countable additivity of hyperreal numbers is neither here nor there when it comes to this fact of life. (For more on this lamentable fact of life see Arntzenius, Elga and Hawthorne 2004).

8. Hyperreal utilities depend on ultrafilters

Consider a world with hyperreal utility $<1,-2, 3,-4, 5,-6 \ldots>$. How big or small is this? Well, suppose that according to the ultrafilter in question the odd numbered locations are a large set. Then the hyperreal $<1,-2, 3,-4, 5,-6 \ldots>$ is the same hyperreal as $<1, 2, 3, 4, 5, 6, \ldots.> = \alpha$, since they agree on a large set of locations. On the other hand suppose that according to our chosen ultrafilter the even numbered locations are a large set. Then <1,–2, 3,–4, 5,–6 . . . > is the same hyperreal as <–1,–2,–3,–4,–5,–6, ...> = $-\alpha$. That is to say, the hyperreal utility of a given world can depend wildly on the choice of ultrafilter. And, either the evens or the odds have to be a large set, but there clearly is no unique correct choice here. So we have a problem.

Here is a less wild example. Consider the following two worlds

TU(W): <2, 4, 2, 4, 2, 4, 2, 4 . . . .>
TU(W'): <1, 1, 1, 1, 1, 1, 1, 1, . . . >.

If the odd-numbered locations are a large set according to our ultrafilter then TU(W) = 2xTU(W'), but if the even-numbered locations are a large set according to our ultrafilter then TU(W) = 4xTU(W'). So we have a problem.

Now one might hope that this kind of ultrafilter dependence only occurs for rather unusual worlds with an unlikely structure to the sequence of its utilities. Unfortunately this is not so.

Suppose that we have one person who lives forever. Suppose his utility is either +1 or –1 during each unit interval, and that there is an (independent) equal chance of each at each time. Then his total utility after n units of time develops like a 1-dimensional random walk. One can prove that with chance 1 such a random walk reaches every finite value infinitely often.

And that means that for every possible finite value there is a choice of ultrafilter such that the hyperreal utility of that world has that finite value.

We have a serious problem.

Here is an obvious suggestion. Let's say that U(W) = U(W'), U(W) < U(W'), U(W) > U(W'), U(W_1) = U(W_2) + U(W_3), U(W_1) = U(W_2) x U(W_3), and U(W_1) = r_1U(W_2)+r_2 iff this is so for *every* choice of ultrafilter. Unfortunately, this will lead to a huge amount of indeterminacy. For if W and W' are such that TU(W in R_i)>TU(W' in R_i) for each of an infinite set of regions {R_i}, and TU(W in R_j) < TU(W' in R_j) for each of some distinct infinite set of regions {R_j}, then their relative TU will be indeterminate. And this, e.g., will be so for 'almost all' (measure 1) pairs of 'random walk' worlds. Suppose, for instance, that we have two worlds fluctuating randomly (in the mathematical sense of a random walk) around some fixed average utility. Then with chance 1 these worlds will have incomparable utilities.

In fact, there is a very close relation between the 'all possible ultrafilters approach' and the Vallentyne-Kagan approach. Suppose that we have two worlds W and W' and we have fixed on a particular increasing sequence of regions <R_i> (identified across the worlds by means of a counterpart relation). And suppose that according to every ultrafilter TU(W) > TU(W'). This means that there does not exist an infinite set {R_j} such that TU(W in R_j) ≤ TU(W' in R_j) for each R_j in that set. For that would mean there was an ultrafilter according to which TU(W) ≤ TU(W'). But if there does not exist such a set then there must be an integer n such that for all k > n, TU(W in R_k) > TU(W' in R_k). (If not, one would keep finding further regions R_k such that TU(W in R_k)≤TU(W' in R_k), and hence there would be infinitely many such regions.) But that means that according to Vallentyne-Kagan dominance U(W)>U(W'). Conversely suppose that according to Vallentyne-Kagan theory U(W)>U(W'). Then for every infinite set {R_i}, TU(W in R_i)>TU(W' in R_i). And that means that according to every ultrafilter TU(W)>TU(W'). So the two theories agree on all utility comparisons.

Is there any remaining advantage to the hyperreal approach? Do we get any residual benefit from some numerical relations that we get on the hyperreal approach and not on the Vallentyne-Kagan approach? Well, the actual numerical value that we get very often depends on the ultrafilter we use. For instance consider

W: 2, 4, 2, 4, 2, 4, 2, 4. . . .
W': 1, 1, 1, 1, 1, 1, 1, 1,. . .

According to every ultrafilter TU(W)>TU(W'). However, whether TU(W) = 2TU(W') or TU(W) = 4TU(W') depends on which ultrafilter we use. And that means that we have nowhere near enough ultrafilter independent numerical relations to be able to apply standard decision theory. So there is no genuine advantage to the hyperreal approach once we admit that we need to get rid of ultrafilter dependence. We are back where we started.

Let's slowly start moving towards my solution.

9. A causal approach

Bostrom (2011) suggests, and argues against, the idea that we may restrict the evaluation of utilities to the 'changes we may affect', to our 'causal sphere of influence'. My main objection to this idea, which is different from Bostrom's objections, is that it is an unclear hybrid of an attempt to give a method for evaluating the utilities of (infinite) worlds, and an attempt to give a method for evaluating the expected utilities of acts in (infinite) worlds. Moreover, suppose that we are uncertain as to what the range of our 'causal sphere of influence is'. What is the relevant 'causal sphere of influence' then? The range that it in fact has? The biggest range that the agent in question has non-zero credence in? The union of all the ranges that the agent has non-zero credence in? Won't the last two most plausibly be infinite? Furthermore, the notion of causation is unclear enough that I would like to avoid appeal to it, and would much prefer only to need to appeal to the subjective probabilities of utilitarian agents, as will be the case in my solution. Finally, this idea, as it stands, does not help very much: for instance, in Newtonian worlds, and in relativistic worlds with an infinite future, on the most natural understanding of what our causal sphere of influence is, it is infinite, so this approach on its own does not avoid the problem of infinities.

10. Expected utility dominance

The basic idea of my solution is simple: don't try to come up with a numerical quantity that represents the utility of infinite worlds, rather, use the Vallentyne-Kagan approach to evaluate the *expected* utilities of decision makers in infinite worlds. For when it comes to expected utilities it is not a problem that we arrive merely at a partial ordering, rather than numerical values, for we have no need to attach numerical values to the expected utilities of acts for purposes of utilitarian decision theory. All we need is a utilitarian ranking of acts. This idea automatically includes a version of the causal approach: those parts of the world which are such that the expected utilities of the acts are the same no matter which act we perform are irrelevant regarding Vallentyne-Kagan dominance applied to expected utilities.

Let me be a bit more precise. Consider a simple case: the agent in question is certain that the world is Newtonian, and has to make a decision between acts A and B. The utilitarian decision procedure is then as follows. First consider all the worlds that are epistemically possible according to the decision maker given either A or B. For any pair of worlds W and W' identify regions across worlds by a counterpart relation. Then say that EU(A) > EU(B) iff for all allowable expansions R_1, R_2, R_3, ... of the decision region R_1 there exists an integer n such that for all k > n, EU(A in R_k) > EU(B in R_k). (And the same goes for EU(A) < EU(B) and EU(A) = EU(B), mutatis mutandis.) As discussed in section 3, an allowable expansion is one that at each time expands at the same rate in each direction in space, and at each location in space expands at the same rate in each direction of time. And, as also discussed in section 3, this suggestion can naturally be extended to the case of neo-Newtonian and relativistic space-times. That is all there is to my solution: Kagan-Vallentyne dominance w.r.t. expected utilities rather than utilities.

Let me now return to the issue of order-dependence. We have seen that Vallentyne and Kagan, and Bostrom, had to make use of a preferred ordering, of temporal intervals and space-time regions. I attempted to justify this use by pointing out that infinite permutation invariance principles conflict with an extremely plausible dominance principle in many cases. However, this was not so much a positive argument that infinite permutation invariance does not hold; rather it was an argument from despair: if we do not violate infinite permutation invariance it seems there can be no solution to the problem with infinite worlds. It is therefore interesting, and important, to note that my solution in most realistic cases only requires a notion of identity of location across the relevant possible worlds, i.e. in most realistic cases my solution has no need for a preferred ordering. Let me explain.

Consider a sequence of disjoint regions R_1, R_2, ... the union of which is the whole of space-time, and consider the sequence of expected utility differences between acts A and B: EU(A in R_1)-EU(B in R_1), EU(A in R_2)-EU(B in R_2), ... A sequence is said to be absolutely convergent iff the sequence of sums of the absolute values of the terms converge to a finite number. One can prove that any such sequence converges to the same finite value independent of the ordering of the sequence. Let me now slightly stretch the standard notion of 'absolute convergence' in order to include cases in which the sequence diverges to either positive infinity and negative infinity, and what it diverges to does not depend on the order in which one sums the sequence. This slightly non-standard notion of 'absolute convergence' amounts to the following: a sequence is 'absolutely convergent' iff either the sum of all positive terms is finite, or the sum of all negative terms is finite, or both.

Now, if the sequence EU(A in R_1)-EU(B in R_1), EU(A in R_2)-EU (B in R_2), ... converges absolutely then my solution will rank the utilitarian merits of acts A and B independent of an ordering of the regions in question: the only thing that we then need is a notion of identity of region across the possible worlds associated with the actions A and B. And this is so in a large class of cases. For instance if the actions in question only make a non-zero probabilistic difference over a finite part of space-time (according to the relevant person's subjective probabilities), and utilities are finite in finite regions. Another case in which the sequence of expected utility differences will be absolutely convergent is the random walk case that I previously mentioned. That is to say, if one's credences in development of worlds are equal to that of a random walk of utilities with equal chance of equal steps, then all expectation values over finite periods of time will be equal to the utility of the starting point, so there will be a well-defined finite difference in expected utilities no matter how one orders the locations. More generally one would expect that in the vast majority of plausible cases due to increasing uncertainty as to what the difference in the utilities will be for two different acts as one considers regions that are further and further away from the decision situation, the differences in the expected utilities will diminish quite fast as one considers regions that are further and further away, so that the expected utility difference will converge absolutely. That is to say, in the vast majority of plausible cases my solution is order-independent. Let me also emphasize that this order-independence crucially relies on the fact that we are considering *expected* utilities rather than utilities. For unless one is certain that the world is finite, or one is certain that the utilities of all epistemically possible worlds can only differ by a finite amount, inevitably there will be among these epistemically possible worlds pairs of worlds such that the sequence of their utility differences will not converge absolutely, so that neither Vallentyne and Kagan nor Bostrom will be able to classify the utilitarian merits of these worlds without making use of some preferred ordering.

Let me next say something about the fact that my solution makes use of a counterpart notion of identity of location across possible worlds. One might worry that such a reliance on a counterpart notion of identity of location is objectionable since, surely, whether a world has higher utility or lower utility should not depend on how one might identify locations across possible worlds, especially since the counterpart relation can be vague. Let me try to alleviate such worries. In the first place, if one accepts a Lewisian account of counterfactuals, the truth values of counterfactuals, and of causal claims (depending a bit on what one takes to be the connection between causal claims and counterfactual claims), also depends on counterpart relations. And this does not seem to be objectionable (to me at least). In the second place notice when one considers the epistemically possible worlds corresponding to the possible acts in a particular decision situation, the counterparts of regions among these worlds will typically be quite determinate since the patterns of occupation in space-time, especially those prior to the acts, will be extremely similar in all such worlds. That is to say, the fact that my solution only requires a notion of identity of regions across worlds which are associated with possible acts in a given decision situation gets rid of much of the indeterminacy associated with the relevant counterpart relations.

There is another worry one might have about my use of counterpart relations. One might worry that I am using the counterparts of regions rather than the counterparts of people, in order to compare the expected utility merits of actions. One might think that there is something fundamentally misconceived about figuring out the relative merits of worlds by comparing how they fare with respect to locations rather than how they fare with respect to people. Now, one reason for dealing with locations rather than people is that the structure of locations can provide us with a natural ordering which allow us to apply dominance reasoning to infinite worlds, while people do not come equipped with any obvious natural ordering. However, we have also seen that in most plausible cases my solution has no need for an ordering, so let me discuss whether in view of this it would be better to focus on people rather than regions.

Here is a way to precisify the idea of focussing on people rather than regions. We could adopt the 'weak people criterion' which says that the expected utility of act A is higher than that of B iff the sum of all expected utility differences across all people is absolutely convergent and greater than 0. This, of course, requires us to identify people across possible worlds, which we can do using a counterpart relation. It also requires us to assign utilities to people in worlds in which they do not exist, i.e. do not have a plausible counterpart. Well, let me stipulate that the utility of a person in a world in which they do not exist is zero. (Of course, one might worry how to determine where that '0' fits on the scale of living persons' utilities, but that is a problem that utilitarianism has to deal with anyhow, and has nothing in particular to do with the problem of infinite worlds.) More formally then, the 'weak people criterion' is

> $EU(A) > EU(B)$ iff $\Sigma_P(EU(P/A)\text{-}EU(P/B))$ is absolutely convergent and >0, where we are summing over all (epistemically possible) people p.

I have called this principle 'weak' because it does not yield a verdict in cases in which the sum of EU differences of people does not absolutely converge. However, just as before in the case of locations, it seems that in most plausible cases it will absolutely converge.

The corresponding 'weak location criterion' is

EU(A) > EU(B) iff $\Sigma_R(EU(R/A)-EU(R/B))$ is absolutely convergent and >0, where we are summing over all (epistemically possible) regions R.

How to choose between these criteria? Well, let us look at a simple case where the two criteria yield differing verdicts. Suppose I am certain that if I do nothing Suzy will live 50 years and have a life worth 100 utiles, and Suzy will then have her only child, Mary, at the end of her life and Mary will then live 50 years and have a life worth 100 utiles, and Mary will then have her only child, Gail, at the end of her life, and so on. But suppose I could, instead of doing nothing, perform an act A, which will make Suzy live twice as long, 100 years rather than 50 years, in a way that will increase the total utility of her life to 150 while reducing her utility per year accordingly.

In short, doing nothing leads to the following (future) 50 year periods

Suzy 100 utiles, Mary 100 utiles, Gail 100 utiles, Robin 100 utiles.....

Act A, on the other hand, leads to the following (future) 50 year periods

Suzy 75 utiles, Suzy 75 utiles, Mary 100 utlies, Gail 100 utiles, ...

By the weak location criterion doing nothing is better, assuming that the pattern of the world prior to the action determines that the temporal counterpart of Suzy's 100 utile life is the first 50 years of Suzy's 150 utile life. By the weak people criterion act A is better, assuming that it is correct to identify Suzy's child, and so on, in both scenarios. I expect that most people will think that the verdict of the weak people criterion is better on the grounds that ethics is fundamentally concerned with people, or at least, with utility bearing living creatures, not with locations. If one can make some people happier and none less happy, who cares if some regions are less happy while the others are just as happy. I myself have no strong views on this matter, and I would prefer the view that it is indeterminate which act is better when the weak people criterion yields a different verdict from the weak location criterion. But I am happy to leave it to readers to make up their own minds, especially since I doubt that we are likely to be faced with a situation in which it matters which of the two criteria we use.

Finally, one might worry that I evaluate the utilitarian merits of acts by looking at *expected* utility differences rather than actual/counterfactual utility differences. For one might think that how good or bad an act objectively is depends not on the subjective expectations of agents (ideal or not), but rather depends on the difference the act in fact made, i.e. depends on the utility difference between what in fact happens and what would have happened had any of the other possible acts been performed. My own view is that in certain contexts (various varieties of) *expected* utility differences provide the relevant moral evaluations of acts, and in certain other contexts (various varieties of) actual/counterfactual utility differences provide the relevant moral evaluations of acts. In any case, I would find it somewhat problematic if one could never make any sense of actual/counterfactual utility differences between acts. But, of course, it should be clear that the Vallentyne-Kagan criterion can be applied in many cases for an evaluation of actual/counterfactual differences. For in most such cases there will be a fairly determinate counterpart relation between locations (and between people), so that one can straightforwardly apply the Vallentyne-Kagan criterion. My main objection to the Vallentyne-Kagan criterion, after all, was only that one could not multiply it with probabilities, so that it could not be used by a utilitarian decision maker, my objection was not that it cannot be used to evaluate the comparative utilitarian merits of pairs of worlds. So, if one is in a context where the relevant evaluation is in terms of actual/counterfactual utilities, then one can use the original Vallentyne-Kagan criterion, if needed with some of the bells and whistles that I added in section 3 in order to deal with the problems I described.

11. Conclusions

The Vallentyne-Kagan approach to the comparative utility of infinite worlds has the major shortcoming that it cannot supply a quantity which we can multiply with probabilities, so that it is hard to see how a Vallentyne-Kagan utilitarian can make any decisions. It also suffers from a number of minor problems associated with the need for a preferred ordering of locations in space-time. The hyperreal approach suffers from the major problem that it becomes effectively equivalent to the Vallentyne-Kagan approach when one gets rid of the objectionable reliance on a choice of ultrafilter. In addition it suffers from minor problems which are similar to the problem that the Vallentyne-Kagan approach has with respect to the need for an ordering of locations. My own suggestion is to apply a (slightly modified version of) the Vallentyne-Kagan criterion to the expected utilities of actions, rather than to the utilities of worlds, so that the non-numerical nature of this evaluation is not a problem. In most realistic situations this will offer a definite verdict which does not rely on a preferred ordering of locations and/or persons.

Notes

[1] They demand that every initial region has an expansion such that one gets dominance for all further expansions and they do not demand that every period is eventually included. They also consider spatio-temporal orderings other than temporal orderings. I discuss spatio-temporal orderings later on in the main text.

[2] Well, actually, this depends on how fast we expand in the temporal dimension relative to the spatial dimension. The simplest way to check that there is a frame of reference relative to which $U(W_1) > U(W_2)$ is by drawing a sequence of square boxes, by adding a single 'band' of utilities around the previous box at each stage

of the sequence, and then seeing what happens to the total utilities in the expanding sequence of boxes.

3. One might instead suggest to use an expanding sequence of convex regions. However, there is no notion of convexity in General Relativity that is useful for our purposes.
4. One can also show that the arithmetic of hyperreals that one obtains in this manner is independent of one's choice of non-principal ultrafilter, in the sense that the arithmetical structures that one obtains for different choices of non-principal ultrafilter are all isomorphic, as ordered fields with zero and unit. This is good news for hyperreal arithmetic but does not help when it comes to the ultrafilter dependence of the hyperreal utility associated with a given world history. I will say much more about this ultrafilter dependence in the main text.
5. I call this world 'bizarre' because the corresponding utilities U_n are <1, –3, 5, –7, 9, –11, . . . >. This means that the difference between the utilities of consecutive periods grows without bound as time progresses.

References

Arntzenius, F., Elga, A. and Hawthorne, J. (2004). "Bayseanism and Binding", *Mind* 113 (450): 251–283.

Benci, V. Horsten, L. and Wenmackers, S. (2011). "Non-Archimedean probability", electronic preprint at http://arxiv.org/abs/1106.1524

Bostrom, N. (2011). "Infinite Ethics", *Analysis and Metaphysics* 10: 9–59.

Broome, J. (1991). *Weighing Goods*. Oxford: Blackwells.

Broome, J. (2004). *Weighing Lives*. Oxford: Oxford University Press.

Vallentyne, P. and Kagan, S. (1997). "Infinite Value and Finitely Additive Value Theory", *Journal of Philosophy* 94 (1): 5–26.