Prudential Longtermism

Petra Kosonen

petra.kosonen@worc.ox.ac.uk

Based on a paper co-authored with Johan E. Gustafsson

Longtermism

I will argue that person-affecting views lead to Longtermism.

Longtermism

Our acts' expected influence on the value of the world is mainly determined by their effects in the far future.

Longtermism is counter-intuitive. It implies that the short-term effects which we normally focus on are outstripped by the effects in the distant future.

When evaluating acts, we can often simply ignore their short-term effects and instead focus on their effects in the far future.

Longtermists might focus on, for example, preventing human extinction due to an asteroid or a pandemic.

If we adopt an additive total utilitarian view, there is a straightforward case for Longtermism.

According to total utilitarianism, an outcome is better than another outcome just in case the total amount of well-being it contains is greater.

Even a small possibility of a very large population living in the far future outweighs the importance of our acts' effects in the near future.

It is less clear whether there is a similar case for Longtermism if we accept a person-affecting view, on which an outcome cannot be better than some other outcome unless it is better for someone.

Our current acts do not only affect the number and quality of future lives, but they also affect who will exist in the future—so that each act we can perform results in different people existing in the future due to the ripple effects of these acts.

So, if it cannot be better or worse for someone to exist than to not exist, it seems that the only people we can make better off are those who already exist.

Thus, if it is certain (or almost certain) that no one alive today will be alive in the far future, then person-affecting views lead to the rejection of Longtermism.

Prudential Longtermism

But, in fact, there is a different path to Longtermism that is perfectly compatible with those views. Instead of total utilitarianism, this path appeals to

Prudential Longtermism

Prudential Longtermism is true for a person *S* if and only if our acts' overall influence on the expected prudential value for *S* is mainly determined by the effects of these acts in the far future.

If Prudential Longtermism is false for all currently existing people, then all normative views on which only these people matter lead to the rejection of Longtermism.

Prudential Longtermism depends mainly on the feasibility of different forms of life extension.

But, as we shall see, it also depends on what relation matters in survival and how we should aggregate personal value in cases of fission—that is, cases in which there are multiple individuals in the future who are all related to you (as you are now) in the way that matters for survival.

We may distinguish between different strengths of Prudential Longtermism:

Weak Prudential Longtermism

Weak Prudential Longtermism is true for a person S if and only if our acts' overall influence on the expected prudential value for S is **mostly** determined by their effects in the far future.

Strong Prudential Longtermism

Strong Prudential Longtermism is true for a person S if and only if our acts' overall influence on the expected prudential value for S is **overwhelmingly** determined by the effects of these acts in the far future.

If Weak Prudential Longtermism is true for you, then the far future matters more than the near future for your prudential value.

If Strong Prudential Longtermism is true for you, then the far future matters overwhelmingly more than the near future for your prudential value, and, for prudential concerns, you could often simply ignore our acts' short-term effects and instead focus on their long-term effects.

I will discuss whether Weak or Strong Prudential Longtermism is true for some currently existing persons, and whether this means that person-affecting views lead to (impersonal) Longtermism.

It is clear that there are things we could do such that we would have no hope of any prudential value after the short-term. So, I will look for acts and technologies that may provide a lot of prudential value in the long-term.

By performing such acts rather than the acts that offer no expectation of long-term prudential value for you, our acts have an enormous influence on your expected prudential value. And then Prudential Longtermism holds for you.

Recap

Longtermism

Our acts' ex ante influence on the value of the world is mainly determined by their effects in the far future.

Prudential Longtermism

Prudential Longtermism is true for a person S if and only if our acts' overall influence on the expected prudential value for S is mainly determined by the effects of these acts in the far future.

Anti-ageing

The case for Prudential Longtermism relies on the feasibility of extreme life extension. There are a number of ways in which we may extend our healthy lifespans.

I will start by discussing anti-ageing.

Anti-ageing is the attempt to stop, or even reverse, ageing.

Research on anti-ageing has recently made significant progress. Could anti-ageing, by itself, lead to Prudential Longtermism?

Anti-ageing

The attempt to stop, or even reverse, ageing.

If it succeeds in stopping or reversing ageing, it could, of course, significantly lengthen our lives.

But, even if we stop ageing, we may still die from other causes.

Given a 0.13 % chance of death per year (the proportion of people aged between 30–31 who died in the U.S. in 2018), your life expectancy is $1/0.0013 \approx 770$ years.

This estimate assumes that the annual background risk of death (from injury or illness) will not change, and it also does not take into account rare events, such as wars, global catastrophes or existential risks.

Is 770 years a sufficiently long life expectancy to lead to Prudential Longtermism?

Let the next 100 years constitute the short-term, and let the long-term start after that.

Let us also assume (somewhat arbitrarily) that a technology leads to Strong Prudential Longtermism if and only if your expected number of life years in the long-term is at least 100,000 times as great as your expected number of life years in the short-term.

This will be true if your life-expectancy is at least 10 million (plus 100) years, assuming that you will certainly live for 100 years.

If there is such a technology, it is plausible that our acts' overall influence on the expected well-being of some currently existing person is overwhelmingly determined by our acts' effects in the far future.

Next, let us assume that a technology leads to Weak Prudential Longtermism if and only if your expected number of life years in the long-term is greater than your expected number of life years in the short-term.

Then, assuming that the long-term does not provide opportunities for far greater or far lower welfare than the short-term, it is likely that our acts' overall influence on some currently existing person's expected well-being is mostly determined by their effects in the far future.

How high must the probability of anti-ageing working be in order for it to lead to Weak Prudential Longtermism?

With p being the constant probability of death each year if anti-ageing works (which I have assumed to be 0.13 %), we have that the expected years of life in the short-term (that is, the next 100 years) if anti-ageing works is

$$\sum_{n=1}^{100} (1-p)^n \approx 93.7$$

Let q be the probability of anti-ageing working. And assume that your current life expectancy without any new life-extension technology is 50 years (the US life expectancy at age 30 in 2018). Then, anti-ageing alone leads to Weak Prudential Longtermism if

$$\left(\frac{1}{p} - 93.7\right) q > 93.7q + 50(1-q).$$

Hence anti-ageing leads to Weak Prudential Longtermism if q, the probability of anti-ageing working, is greater than 8%. Then, your expected number of life years in the long-term is greater than your expected number of life years in the short-term.

But anti-ageing alone does not lead to Strong Prudential Longtermism.

Even assuming that anti-ageing is guaranteed to work, the expected number of life years in the long-term is less than 8 times greater than the expected number of life years in the short-term, given a 0.13 % yearly chance of death.

Of course, we may be able to decrease our yearly risk of death in the future and thereby improve our chances of survival significantly.

In order to get $100,000$ times as great the expected number of life years in the long-term as in the short-term, we need the annual risk of death to be at most one-in-10-million.

This, of course, assumes that anti-ageing works.

But since there is uncertainty about the feasibility of anti-ageing, the annual risk of death needs to be even lower for anti-ageing alone to lead to Strong Prudential Longtermism.

Recap

If we stop ageing, your life expectancy is 770 years (given a 0.13 % chance of death per year).

Anti-ageing leads to Weak Prudential Longtermism with reasonable probabilities of anti-ageing working (at least 8%).

But it does not lead to Strong Prudential Longtermism.

Weak Prudential Longtermism

Weak Prudential Longtermism is true for a person *S* if and only if our acts' overall influence on the expected prudential value for *S* is **mostly** determined by their effects in the far future.

Strong Prudential Longtermism

Strong Prudential Longtermism is true for a person S if and only if our acts' overall influence on the expected prudential value for S is **overwhelmingly** determined by the effects of these acts in the far future.

Cryonics

Cryonics is the process of storing a person's brain (or whole body) at very low temperature after their (legal) death in the hope that they may one day be revived.

One worry about cryonics is whether it can preserve memories. Many philosophers believe that psychological continuity is what matters in survival.

Psychological continuity

A future prospect is as bad as death for you unless you are psychologically continuous with someone in that future prospect.

Psychological continuity in turn consists in overlapping sequences of psychological connections.

And these connections are usually taken to be memory relations, that is, the relation of your current experiences being remembered by the person at the future time.

So, on these views, cryonics does not preserve what matters in survival if it does not preserve your memories.

Yet there are other candidates for what matters in survival.

Some people believe that physical continuity is what matters.

Physical continuity

A future prospect is as bad as death for you unless you have the same brain (or enough of the same brain) as someone in that future prospect.

Thus, cryonics could preserve what matters in survival even if it does not preserve memories (or any other psychological connections)—as long as it is possible to revive the same spatio-temporally continuous brain.

Does cryonics lead to Strong Prudential Longtermism?

Even if cryonics leads to a successful revival, it is still open to worries about fatal injuries that permanently destroy the brain after the revival.

So, even if it was possible to revive the spatio-temporally continuous brain after cryopreservation and this brain could be given a new body, that brain may still be damaged beyond the possibility of revival.

The annual risk of brain destruction (during those years in which the brain is not cryopreserved) would have to be at most one-in-10-million in order to get at least 10 million life years in expectation.

This risk might still be too low for Strong Prudential Longtermism to be true for anyone.

Hence cryonics alone (or in combination with anti-ageing) at most gives us Weak Prudential Longtermism.

Still, cryonics (like anti-ageing) might buy us time for finding better ways of extending life.

Recap

Cryonics is the process of storing a person's brain (or whole body) at very low temperature after their (legal) death in the hope that they may one day be revived.

Cryonics alone (or in combination with anti-ageing) at most gives us Weak Prudential Longtermism.

Psychological continuity

A future prospect is as bad as death for you unless you are psychologically continuous with someone in that future prospect.

Physical continuity

A future prospect is as bad as death for you unless you have the same brain (or enough of the same brain) as someone in that future prospect.

Uploading

Uploading

Uploading is the process of scanning our brains and loading the information on to computers, where our brains are then simulated.

A standard worry about uploading is whether the simulation will be conscious. A zombie simulation would not (at least on hedonism) have any well-being so it would be prudentially worthless.

Another worry is whether you would stand in the relation that matters in survival to your simulation.

The views of personal identity on which you could plausibly be identical to your simulation are reductionist views where personal identity just consists in an impersonal mental relation holding uniquely.

Here, an impersonal relation is a relation that can be completely described without mentioning people.

But, if personal identity can be reduced to an impersonal relation (such as psychological continuity) holding uniquely, it seems that we should also care about this relation when it holds from one to many (fission cases) rather than only in case it holds from one to one.

The most influential reductionist view is that psychological continuity is what matters in survival.

Psychological continuity (represented by Relation R) is the holding of overlapping sequences of psychological connectedness (represented by Relation C). Psychological connectedness is a direct psychological connection between a person at one time and a person at another time, such as the person at the latter time remembering the experiences of the person at the earlier time.

Person-stages=temporal parts of people

Psychological connectedness

Person-stage x is C-related to person-stage y $(xCy) =_{\text{df}}$ x is psychologically connected to y with the right kind of cause and x is present either simultaneously with y or earlier than y.

Psychological continuity

Person-stage x is R-related to person-stage $y =_{\text{df}}$ either xCy or yCx, or there are person-stages $z_1, z_2, \ldots, z_n$ such that either

(i) $xCz_1, z_1Cz_2, \ldots, z_{n-1}Cz_n, z_nCy$ or

(ii) $yCz_1, z_1Cz_2, \ldots, z_{n-1}Cz_n, z_nCx$.

A very long simulation

One reason to think that uploading may lead to Prudential Longtermism is that the uploads can live on for a very long time.

Yet, since prudential concern is plausibly forward (rather than backward) looking, the simulations need not have any special interest in continuing to be directly psychologically connected to you. So we may suspect that they will gradually let go of their memories of you in order to make room (in computer memory) for more useful knowledge.

More generally, it seems that each stage of the simulation would like to be remembered by the next stage, but they do not have any special interest in remembering earlier stages. So they may opt to forget earlier stages to free computer memory.

So it seems that there would be no connectedness between distant stages of the simulation (nor between your stages and those of simulations in the far future).

So, if Relation C is what matters, it seems that uploading would not lead to Prudential Longtermism in virtue of a very long-lasting simulation.

But, if Relation R is what matters in survival, it seems that, as long as the simulation is kept running, your relation to your simulation contains what matters in survival. And, if civilization survives and people have some interest in keeping the simulation running, then the simulation may run for a very long time.

Assuming that you, as you are now, are R-related to a large number of person-stages of a long-lasting simulation, how much prudential value does this provide?

This depends on three factors: (i) how much your relation to each of these person-stages matters, (ii) how well-off these person-stages are, and (iii) how the well-being of these person-stages should be aggregated.

First, regarding the aggregation of the well-being of the future person-stages, consider

The Intrapersonal Total View

In the absence of fission, the overall prudential value of the future is the sum total, for all future person-stages, of the well-being of that stage multiplied by the weight of the R-relation between that stage and you as you are now.

On this view, your future momentary well-being is added up, in proportion to the weights of the R-relations, to get the prudential value of your future.

Moreover, we can defend this view with a mere-addition argument:

Adding a long life that is at each point minimally positive in well-being to your lifespan seems to be at least as good for you as your life without that addition.

Then, making your life equal in quality throughout while increasing the average level of well-being a little bit seems to be good for you.

Then, we find that the end result—that is, a life that is at each time barely worth living—would be better for you than your current lifespan (no matter how good your current lifespan is).

Next, consider how much your relation to each of the future person-stages matters.

Relation C has a straightforward weighting: the proportion of how much of the earlier person-stage's psychological state the later person-stage shares or remembers.

The weighting of Relation R is less straightforward. Since Relation R holds in virtue of overlapping sequences of C-related person-stages, it seems natural to adopt following view:

> **The Multiplicative View of Continuity Strength** Let a *weight-product* of a sequence of C-related person-stages be equal to the product of the weights for each C-relation in the sequence. The weight of Relation R holding between person-stages x and y is equal to the maximum weight-product of any sequence xCy or yCx or a sequence via person-stages $z_1, z_2, \dots, z_n$ such that either
> (i) $xCz_1, z_1Cz_2, \dots, z_{n-1}Cz_n, z_nCy$ or
> (ii) $yCz_1, z_1Cz_2, \dots, z_{n-1}Cz_n, z_nCx$.

The idea is that you just multiply the weights of the C-relations to get the weight of the R-relation. For example, if B shares 99% of A's psychology, and C shares 99% of B's psychology, then C's well-being matters $0.99 \times 0.99 \approx 0.98$ to A. (E.g. if C's happiness is 100, then the value of A's future is 98.)

Does this view lead to Strong Prudential Longtermism given a successful upload with a long-lasting simulation?

Let every person-stage of a simulation be a year long. Suppose that the well-being of each person-stage is constant at u. Let the weight of each C-relation be w. Then, given the Multiplicative View of Continuity Strength, the prudential value of an x years long simulation is

$$\sum_{i=1}^{x} uw^{i} = \frac{uw\left(w^{x}-1\right)}{w-1}$$

As the simulation lasts longer, this converges to

$$\sum_{i=1}^{\omega} uw^{i} = -\frac{uw}{w-1}$$

To see that this does not favour Strong Prudential Longtermism, note that (given a positive well-being u and given that the weight w for the C-relations is positive and not greater than 100 %) the infinite number of years after the first 100 years do not contribute 100,000 times more to the prudential value of the future than the first 100 years of the simulation unless 99.99999 % of each person-stage's psychology is retained each year.

Would it be in each person-stage's interest that the next person-stage of the simulation remembers them to this extreme extent?

It may seem that it would, because the more the next person-stage remembers them the more the next stage (and the future) matters to them.

But, if each stage needs to remember the last one completely, it seems that the simulation would constantly need more memory in order to store new knowledge. (Computational resources could also be used to create more simulations.)

So it would make sense at some points to forget the last person-stage to some extent.

But, if so, a long-lasting simulation does not (by itself) lead to Strong Prudential Longtermism.

Recap

A single long-lasting simulation does not (by itself) lead to Strong Prudential Longtermism—unless each person-stage retains the psychology of the previous person-stages almost perfectly.

Psychological continuity

Psychological continuity (represented by Relation R) is the holding of overlapping sequences of psychological connectedness (represented by Relation C).

Psychological connectedness

Psychological connectedness is a direct psychological connection between a person at one time and a person at another time, such as the person at the latter time remembering the experiences of the person at the earlier time.

Recap

The Intrapersonal Total View

In the absence of fission, the overall prudential value of the future is the sum total, for all future person-stages, of the well-being of that stage multiplied by the weight of the R-relation between that stage and you as you are now.

The Multiplicative View of Continuity Strength

Multiply the weights of the C-relations to get the weight of the R-relation.

Branching simulations

So far we have only considered a single simulation. But, if we can make one simulation, we can make many. How should we aggregate the well-being of future person-stages in branching cases (that is, fission cases)?

Suppose that you will undergo uploading and that either (A) one simulation would be created and it would enjoy four years of high momentary well-being or (B) that simulation and a separate simulation would be created and each of these simulations would enjoy three years of high momentary well-being (at the same momentary well-being level as in A):

	S_1	S_2
A	4	/
B	3	3

One possibility is to stick to the total view even in fission cases:

The Prudential Total View

The prudential value of a prospect for you is equal to the sum total of the well-being of every person-stage that you, as you are now, are related to by the relation that matters, where the well-being of each stage is weighted by the strength of that relation.

On this view, you would be better off if two three-year simulations were created instead of one four-year simulation, that is, B is prudentially better than A.

	S_1	S_2
A	4	/
B	3	3

We can contrast this total view with an average view.

The Prudential Average View

Evaluate the prudential value of each simulation by the Intrapersonal Total View. Assume that fission stages are followed by a chance node with an equal probability of being followed by each of that stage's successors. Hence we transform prospects with fission into prospects of uncertainty. The prudential value of a prospect is equal to your expected well-being in the transformed prospect.

On this view, we treat the prospect of the two three-year simulations as if it were a fifty-fifty lottery between each of the two simulations being implemented on its own without the other. Hence, on the Prudential Average View, the prudential value of the two three-year simulations is the same as the prospect of a single three-year simulation, which is worse than the single four-year simulation.

But, there is a straightforward argument for the answer of the Prudential Total View.

Consider, in addition to A and B, a third prospect A^+ that is just like A except that a second simulation is also implemented and this additional simulation has the same momentary well-being level as the first simulation but is only run for one year:

	S_1	S_2
A	4	$\diagup$
A^+	4	1
B	3	3

It seems that, if simulation S_1 in A provides what matters in survival, then the same simulation in A^+ should also provide what matters in survival. The only difference in A^+ is that, in addition to S_1, there is another simulation to which you also stand in the relation that matters. So, from the perspective of what matters in survival, A^+ should be at least as great a success as A. Consequently, A^+ must be at least as good as A for you.

Next, compare A^+ and B. Prospect B differs from A^+ in that S_1 lives for one year less but S_2 lives for two more years.

Given that you stand in the relation that matters to *both* simulations, in terms of prudential value the two extra years for S_2 in B should outweigh the single extra year for S_1 in A^+. So B is better than A^+ for you.

Then, by the transitivity, we have that B is better than A for you.

	S_1	S_2
A	4	/
A^+	4	1
B	3	3

If we adopt the Prudential Total View, rather than the Prudential Average View, we seem to have a route to Strong Prudential Longtermism.

If we create not just one simulation of you but a large number of simulations, your prudential value from these simulations increases in proportion to the number of simulations. Moreover, each one of these simulations is in much the same situation, as they also increase their prudential value from the future the more simulations there will be of them.

And, in turn, these simulations of simulations are in much the same situation, as they can increase their prudential value by creating even more simulations of themselves. Hence it seems that we would get an explosion of more and more simulations that all have what matters in relation to you.

Since this increase in the number of simulations will outweigh the diminishing weight of the *R*-relation between you, as you are now, and the simulations as they get more distant from you, you will (at least in expectation) get most of your prudential value from this enormous amount of simulations in the far future.

Hence, if we have a sufficiently high credence that uploading works, we get Strong Prudential Longtermism.

So far in our discussion of uploading, we have assumed that psychology is what matters in survival.

If (i) psychological continuity is what matters, (ii) uploading technology is feasible, and (iii) simulations of brains would be conscious, it follows that you would have what matters in survival to a simulation from your uploaded brain. Or, at least, it follows if this continuity need not have its normal cause: being caused by the continued existence of your brain.

But there is at least one rival to the psychological view that may also allow that uploading provides what matters.

The phenomenal view

You have what matters in relation to a future person-stage if and only if you are phenomenally continuous with that future person-stage.

Phenomenal continuity is, basically, the relation of partaking of the same stream of consciousness.

In the same way as psychological continuity is the holding of overlapping sequences of psychological connectedness, phenomenal continuity is the holding of overlapping sequences of phenomenal connectedness. Phenomenal connectedness, in turn, is the relation of experienced togetherness—that is, the relation of one experience being experienced together with another experience in the same conscious state.

This relation can hold between experiences at a time, such as your current visual experience and your current auditory experience. But this relation of experienced togetherness can also hold over time. For example, when you are listening to music and you hear one note transition into the next, or more generally, each one of your experiences flows into the next.

On this phenomenal view, a simulation provides what matters in survival if and only if a stream of consciousness you currently partake in will include the experiences of the simulation.

If simulations can be conscious and have experiences at all, could we get your stream of consciousness to transfer to a simulation?

Barry Dainton suggests that we can. By gradually replacing more and more of your brain with functionally equivalent digital silicon-based parts, he suggests that your stream of consciousness would continue intact.

Then, if done gradually, uploading may still provide what matters in survival even on the phenomenal view.

Recap

If we adopt the Prudential Total View, rather than the Prudential Average View, we seem to have a route to Strong Prudential Longtermism via a large number of simulations.

Since this increase in the number of simulations will outweigh the diminishing weight of the *R*-relation between you, as you are now, and the simulations as they get more distant from you, you will (at least in expectation) get most of your prudential value from this enormous amount of simulations in the far future.

The Prudential Total View

The prudential value of a prospect for you is equal to the sum total of the well-being of every person-stage that you, as you are now, are related to by the relation that matters, where the well-being of each stage is weighted by the strength of that relation.

The phenomenal view

You have what matters in relation to a future person-stage if and only if you are phenomenally continuous with that future person-stage.

If Prudential Longtermism is true, then a large number of theories that otherwise would not lead to (impersonal) Longtermism may turn out to do so.

- Person-affecting views on which we should minimize the strongest complaint would lead to Longtermism. This is so, since the strongest complaints will come from people for whom Prudential Longtermism is true.
- Common-sense morality, on which you should prioritize your family and friends, would lead to Longtermism if Prudential Longtermism is true for a sufficient number of your family and friends (and a sufficient number of strangers).
- Self-interest theories would lead to Longtermism if Prudential Longtermism is true for the agent.
- Person-affecting utilitarianism would lead to Longtermism if Prudential Longtermism is true for a sufficient number of current people.

The practical implications of Longtermism based on Prudential Longtermism would differ in some respects from those of Longtermism based on total utilitarianism.

In addition to prioritizing the reduction of existential risk to safeguard humanity as a whole, Longtermism based on Prudential Longtermism would also prioritize speeding up technological progress in the areas that may help life extension.

Given the common view that the badness of death consists largely in how much better you life would have been in expectation if you had lived on, Prudential Longtermism makes avoiding an early death all the more pressing.

A side-effect of Prudential Longtermism and a general increase in people's life expectancy is likely to be that people become more invested in the long-term, since they now have a personal stake in it.

Conclusion

Anti-ageing leads to Weak Prudential Longtermism with reasonable probabilities of anti-ageing working (at least 8%).

But it does not lead to Strong Prudential Longtermism.

Cryonics alone (or in combination with anti-ageing) at most gives us Weak Prudential Longtermism.

A single long-lasting simulation does not (by itself) lead to Strong Prudential Longtermism—unless each person-stage retains the psychology of the previous person-stages almost perfectly.

Conclusion

If we adopt the Prudential Total View, rather than the Prudential Average View, we seem to have a route to Strong Prudential Longtermism via a large number of simulations.

Since this increase in the number of simulations will outweigh the diminishing weight of the *R*-relation between you, as you are now, and the simulations as they get more distant from you, you will (at least in expectation) get most of your prudential value from this enormous amount of simulations in the far future.

The practical implications of Longtermism based on Prudential Longtermism would differ in some respects from those of Longtermism based on total utilitarianism.

In addition to prioritizing the reduction of existential risk to safeguard humanity as a whole, Longtermism based on Prudential Longtermism would also prioritize speeding up technological progress in the areas that may help life extension.