Prudential Longtermism

Petra Kosonen

Joint work with Johan 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.

- Longtermists usually focus on preventing human extinction due to AI or a pandemic.

- Person-affecting views seem to provide an argument against Longtermism because, on these views, an outcome cannot be better than some other outcome unless it is better for someone.
- Each act we perform results in different people existing in the future due to the ripple effects of these acts.[1]
- 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 that no one alive today will be alive in the far future, then person-affecting views lead to the rejection of Longtermism.

[1] Parfit (1984).

Prudential Longtermism

- There is a different path to Longtermism that is perfectly compatible with those views:

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.

- 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.

- If Weak Prudential Longtermism is true for you, then the far future matters more than the near future for your prudential value.
- 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.
- Let the next 100 years constitute the short-term, and let the long-term start after that.

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 Strong Prudential Longtermism is true for you, then the far future matters overwhelmingly more than the near future for your prudential value.
- Let us 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.

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: 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 30-year-olds who died in the U.S. in 2018), your life expectancy is $1/0.0013 \approx 770$ years.
- Is 770 years a sufficiently long life expectancy to lead to Prudential Longtermism?

- The probability of anti-ageing working must be at least 8% in order for it to lead to Weak Prudential Longtermism.
- Then, your expected number of life years in the long-term is greater than your expected number of life years in the short-term.

SKIP

The expected years of life in the short-term (that is, the next 100 years) if anti-ageing works is

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

Here, p=the constant probability of death each year if anti-ageing works (which I have assumed to be 0.13 %.)
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%.

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

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.

Psychological continuity

- Many philosophers believe that psychological continuity (represented by Relation R) 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 connectedness (represented by Relation C).
- 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.

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 looking, the simulations need not have any special interest in continuing to be directly psychologically connected to you.
- So they might gradually let go of their memories of you in order to make room (in computer memory) for more useful knowledge.
- So there would be no connectedness between distant stages of the simulation.

- So, if Relation C is what matters, it seems that uploading would not lead to Strong 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.
- 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.

- 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 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.

- Does this view lead to Strong Prudential Longtermism given a successful upload with a long-lasting simulation?
- No, unless 99.99999 % of each person-stage's psychology is retained each year.

SKIP

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.

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

Branching simulations

- 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	-
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.

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

- 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 a large number of simulations of you, 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—leading to an explosion of more and more simulations.
- This increase in the number of simulations can outweigh the diminishing weight of the R-relation between you, as you are now, and the simulations as they get more distant from you.
- So, you will (at least in expectation) get most of your prudential value from this enormous amount of simulations in the far future.

- So, if we have a sufficiently high credence in uploading working, we get Strong Prudential Longtermism.
- 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:
    1. Person-affecting views on which we should minimize the strongest complaint would lead to Longtermism. The strongest complaints will come from people for whom Prudential Longtermism is true.
    2. Common-sense morality, on which you should prioritize your family and friends, might lead to Longtermism.
    3. Self-interest theories would lead to Longtermism if Prudential Longtermism is true for the agent.
    4. Person-affecting utilitarianism would lead to Longtermism if Prudential Longtermism is true for a sufficient number of current people.

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.
- 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.
- If we adopt the Prudential Total View, we seem to have a route to Strong Prudential Longtermism via a large number of simulations.
- Person-affecting views (and others) would then imply Longtermism.

References I

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