The Moral Case for Long-Term Thinking

Hilary Greaves, William MacAskill, and Elliott Thornley

In *The Long View*[*]

Summary: This chapter makes the case for *strong longtermism*: the claim that, in many situations, impact on the long-run future is the most important feature of our actions. Our case begins with the observation that an astronomical number of people could exist in the aeons to come. Even on conservative estimates, the expected future population is enormous. We then add a moral claim: all the consequences of our actions matter. In particular, the moral importance of *what* happens does not depend on *when* it happens. That pushes us toward strong longtermism.

We then address a few potential concerns, the first of which is that it is impossible to have any sufficiently predictable influence on the course of the long-run future. We argue that this is not true. Some actions can reasonably be expected to improve humanity's long-term prospects. These include reducing the risk of human extinction, preventing climate change, guiding the development of artificial intelligence, and investing funds for later use. We end by arguing that these actions are more than just extremely effective ways to do good. Since the benefits of longtermist efforts are large and the personal costs are comparatively small, we are morally required to take up these efforts.

Introduction

The future is big. Our planet currently hosts around eight billion people. This century will see the birth of more than ten billion. If that number holds steady for just ten more centuries, we have a hundred billion people ahead of us. But humanity could last much longer than that. If all goes well, we can expect our descendants to outnumber us by an even greater margin. We residents of the twenty-first century could turn out to be a drop in the ocean.

It is hard to grasp the size of humanity's potential. We are not used to thinking on the necessary timescales. Harold Wilson said that a week was a long time in politics, and the remark seems true of many other domains too.[1] Our world

[*] Published by FIRST, 2021. Open-access version: tinyurl.com/thelongview2021

[1] (Rees 1994)

changes so quickly that the consequences of our actions even a few years from now are tough to predict,[2] so it is no surprise that we rarely consider how our decisions might affect people living hundreds, thousands, or even millions of years in the future.

Nonetheless, we — the authors — believe that this neglect of the long-term future is a grave moral error. The view recently dubbed *longtermism* serves as a corrective.[3] According to this view, we should be particularly concerned with ensuring that the long-term future goes well. In this contribution, we argue for both longtermism and a further claim that we call *strong longtermism* which states that, in many situations, impact on the long-term future is the most important feature of our actions today.

Where exactly the short-term future ends and the long-term future begins is not important. We claim that the view is true even when we draw the line a surprisingly long time from now — say, a hundred years. The claim, then, is that the moral value of our actions depends primarily on their consequences arising more than a century in the future. That means that the predicted short-run value of our actions should not weigh heavily in our decision-making. Instead, our choices should be driven mainly by long-run considerations. Short-term effects matter, but they matter primarily as mediators of long-term effects.

We believe that strong longtermism has practical implications for individuals, charities, and governments. Humanity's future could be extraordinarily valuable, and it currently hangs in the balance. If these facts were widely recognised, many of our priorities would change.

The case for strong longtermism

The case for strong longtermism begins with the observation that our future could be vast. Astronomical numbers of people could exist in the aeons to come. Of course, the exact number is uncertain. The range of possibilities is wide. But, for our purposes, we can work with the *expected* number of future people. We calculate this figure in the same way that we calculate the expected value of a lottery ticket. Suppose that a ticket offers a 1% chance of winning £300. Then its expected value is $0.01 \times £300 = £3$.

Whether a lottery ticket is worth buying depends on the numbers, and the same is true of our argument for longtermism. Here, as below, we will endeavour to be conservative in our estimates, erring on the side of underestimating the expected number of future people. If the case for longtermism is strong on these numbers, it will be even stronger on less cautious estimates. In that spirit, suppose that the chance that humanity survives until the Earth becomes uninhabitable —

[2] (Tetlock 2005)

[3] (MacAskill 2019a; Greaves and MacAskill 2019)

one billion years from now[4] — is just 0.1%. The future is hard to predict, so being more than 99.9% confident that we will not make it that far seems hubristic. Suppose also that ten billion people live in each century. In that case, the expected number of future people is at least 100 trillion (10^{14}) — over 10,000 times the number of people alive today.

The size of that number might lead you to think that our estimates were not conservative after all. But note that the above calculation leaves out many opportunities for further inflation. Perhaps the most significant is the chance of space settlement. There are around 250 billion stars in the Milky Way, some of which will last for trillions of years.[5] If we judge that there is even a tiny chance that our descendants settle just a small fraction of these solar systems, the expected number of future people balloons upward.

Suffice it to say, the expected future population is large indeed. That is the first component of our argument for strong longtermism. The second component is a moral claim: all the consequences of our actions matter. More specifically, the moral importance of *what* happens does not depend on *when* it happens. Agony and ecstasy occurring a hundred years from now matter just as much as agony and ecstasy occurring ten years from now.

This claim rules out what economists call a 'positive rate of pure time preference': preferring that good things occur at earlier rather than later times purely because they are earlier. To be sure, time preferences are appropriate in some domains. A pound now is preferable to a pound in ten years' time. But that is because we expect to be richer in the future, and pounds have diminishing marginal utility. Features of our lives that are intrinsically good or bad — things like joy and sadness — do not have diminishing marginal utility, so time preferences concerning these things are out of place. Consider an example. Suppose that you can save one person from torture ten years from now or two people from torture a hundred years from now, and that your decision will have no other consequences. It seems clear that, in this case, you should save the two people. Their pain should not be discounted simply because it occurs further in the future.[6]

Together, our argument's two components — the future is vast and all consequences matter — push us toward our conclusion: we can have a much bigger effect on the value of the future by trying to change its long-term rather than its short-term value. That in turn suggests that we should devote much more of our focus to considering the long-run effects of our decisions, and makes plausible the

[4] (Adams 2008)

[5] Ibid.

[6] Greaves (2017) surveys further arguments for and against a positive rate of pure time preference.

strong longtermist claim that, in many situations, we ought to perform the action that we expect will have the best effects on the long-term future.

This claim is only strengthened by the observation that, so far, few people have recognised the importance of this longtermist insight. Most people and institutions are biased towards the short term.[7] If we direct our focus on the next few years, we enter a crowded field in which many of the best opportunities have already been taken and further progress is difficult. But if we instead cast our sights further, we find fresh ground. Any opportunities here are less likely to have been taken, so we can expect to have an outsized impact.

Longtermist initiatives

But can we predictably improve the long-run future? One might think not, reasoning along the following lines:

> Our world is so complex that it is impossible to foresee what effects our actions will have decades from now, let alone centuries. Since the long-run consequences of our actions are so uncertain, we cannot reasonably expect that any of our actions will make the long-term future better rather than worse. In light of this uncertainty, we should focus on the near future where effects are easier to predict.

We agree that the long-run value of many actions is hard to predict. But, importantly, this is not true of *all* actions. Some actions *can* be reasonably expected to improve humanity's long-term prospects, and this is enough to make strong longtermism true.

To explain one set of such actions, we first need to introduce an idea. Imagine a golf ball blown around a putting green by blustering winds. While the ball is on the turf, it will roll back and forth. The state of the scene will be constantly changing. But if the ball falls into a hole, it will remain there. The ball's being in the hole is what we call a *persistent state*. It is a state which, upon coming about, tends to persist for a long time.

Our world is like this windy putting green. It too has persistent states. Human extinction is one of them. The chances of humanity evolving all over again, post-extinction, are tiny. Human survival is another persistent state, albeit to a lesser extent. While the risks of extinction are real, there is at least a strong tendency for humanity to endure. These two persistent states differ in their long-run value. Our survival through the next thousand years and beyond is, plausibly, better than our

[7] (Frederick, Loewenstein, and O'Donoghue 2002)

extinction in the near future. So, if we can reduce the chance of human extinction, we can predictably improve the long-term future.[8]

And it is increasingly recognised that we *can* reduce the chance of extinction. Matheny (2007), for instance, estimates that a $20 billion asteroid deflection system could halve the probability of an extinction-level asteroid hitting the Earth this century, reducing the risk from one-in-one-million to one-in-two-million. That decrease may seem small in absolute terms, but it makes an enormous difference to the expected number of future people. Recall that our calculation from the previous section gave us an expected future population of 100 trillion. Increasing the chance that this population gets to exist by just one-in-two-million is equivalent to saving 50 million lives in expectation. That comes out at $400 per life saved. And this is just one example. Combatting other extinction threats — such as those arising from new or engineered pandemics — might be even more cost-effective.[9] This we could achieve by funding biosecurity work at the Johns Hopkins Center for Health Security,[10] for instance, or the Future of Humanity Institute.[11]

That said, the case for reducing extinction risk hangs on our moral view. If we embrace a person-affecting approach to future generations — on which we care about *making lives good* but not about *making good lives*[12] — then extinction would not be so bad. It might even be judged good.[13] An asymmetric moral view — according to which bad lives get more weight than good lives — might lead us to a similar verdict.[14] And even on more standard views about the value of bringing new generations into existence, we might worry that future lives will be bad overall, so that extinction would be the lesser evil.[15]

Nevertheless, we argue, those drawn to person-affecting, asymmetric, and pessimistic views should still be strong longtermists. That is because extinction is not the only persistent state whose likelihood we can affect. Artificial intelligence presents another opportunity. Experts judge that there is a real chance that we develop advanced AI this century, with capabilities exceeding our own across a wide range of domains.[16] As a result of their superior intelligence, these artificial

---

[8] Bostrom (2013) presents this argument in more detail.

[9] (Millett and Snyder-Beattie 2017)

[10] www.centerforhealthsecurity.org/

[11] www.fhi.ox.ac.uk

[12] This slogan is a rephrasing of Narveson (1973, 80). See Roberts (2011) for a more recent discussion of person-affecting approaches.

[13] Thomas (2019) canvasses a range of possibilities.

[14] (Hurka 2010)

[15] (Althaus and Gloor 2019)

[16] (Bostrom 2014, chaps 1–2; Müller and Bostrom 2016; Grace et al. 2018)

agents may come to exert significant control over human affairs: making important decisions on behalf of individuals, governments, and other institutions. These agents might also endure indefinitely. Since their underlying code could be copied, they could outlast any given piece of hardware.[17] These two features of AI systems — their influence and their staying power — mean that they are likely to have substantial and lasting effects on the future.[18] That in turn suggests that we can have a beneficial influence on the long-term by increasing the chances that these systems are aligned with the right values. Work underway at OpenAI[19] and the Center for Security and Emerging Technology[20] — to take just two examples — aims to achieve exactly that.

Another set of persistent states relates to climate change. A warmer climate could slow long-run economic growth, leaving future civilisation worse-off indefinitely.[21] It could also lead to the extinction of species, the destruction of coral reefs, and other forms of irreversible damage to our ecosystem.[22] Because these potential harms are near-permanent, we can expect that fighting climate change will have enduring effects on the future.

In sum, humanity finds itself in a delicate position. Our civilisation is currently poised between a range of persistent states. Falling into one of these states would likely have immense effects on the long-term future. Through the judicious use of time and resources, we can alter the chances that these states come about. As a consequence, we have the power to make our world better for generations to come.

But, as noted above, the case for strong longtermism hinges on the numbers. Not every lottery ticket is worth buying, and the same could be true of our proposed longtermist interventions. However, we argue that — even on conservative figures — the opportunities we list above are well worth the expense. Matheny's proposed asteroid deflection system is one example. Another concerns artificial intelligence. If £1 billion of grants could reduce the chance of a catastrophic AI outcome — in which humanity's future is rendered near-worthless — by just 0.001%, then a £10,000 donation can do as much good as saving 10,000

[17] (Hanson 2016, 57–58)

[18] (Chalmers 2010; Bostrom 2014; Ord 2020)

[19] www.openAI.com

[20] www.cset.georgetown.edu

[21] (Stern 2007; Pindyck 2013)

[22] (IPCC 2014, 1052–54)

lives.[23] By contrast, it is widely agreed that the best available human-centric short-term interventions save roughly 4 lives per £10,000, at least in the short term.[24]

And there is still much we do not know. Even if, at present, we cannot reasonably expect any of these longtermist initiatives to have better consequences than the most effective short-term actions, it remains possible that extra information would tip the scales in favour of a longtermist option. In that case, funding research into the long-run effects of various initiatives may be our best move. Since future people would likely take note of this research, we can expect our donations to increase the effectiveness of humanitarian efforts for many years to come.

Another option is to save our money.[25] We could set up a foundation or donor-advised fund with an explicitly longtermist mission. This fund would pay out if and when a good opportunity to shape the long-term future arises. The phase before the widespread deployment of advanced artificial intelligence would be one such opportunity. Since both the value of this longtermist fund and our knowledge about the efficacy of various actions is likely to grow over time, we can expect its impact to be especially substantial.

The upshot is that we have a whole array of opportunities to benefit the generations who could exist in centuries to come. Even on the most cautious estimates, we should expect longtermist initiatives to do many times as much good as it is possible to do in the short term. So, if our aim is to do good, we should focus on the long term.

Are we morally required to be longtermists?

The argument above will motivate many people to set their sights on the long term. But others might want to hear more about the precise moral status of longtermist initiatives. For even if longtermist actions like reducing extinction risk have the *best effects* on the future, that does not immediately imply that we are *morally required* to reduce extinction risk. For those people motivated mainly by a desire to avoid acting immorally, and not by a more wide-ranging desire to do good, this last step is important.

Let us expand on this point. On some moral views, doing what has the best effects is not always morally required.[26] Consider an example. Suppose that you are walking by a shallow pond, and a child asks you to wade in and get her football.

---

[23] (Greaves and MacAskill 2019)

[24] This figure comes from GiveWell's 2020 charity assessments. Their model can be found here: www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models. Note that GiveWell's focus is on health in the developing world. Short-term initiatives aimed at helping animals are plausibly even more cost-effective. See www.animalcharityevaluators.org for more.

[25] (Christiano 2014; MacAskill 2019b; Trammell 2020)

[26] See, e.g., Scheffler (1982)

You judge that the joy the child would feel in getting her ball back outweighs the frustration you would feel in ruining your clothes, so wading in would have the best consequences. Nevertheless, one might well claim, you are not morally required to wade in. You need not feel bad about staying dry. Similarly, one might argue, we are not morally required to do what has the best effects on the long-run future. We can instead devote our time and resources to other things.

Perhaps there are cases where we need not do what is impartially best.[27] We can allow that. But even so, we maintain that longtermist actions are morally required. This conclusion is implied by the following plausible claim:

> When the action with the best effects has effects *much* better than other available actions, and any difference in personal costs is comparatively small, we are morally required to perform the action with the best effects.[28]

This claim is compatible with the judgement that you need not wade into the pond. Although wading in would have better effects than staying dry, the effects are presumably not *much* better. The claim also makes sense of our judgements in an amended version of the case. Suppose instead that the child is drowning in the shallow pond. Then it seems undeniable that you are morally required to wade in and save her. Precisely because the stakes are high and the cost to you is small, you must do what is best.[29]

Our situation is closer to the drowning case than it is to the football case. Longtermist initiatives like preventing future pandemics are not merely slightly more effective than the best short-term initiatives. They are many times more effective.[30] Since the consequences of longtermist efforts are so much better in expectation, and the personal costs of a long-run focus are small, we are morally required to take up longtermist efforts.

Conclusion

Humanity's potential is vast and yet fragile. We could be on the verge of a long and magnificent future in which our descendants flourish for aeons to come. We could also be headed for an untimely end, or a drop into a permanent rut. Our fate is as yet undetermined. Influencing the chances that these futures come to pass is within our power.

These facts, in combination with a couple of plausible moral claims, have led us to a surprising conclusion: in many situations, effects on the long-run future are

---

[27] That is to say, perhaps maximising consequentialism is false. See Sinnott-Armstrong (2019, sec. 6)

[28] Greaves and MacAskill (2019) discuss this claim in more detail.

[29] This case was first presented by Singer (1972).

[30] That is unless the best short-term initiatives happen to have long-run benefits comparable to those of the longtermist interventions discussed above. This condition seems to us unlikely.

the most important feature of our actions today. This shift to a strong longtermist perspective is of no small importance. In fact, it has many practical implications. It directs us to spend significantly more of our time and resources on reducing extinction risk, preventing climate change, guiding AI development, improving institutional decision-making, fostering international cooperation, researching the long-run efficacy of various initiatives, investing funds for later use, and — almost certainly — many other things besides.

References

Adams, Fred C. 2008. 'Long-Term Astrophysical Processes'. In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković. Oxford: Oxford University Press.

Althaus, David, and Lukas Gloor. 2019. 'Reducing Risks of Astronomical Suffering: A Neglected Priority'. *Center on Long-Term Risk* (blog). https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/.

Bostrom, Nick. 2013. 'Existential Risk Prevention as Global Priority'. *Global Policy* 4 (1): 15–31.

———. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Chalmers, David J. 2010. 'The Singularity: A Philosophical Analysis'. *Journal of Consciousness Studies* 17 (9–10): 7–65.

Christiano, Paul. 2014. 'We Can Probably Influence the Far Future'. *Rational Altruist* (blog). https://rationalaltruist.com/2014/05/04/we-can-probably-influence-the-far-future/.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. 'Time Discounting and Time Preference: A Critical Review'. *Journal of Economic Literature* 40 (2): 351–401.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. 'When Will AI Exceed Human Performance? Evidence from AI Experts'. *Journal of Artificial Intelligence Research* 62: 729–54.

Greaves, Hilary. 2017. 'Discounting for Public Policy: A Survey'. *Economics & Philosophy* 33 (3): 391–439.

Greaves, Hilary, and William MacAskill. 2019. 'The Case for Strong Longtermism'. *GPI Working Paper No.7-2019*. https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism/.

Hanson, Robin. 2016. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford: Oxford University Press.

Hurka, Thomas. 2010. 'Asymmetries in Value'. *Noûs* 44 (2): 199–223.

IPCC. 2014. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment*

Report of the Intergovernmental Panel on Climate Change. Cambridge, UK and New York, NY: Cambridge University Press.

MacAskill, William. 2019a. 'Longtermism'. *The Effective Altruism Forum* (blog). 2019. https://forum.effectivealtruism.org/posts/qZyshHCNkjs3TvSem/longtermism.

MacAskill, William. 2019b. 'When Should an Effective Altruist Donate?'. *GPI Working Paper* 8-2019.

Matheny, Jason G. 2007. 'Reducing the Risk of Human Extinction'. *Risk Analysis* 27 (5): 1335–44.

Millett, Piers, and Andrew Snyder-Beattie. 2017. 'Existential Risk and Cost-Effective Biosecurity'. *Health Security* 15 (4): 373–83.

Müller, Vincent C., and Nick Bostrom. 2016. 'Future Progress in Artificial Intelligence: A Survey of Expert Opinion'. In *Fundamental Issues of Artificial Intelligence*, edited by Vincent Müller, 553–571. Springer.

Narveson, Jan. 1973. 'Moral Problems of Population'. *The Monist* 57 (1): 62–86.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.

Pindyck, Robert S. 2013. 'Climate Change Policy: What Do the Models Tell Us?' *Journal of Economic Literature* 51 (3): 860–72.

Rees, Nigel. 1994. *Brewer's Quotations: A Phrase and Fable Dictionary*. London: Weidenfeld & Nicolson.

Roberts, Melinda A. 2011. 'An Asymmetry in the Ethics of Procreation'. *Philosophy Compass* 6 (11): 765–76.

Scheffler, Samuel. 1982. *The Rejection of Consequentialism*. Oxford: Clarendon Press.

Singer, Peter. 1972. 'Famine, Affluence, and Morality'. *Philosophy and Public Affairs* 1 (3): 229–243.

Sinnott-Armstrong, Walter. 2019. 'Consequentialism'. In *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/sum2019/entries/consequentialism/.

Stern, Nicholas. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge: Cambridge University Press.

Tetlock, Philip E. 2005. *Expert Political Judgment*. Princeton: Princeton University Press.

Thomas, Teruji. 2019. 'The Asymmetry, Uncertainty, and the Long Term'. *GPI Working Paper No.11-2019*. https://globalprioritiesinstitute.org/teruji-thomas-the-asymmetry-uncertainty-and-the-long-term/.

Trammell, Philip. 2020. 'Discounting for Patient Philanthropists'. https://philiptrammell.com/static/discounting_for_patient_philanthropists.pdf.