William MacAskill, *What We Owe The Future: A Million-Year View*

(One World Publications, London, 2022), pp. 246.

Michael Plant

Wellbeing Research Centre, University of Oxford, Oxford, UK and The Happier Lives Institute, London, UK

In *What We Owe The Future (WWOTF)*, William MacAskill makes the case for *longtermism*, the idea that positively influencing the longterm future is a key moral priority of our time. By 'longterm', MacAskill means the really longterm: the book is subtitled 'A million-year view'. MacAskill says his case is based on three premises:

(1) Future people count.
(2) There could be lots of them.
(3) We can make their lives go better.

He remarks 'these premises are simple, and I don't think they are particularly controversial. Yet taking them seriously amounts to a moral revolution' (p. 9). His main proposals are to focus on reducing the chance of premature extinction, allowing continued moral development by strengthening liberal institutions, and growing longtermism as a research field.

The book certainly marks an evolution in MacAskill's own thinking: he is a leading light in effective altruism, the research field and social movement that aims to find the best ways to help others. MacAskill recounts that he used to believe that this meant focusing on the global poor, but others eventually persuaded him of longtermism. Although MacAskill states his aim was to 'write the case for longtermism that would have convinced me a decade ago' (p. 6), the book is clearly aimed at the general public, not academic philosophers. Instead of dense, technical text and a creeping barrage of thought experiments, we are treated to flowing prose and a whistlestop tour of history; it was joyful, even moving, to read.

Given the objective of persuading others, the book must count as a runaway success. For its launch, MacAskill pulled off a media blitzkrieg, with either a profile of himself, or a review of the book – in either case usually glowing – seeming to materialise in every outlet this author had ever heard of. He even featured on a US late-night talk show, not the normal domain of philosophers.

However – and although I wanted to share MacAskill's enthusiasm for longtermism – I found the case unpersuasive. Further, it seems too bold to claim that the premises are simple or uncontroversial or, if taken seriously, would amount to a moral revolution.

To be clear, my concern is not that MacAskill does not treat his topic with the painstaking rigour he is clearly capable of – that would be unreasonable, given he is writing for a general audience. Rather, it is simply that MacAskill does not do enough to identify or anticipate, then address, the weaknesses in his argument. At times, I found the book uncomfortably polemical, as if MacAskill had set out to convince the reader, as effectively as possible, to share his conclusion, even if they would not fully understand the reasons for it and the challenges to them. Before I elaborate on my concerns, I will summarise the book.

https://doi.org/10.1017/S0953820823000109 Published online by Cambridge University Press

Summary

Chapter 1 introduces the three-premise case for longtermism given above and motivates the first two premises. 'Future people count' is animated with the *Hiker* thought experiment borrowed from Derek Parfit: suppose I am hiking and consider dropping a shard of glass, knowing that a child may cut herself on it. MacAskill says 'Should I care whether it's a week, a decade, or a century from now? No. Harm is harm. Whenever it occurs' (p. 9). He goes on, 'Distance in time is like distance in space. People matter even if they live thousands of miles away. Likewise, they matter even if they live thousands of years hence.' The second premise requires no elaboration: if humanity continues, many could live.

Chapter 2 points out that we clearly can shape history: one example given is that our ancestors did it by spreading across the globe and hunting large animals ('megafauna') to extinction. MacAskill provides a three-factor formula for assessing the longterm value of outcomes: we multiply their *significance* (the average value of an outcome), *persistence* (how long it lasts) and *contingency* (whether, and for how long, the outcome would have occurred anyway). MacAskill hypothesises that civilisation is at a moment of 'plasticity': it is currently malleable but will soon 'set' in the form we leave it in.

Chapters 3–7 burnish MacAskill's practical thesis: that we *can* change the future. Chapters 3 and 4 present some examples from history including a long discussion about how slavery may never have ended were it not for the work of the early abolitionists; chapter 4 suggests that the rise of artificial general intelligence (AGI) may make now a particularly 'plastic' time.

Chapters 5–7 look forward to potential risks to humanity's survival and flourishing such as engineered pathogens, great-power wars, civilisational collapse, technological stagnation; the aim is more to identify these than to provide novel solutions.

Chapters 8 and 9 develop the moral thesis: that we *should* change the future. These are only the places ethics are discussed in *WWOTF* outside chapter 1 (where they are not considered in any depth). I recognise MacAskill has much ground to cover, but this nevertheless felt too brief. Chapter 8 focuses on population ethics, which he notes is a notoriously confounding area of philosophy. MacAskill introduces the *intuition of neutrality*: the view that, in slogan form, says 'morality is in favour of making people happy but indifferent about making happy people' (this famous quote, sometimes called 'Narveson's Dictum', is from Jan Narveson, 'Moral Problems of Population', *The Monist*, 57(1) (1973) 80). If correct, this would present a severe challenge to longtermism: we would be indifferent about bringing about those future (happy) generations. MacAskill does not motivate or defend this view and quickly dismisses it as untenable. Ultimately, he argues for the *Total View* in population ethics, on which the value of an outcome is the sum total of wellbeing in it, so adding happy lives is good and future people matter just as much as those alive today. (Technically, given MacAskill's views on moral uncertainty, he adopts a *Critical Level View* with a 'low but positive critical level' (p. 187). The details aren't important: at a low level, the view is practically identical to the Total View; higher critical levels progressively weaken the case for longtermism if one also assumes, which MacAskill does, that future people would generally have good lives. Hence, for simplicity, I take MacAskill as adopting the Total View.)

Chapter 9 considers the world's wellbeing trajectory – if it would be negative, should we worry about going extinct? – and concludes, speculatively, that that future is bright.

Finally, chapter 10 turns to what we should do. Many of MacAskill's comments here, such as about choosing charities and careers, are not specific to longtermism and I will

https://doi.org/10.1017/S0953820823000109 Published online by Cambridge University Press

not mention. Regarding longtermism, MacAskill recognises we are 'often in a position of deep uncertainty with respect to the future', observing that, if someone born in 1500 had sought to improve today, they would have had little idea of what to do (p. 225). MacAskill's move here is to compare longtermism to an expedition into the unknown: we do not know exactly what we will face, but we can still prepare by doing things that seem broadly useful. In particular, he recommends more work on various priorities where we are confident what the good outcomes are, such as reducing climate change and wars and increasing pandemic preparedness. For areas where the priorities are unclear – he cites AGI safety and longtermism itself – he suggests developing those as research fields so we are in a better place to know what to do and to take action later.

Concerns

I have four challenges to the nature or presentation of the case.

First, MacAskill's three-premise 'case' is not a valid argument for longtermism. It does not follow, merely from recognising that we *can* help one large group of beings, that they are a *key priority*. Consider this alternative:

(1a) Philosophers matter.
(2a) There could be many of them.
(3a) We can make their lives go better.

Does this form a valid argument for 'philosopherism', the idea that positively influencing the lives of philosophers is a key moral priority? Presumably not. Note that we can replace 'philosophers' with many other groups and so, by the same logic, seemingly conclude that nearly everything is a 'key priority', rendering the term meaningless. This would be less of a problem if MacAskill defined 'key priority' or went on to compare longtermism to some other putative key priorities – such as global poverty – but neither happens in *WWOTF*. Although MacAskill says that taking longtermism seriously *would* amount to a moral revolution, he does not tell us how seriously we *should* take longtermism.

Second, when it comes to the thesis we should influence the longterm, MacAskill presents his premises as simple and uncontroversial when they are not; I have in mind here premises (1) and (3).

In chapter 1, as noted, MacAskill's claim is that people distant in time matter just as much as people distant in space. This implies there is no possibly relevant distinction that might ground a lesser concern for future people than for distant people.

Yet, there is one. The people of the far future are hypothetical. They might exist, they might not. If we go extinct, none will exist. In contrast, people on the other side of the world actually exist and necessarily exist (they exist whatever we do). They have physical bodies. They can be benefitted or harmed. They count. Intuitively, if a particular person will exist in the future, we should count them just like a present person. But almost all far future people are not like this: they are hypothetical, because their existence is contingent on our actions. It is far less clear that hypothetical people count. How can we benefit or harm someone unless they exist? Can we harm James Bond? (Bond is admittedly fictional, not hypothetical, but this may illustrate the point nevertheless.) Hence, we could say future people count *in theory*, in the sense that time *per se* is not relevant. But far future people might little count *in practice*, because

https://doi.org/10.1017/S0953820823000109 Published online by Cambridge University Press

they are largely hypothetical, and hypothetical people may not count. The premise 'future people count' is not straightforward.

MacAskill is aware of all these subtleties, as his discussion much later, in chapter 8, on population ethics, shows. The intuition of neutrality is based on the notion there is a morally relevant distinction between people who do exist and those who *could* exist. MacAskill may not be sympathetic to the intuition of neutrality, but many philosophers think, after serious reflection, it is approximately correct (see, for instance, Johann Frick, 'Conditional Reasons and the Procreation Asymmetry', *Philosophical Perspectives*, 31 (2020) 53–87, and Melinda Roberts, 'The Asymmetry, A Solution', *Theoria*, 77 (2011) 333–367). In framing the debate as being about future people, MacAskill has distractingly mauled a strawman. The more accurate, but difficult, premise for him to have used would be 'hypothetical people count'; readers may have raised their eyebrows at this.

What's more, as MacAskill observes, identity is fragile (p. 173). Our actions today will change who exists later. If we enact some policy, then Angela will never exist, but Bob will. The result is that we cannot make the far future better *for anyone in particular*: it is not better for Angela not to exist, and it is not good for Bob to exist (intuitively, existence is never better for a person); this is the infamous *non-identity problem*. A direct implication of this, however, is that (3) seems false: we cannot make the lives of future people go better: all we can do is cause someone not to exist and someone else to exist instead. In one sense then, people far away in time are just like those far away in space: we are powerless to help them.

MacAskill is mindful of the non-identity problem – though not perhaps the trouble it poses for (3) – and ultimately opts, as noted, for the Total View. Assuming Bob would have a happier life than Angela, the Total View delivers the verdict that it is better that we create Bob; this is not better 'for Bob', but we could say it is better 'for the world'. MacAskill's defence of the Total View is brief, about four pages of text. I have some sympathy with this. As he observes, 'there is still deep disagreement within philosophy about what the right view of population ethics is' (p. 186) and he could not expect to convince his critics of the Total View even after a book-length defence. Yet, for exactly this reason, I am unsympathetic to his portrayal of the Total View as uncontroversial. Famously, it entails the *repugnant conclusion*, which I lack the space to explain. MacAskill tells us that he, along with 28 other philosophers, signed a public statement arguing that the fact a view entails the repugnant conclusion is not a decisive reason to reject that view (p. 180). This might give the misleading impression there is a begrudging consensus among philosophers that we must accept the Total View; yet, many (still) regard it as a non-starter. One hardly needs to proclaim something should not be controversial unless it is.

Hence, it seems objectionable to describe the premises as simple and uncontroversial, especially when the readers are primarily non-philosophers who are liable to take MacAskill at his word. I appreciate MacAskill is trying to spare the reader the intricacies of population ethics. Yet, there is a difference between saying 'This is incredibly complicated, but everyone ultimately agrees' and 'This is incredibly complicated, people disagree furiously, and my argument (may) rely on a controversial view I will only briefly defend.'

We now move to my third concern, which is about the argument that we *can* influence the longterm. An issue here is that *WWOTF* does not contain a historical example that convincingly meets MacAskill's own three criteria (significance, persistence, contingency). I will just consider the example MacAskill spends the most space on: that,

https://doi.org/10.1017/S0953820823000109 Published online by Cambridge University Press

if not for the early abolitionists, slavery would have ended later, and may still be with us today.

I am no historian but I find it hard to believe abolition would never have happened. If we consider other rights movements – womens' rights, LGBT+ rights, animal rights – it seems that lots of people had the same ideas around the same time. If Peter Singer had not written *Animal Liberation* in 1975, would these ideas never have occurred to anyone else? The contingency of these issues seems low, on the order of decades. Suppose, generously, the early abolitionists caused slavery to end 300 years earlier. That would be a magnificent achievement, but disappointingly short-term if we wanted to show, as MacAskill does, we can contingently influence the future one million years hence.

I would like to have seen MacAskill explore the extent to which the case for longtermism depends on being able to point to past successes. We should be sceptical that we will succeed if others have failed – unless we can identify what is different now. One claim in *WWOTF* is that, due to the prospect of AGI, we are living at a unique time. Yet, this sounds worryingly like special pleading; do not people always think their time is special? Indeed, in other work, MacAskill himself argues against the 'hinge of history' hypothesis, the notion we are living through a pivotal period (MacAskill, 'Are We Living at the Hinge of History?', in Jeff McMahan and others (eds), *Ethics and Existence* (Oxford University Press, 2022, 331–357)). I also doubt we are living at the hinge. Unless we are, or unless there is good evidence that others have successfully shaped the past, I do not see why we should expect that we can shape the future. Changing the present is challenging but within our grasp; changing the far future seems beyond it.

It is worth pressing a worry about wishful thinking here: could the seductive allure of longtermism come from the fact we would like to believe we can alter history, not that we genuinely can? When people say they want to 'change the world', we might roll our eyes. Those who claim they want to change the whole course of history – persistently, significantly, and contingently – must merit an incredulous stare. Longtermism looks dangerously like altruism for megalomaniacs.

Fourth and finally, would longtermism, if true, amount to a 'moral revolution – one with far-reaching implications for how […] all of us should think and act' (p. 9)?

Whilst longtermism greatly expands the moral circle in theory, it is much less clear what it changes in reality. The specific priorities MacAskill gives – focus on AGI, pandemics, wars, liberal values, and so on – are all things we already have reason to care about. Longtermism might give us *more* reason to care about them but it is neither necessary nor sufficient to make those the priorities (indeed, with the arrival of GPT-4 between the publication of *WWOTF* and the time of this writing, AGI now seems a very present concern). We need a further argument about how accepting longtermism alters the priorities. I am not sure developing new research fields – in longtermism and AGI safety – constitutes a moral revolution. The general challenge for longtermism is that we cannot confidently see far into the future. Hence, especially if we follow MacAskill's advice of focusing on things that seem robustly good now, it is hard to see how our near- and long-term priorities would radically diverge.

MacAskill opens and closes *WWOTF* by painting longtermism as an expedition into the unknown (p. 6, 226). I ended up feeling about longtermism much the way I would if a cheery Victorian explorer tapped me on the shoulder and invited me to join his trip to discover El Dorado: it sounds exciting, noble, but also unwise. Why should I go on this adventure when its destination is unknown, its prospects uncertain, and its value

https://doi.org/10.1017/S0953820823000109 Published online by Cambridge University Press

unclear, particularly when there is so much we can do here and now? Good luck – but I think I will sit this one out, thanks.

Financial support. This research was funded by a grant from the Forethought Foundation as well support from the Wellbeing Research Centre, Oxford University and the Happier Lives Institute. Open Access publication was jointly funded by the Wellbeing Research Centre and the Happier Lives Institute.

doi:10.1017/S0953820823000109

Michael Pelczar, *Phenomenalism: A Metaphysics of Chance and Experience*

(Oxford: Oxford University Press, 2023), pp. xiii + 210.

Jonathan Riley

Philosophy, Tulane University, New Orleans, LA, USA

Michael Pelczar offers a rare defense of phenomenalism, a metaphysical theory outlined by John Stuart Mill in his study of Sir William Hamilton's philosophy but now largely ignored. Pelczar is enthusiastic about phenomenalism and seeks to restore it to a prominent place in the literature from which it has largely vanished after falling into disrepute about mid-twentieth century. As he remarks, Mill was influenced by traditional idealists such as Leibniz and Berkeley and by Kant's critical idealism, although mention should also be made of the influence of the British tradition of hedonistic associationist psychology as it gradually emerged in the works of empiricists including Locke, Hume, Hartley, James Mill, and Bain. In any case, Mill's phenomenalism is a distinctive metaphysics, which he calls the "psychological theory" and contrasts with Kantian "realism" so-called because Kant posits the reality of noumena or things-in-themselves. Pelczar assumes that noumena have for Kant some power to shape our experience. But Kant is commonly read as maintaining that we do not and cannot know anything about noumena, including whether they play any role in our experience of phenomena.

According to Mill, a physical object such as a table can be re-described as a "permanent possibility of sensation" which, when perceived by a conscious person, interacts with her physical nervous system (which itself requires translation into phenomenalist terms) to produce in her a group of sensations (of extension, shape, color, touch, perhaps pleasure or pain, and so on) appearing more or less simultaneously. The permanent possibility is identified only by the sensations it makes possible, and it exists for as long as does the physical object which it merely re-describes. When not perceived, the possibility continues to exist. It can be perceived again at any time to yield another group of very similar sensations exhibiting the same regularity, although the sensations are not in fact the same ones as before since sensations are fugitive feelings. When the possibility or, in other words, the object is perceived, the sole information received by the conscious subject is the group of simultaneous sensations or perhaps only a part of