William MacAskill

What We Owe the Future: A Million-Year View

London: Oneworld Publications, 2022; hardback, 352 pp., £20.00,
ISBN: 9780861546138

B.V.E. Hyde

Moral circle expansion has been occurring faster than ever before in the last forty years, with moral agency fully extended to all humans regardless of their ethnicity, and regardless of their geographical location, as well as to animals, plants, ecosystems and even artificial intelligence. This process has made even more headway in recent years with the establishment of moral obligations towards future generations. Responsible for this development is the moral theory – and its associated movement – of longtermism, the bible of which is *What We Owe the Future* (London: Oneworld, 2022) by William MacAskill, whose book *Doing Good Better* (London: Guardian Faber, 2015) set the cornerstone of the effective altruist movement of which longtermism forms a part.

Longtermism was perhaps first brought to prominence by Toby Ord in *The Precipice* (London: Bloomsbury, 2020) who defined it as a 'moral re-orientation toward the vast future' (p. 52). Longtermists argue that the (utilitarian) principle of impartiality, or the equal consideration of interests, means that, as Peter Singer, perhaps the principal utilitarian philosopher of our time, says: 'it makes no moral difference whether the person I can help is a neighbor's child ten yards away from me or a Bengali whose name I shall never know, ten thousand miles away' (*Philos. Public. Aff.* vol. 1, no. 3, pp. 229-243; 1972). For Mr. MacAskill, 'distance in time is like distance in space' (p. 10) so, if we are to care about a Bengali ten thousand miles away, then we ought to care about one ten thousand years into the future.

There are some problems with the utilitarian principle of impartiality – and they are not new problems either – none of which are mentioned by Mr. MacAskill, but he seems to be aware of them, because he clunkily adds to his justification of longtermism a deontological principle completely opposed to utilitarianism. He says of future generations that, 'if we recognize they are real people… then we have a duty to consider how we might impact the world they inhibit' (p. 19). This is a rehashed version of Immanuel Kant's 'formula of humanity' which he laid out in the *Groundwork of the Metaphysics of Morals* (Riga:

ISSN (print) 2281-8618-ETS

Johann Friedrich Hartknoch, 1785): 'act that you treat humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means' (p. 429). It does not seem to strike Mr. MacAskill as problematic that Immanuel Kant was referring to conscious persons with moral autonomy who are, crucially, alive, and not to the mere idea of possible people who do not exist but might yet still.

Mr. MacAskill thinks that there is a 'tyranny of the present over the future' that needs to be toppled (p. 9). However, one of the chief difficulties for longtermism is that future people do not exist yet, so he must justify why it is good to make happy people. To do so, he tackles the 'intuition of neutrality' (p. 171) which is, in the words of Jan Narveson (*Monist*, vol. 57, no. 1, pp. 62-86; 1973), that 'we are in favour of making people happy, but neutral about making happy people' (p. 80). Mr. MacAskill has four arguments against this intuition.

The first argument begins with the assumption that our intuition is asymmetrical, meaning that we are indifferent about creating happy people but believe it is morally wrong to bring a miserable new person into existence. If our intuitions truly exhibit this asymmetrical nature, then any argument supporting the notion that it is wrong to create an unhappy person should also apply to the idea that it is good to bring a happy person into the world (p. 172).

The second argument is simply that, because it is intuitive to him that the future is better because of the existence of his happy nephews and nieces, it follows that the world is in fact better with the creation of happy people (p. 172).

The third argument departs from the previous two by relying on empirical findings instead of logical reasoning. He refers to a recent study in psychology that discovered that our moral intuitions regarding the creation of happy or unhappy individuals are actually symmetrical, suggesting that we generally believe it is positive to bring happy people into existence and negative to bring unhappy ones (*Cognition*, vol. 218, art. 104941; 2022).

The fourth argument is that, because a minor shift in timing could have led to a different individual being born instead of you, the sperm responsible for your existence having only a one in two hundred million chance of fertilizing an egg, we are 'like clumsy gods' (p. 174), dramatically altering history's trajectory with each passing moment. From what he calls the 'fragility of identity; (p. 173), the implication is that today's policies will impact the future, not by enhancing the lives of people who would have existed regardless, but by *creating* a new future with individuals who are somewhat happier. Moreover, because it is intuitive that we have indeed improved the future, it must be true that adding people with happier lives is good, thereby disproving the intuition of neutrality (p. 176).

The fifth and final argument offered by Mr. MacAskill is the most sophisticated, but it is not his anyway: he admits by way of an endnote that he takes it

from John Broome's book, *Weighing Lives* (Oxford: Oxford University Press, 2004). Say that in world$_1$ you are not born, in world$_2$ you live in suffering and in world$_3$ you live blissfully. The intuition of neutrality says that world$_1$ is neither better nor worse than world$_2$, which means that world$_1$ is equal in value to world$_2$. What licenses this inference is that John Broome assumes the comparative value relation is complete (§10.1), which means that if something is neither better nor worse than something else, the two are equally as good as one another. From the intuition of neutrality it also follows that world$_1$ is just as good as world$_3$. If values are transitive, which Mr. MacAskill assumes they are, then it follows that world$_2$ is just as good as world$_3$ which, according to Mr. MacAskill, is a 'contradiction' because it cannot be the case that creating a life of suffering is just as good as creating a life of bliss (p. 177). Therefore, it must be good to create good lives.

None of these arguments are sound. More than one begs the question. The strongest is the evidential one, but it does not follow from evidence that we *do* think it good to create happy people that we *should* think so. This runs afoul of David Hume's law, which he explicated in *A Treatise of Human Nature* (London: John Noon, 1739): that no moral statement can be inferred from non-moral ones (bk. iii, pt, i, §1).

Because creating good lives is good, Mr. MacAskill recommends that we ought to have children (p. 187) and to ensure that civilization lasts as long as possible and is as big as possible (p. 188). The bigger the future, the better the future, which is why 'the early extinction of the human race would be a truly enormous tragedy' (p. 189). This is why Mr. MacAskill argues that we are morally obliged to mitigate existential risks, which Nick Bostrom defines as a threat to the premature extinction of intelligent life on earth or the permanent and drastic destruction of its potential for desirable future development (*Global Policy*, vol. 4, no. 1, pp. 15-31; 2013).

The principal existential risks are, according to Mr. MacAskill, engineered pathogens (p. 107), war between great powers (p. 114), climate change (p. 134) and fossil fuel depletion (p. 138). Many futurological researchers, like Mr. Bostrom, in *Superintelligence* (Oxford: Oxford University Press, 2014), are most concerned by existential risk from artificial general intelligence, where humans could be replaced as the dominant lifeform on earth were machine brains to surpass human brains and become superintelligent. Some are skeptical of this alarmism, like Michio Kaku who, in *Physics of the Future* (New York: Doubleday, 2011), said that he believed we will find intelligent robots benevolent and friendly. Mr. MacAskill is both an alarmist and an optimist, for he believes that artificial intelligence might wipe out the human race, but that it still represents intelligent life with moral value, so even its destruction of humanity would not

be a crisis so long as the artificial civilization that advances into the future is not morally bankrupt (p. 87).

Despite the threat of annihilation, Mr. MacAskill thinks that we should be optimistic about the future (p. 193), in part because the world is already good. Mr. MacAskill commissioned psychologists to run a survey which found that, although around 10% of the global population have lives below neutral well-being, most people have positive lives (p. 201). Moreover, the world is getting better. Richard A. Easterlin published a very famous study in a chapter in *Nations and Households in Economic Growth* (New York: Academic Press, 1974) in which he showed that people and countries do not get happier as they get richer over time. However, it has since been revealed that the Easterlin Paradox does not exist. More recent work with better data strongly supports the hypothesis that countries get happier as they get richer (*Brook. Pap. Econ. Act.* no. 1, pp. 1-87; 2008). Likewise, contrary to the common belief, originating with the psychologist Philip Brickman and his colleagues (*J. Pers. Soc. Psychol.* vol. 36, no. 8, pp. 917-927; 1978), that lottery winners are unhappy, Andrew Oswald and Rainer Winkelmann have shown in a chapter in *The Economics of Happiness* (Cham: Springer, 2019) that winning the lottery does increase one's happiness. If the world continues to get richer, we can expect the future to be even happier.

The future can only be good if good values permeate it, though. Values, Mr. MacAskill thinks, can persist for extremely long periods of time through 'value lock-in' (p. 78), of which Confucian influences on the Orient today and Christian influences on the modern Occident are exemplary. The permanence of values is determined by an 'early plasticity, later rigidity' cycle (p. 42). According to Mr. MacAskill, history is like glass that is sometimes hot and sometimes cold. When it is hot, it can be reshaped, but the colder it gets the harder it becomes. As Derek Parfit wrote in his book *On What Matters* (Oxford: Oxford University Press, 2011), we 'live during the hinge of history' (vol. 2, p. 611). The present age is one of plasticity, but longtermists warn that a period of rigidity is on the horizon. What will cause it, Mr. MacAskill says, is artificial intelligence: because it is immortal and has the potential to cause rapid technological progress, whatever values it holds, or whatever values are instilled within it, could last a very long time (p. 83). This means that our values could define the future, which is why changing them for the better is one of the most important longtermist tasks (p. 52).

Really, we should try to avoid value lock-in (p. 88) and have a 'long reflection' (p. 98) where we can work out what a flourishing society would look like. This should give us a 'morally exploratory world' in which better morals win over time such that we converge on the best society (p. 99). There are a few things we need to do to avoid value lock-in. One: we must prioritize the prevention of value lock-in, even at the expense of delaying advancement such as space ex-

ploration or development of artificial intelligence. Two: we must be politically experimental and ensure that our society is culturally and intellectually diverse to avoid premature convergence. Three: we must somehow ensure that cultural evolution results in moral evolution. What we end up with is a 'lock-in paradox' (p. 101): we need to lock-in some institutions and values to prevent a more thoroughgoing lock-in of values.

What We Owe the Future is a well-researched book, bringing to attention lots of diverse and interdisciplinary evidence, interesting facts, and historical cases to support its arguments. It also contains some original empirical research, and several well-designed illustrations have been produced to make some of the more challenging aspects of the book easier to understand and to make some of the more grandiose claims seem even more impactful. The book has its own website (whatweowethefuture.com) where the bibliography is to be found, rather than in the book, which is odd and worth mentioning. The website also contains some supplements, press about the book, and links to established effective altruist organizations that readers are pointed towards in the book, like 80,000 Hours and the Longtermism Fund. Clearly, lots of hard work has gone into the book.

Lots of it, though, is not Mr. MacAskill's. He admits that the book is an extremely collaborative effort and even that 'many sections of the book were essentially coauthored' (p. 247). If you compare his enormous acknowledgements section with the American Psychological Association (APA) author determination guidelines, you might be surprised that only one name is on the front cover of the book. Even the more stringent International Committee of Medical Journal Editors (ICMJE) recommendations would suggest that some of those acknowledged have been cheated out of authorship. Mr. MacAskill is really the book's editor, not the sole author, and there is definitely a looming question over it about the extent to which his claim to sole authorship represents a questionable research practice. You get the impression that the research for the book was done by a team of researchers, whereas the philosophic arguments are the work of the one, which is perhaps why the historical work is much more impressive than the philosophic, which is not well thought out at all.

In fact, Mr. MacAskill's arguments for longtermism represent some of the poorest for what is perhaps the most popular philosophical movement in the world right now. He argues almost entirely by catching the reader in a provocative literary style that has captured so many established academic celebrities like Stephen Fry and Sam Harris. It is the same attractive, optimistic style that was applauded by reviewers such as Amia Srinivasan with respect to some of his earlier books (*Lond. Rev. Books*, vol. 37, no. 18; 2015). But not everyone has been caught in the excitement conjured up by Mr. MacAskill, and some other

reviewers have also criticized his book for being 'replete with highfalutin truisms, cockamamie analogies and complex discussions leading nowhere' (*Wall Str. J.* 26 August 2022) – to which we must add appeals to intuition, inferences from anecdotal evidence, unjustified assumptions, question begging and, of course, the intellectual crime of utter thoughtlessness: more than half the time, Mr. MacAskill is totally unaware of the positions he is committing himself to, and he often prefers cheap tricks in place of proper philosophic argumentation.

Laid plain of his alluring narrative, there is no philosophic substance to the book in the slightest. It is just another episode in the rehashing of an old and outworn utilitarian theory in a contemporary jacket. The ethical wing of effective altruism and longtermism, as they both currently stand, is nothing but utilitarianism with a vocabulary updated to include buzzwords like climate crisis, global poverty, and artificial intelligence. Perhaps these positions on ethics, philanthropy and global priorities can be put right, but it is very unfortunate that a foundational text is so inadequate; in this regard, the movement's future looks bleak, and it will be forced to choose between objectivity and dogma at its current rate. What we owe the future is a better explanation. Or, at least, William MacAskill does.