Three mistakes in the moral mathematics of existential risk

Abstract

Longtermists have recently argued that it is overwhelmingly important to do what we can to mitigate existential risks to humanity. I consider three mistakes that are often made in calculating the value of existential risk mitigation: focusing on cumulative risk rather than period risk; ignoring background risk; and neglecting population dynamics. I show how correcting these mistakes pushes the value of existential risk mitigation substantially below leading estimates, potentially low enough to threaten the normative case for existential risk mitigation. I use this discussion to draw four positive lessons for the study of existential risk: the importance of treating existential risk as an intergenerational coordination problem; a surprising dialectical flip in the relevance of background risk levels to the case for existential risk mitigation; renewed importance of population dynamics, including the dynamics of digital minds; and a novel form of the cluelessness challenge to longtermism.

1 Introduction

Suppose you are an altruist. You want to do as much good as possible with the resources available to you. What might you do? One option is to address pressing short-term challenges. For example, GiveWell (2021) estimates that $5,000 spent on bed nets could save a life from malaria today.

Recently, a number of longtermists (Greaves and MacAskill 2021; MacAskill 2022b) have argued that you could do much more good by acting to mitigate existential risks: risks of existential catastrophes involving "the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development" (Bostrom 2013, p. 15). For example, you might work to regulate chemical and biological weapons, or to reduce the threat of nuclear conflict (Bostrom and Ćirković 2011; MacAskill 2022b; Ord 2020).

Many authors argue that efforts to mitigate existential risk have enormous value. For example, Nick Bostrom (2013) argues that even on the most conservative assumptions, reducing existential risk by just one-millionth of one percentage point would be

as valuable as saving a hundred million lives today. Similarly, Hilary Greaves and Will MacAskill (2021) estimate that early efforts to detect potentially lethal asteroid impacts in the 1980s and 1990s had an expected cost of just fourteen cents per life saved. If this is right, then perhaps an altruist should focus on existential risk mitigation over short term improvements.

There are many ways to push back here. Perhaps we might defend population-ethical assumptions such as neutrality (Naverson 1973; Frick 2017) that cut against the importance of creating happy people. Alternatively, perhaps we might introduce decision-theoretic assumptions such as risk aversion (Pettigrew 2022), ambiguity aversion (Buchak forthcoming) or anti-fanaticism (Monton 2019; Smith 2014) that tell against risky, ambiguous and low-probability gambles to prevent existential catastrophe. We might challenge assumptions about aggregation (Curran 2022; Heikkinen 2022), personal prerogatives (Unruh forthcoming), and rights used to build a deontic case for existential risk mitigation. We might discount the well-being of future people (Lloyd 2021; Mogensen 2022), or hold that pressing current duties, such as reparative duties (Cordelli 2016), take precedence over duties to promote far-future welfare.

These strategies set themselves a difficult task if they accept the longtermist's framing on which existential risk mitigation is not simply better, but orders of magnitude better than competing short-termist interventions. Is it really so obvious that we should not save future lives at an expected cost of fourteen cents per life? While some moves, such as neutrality, may carry the day against even astronomical numbers, many of the moves on this list would be bolstered when joined with a competing maneuver: questioning the longtermist's moral mathematics.

In this paper, I argue that many leading models of existential risk mitigation systematically neglect morally relevant considerations in determining the value of existential risk mitigation. This has two effects. First, debates about the value of existential risk mitigation are mislocated, because many of the most important parameters are neither modeled nor discussed. Second, the value of existential risk mitigation is inflated by

many orders of magnitude. I look at three mistakes in the moral mathematics of existential risk: mishandling of cumulative risk (Section 3), background risk (Section 4), and population dynamics (Section 5). This will help us to gain a better understanding of the factors relevant to valuing existential risk mitigation. And under many assumptions, once these mistakes are corrected, the value of existential risk mitigation will be far from astronomical.

Reflecting on these mistakes in the moral mathematics of existential risk raises at least four classes of positive lessons for longtermism and the study of existential risk, discussed in Section 5. There, we will see the importance of treating existential risk mitigation as a difficult intergenerational coordination problem (Section 6.1); a surprising dialectical flip in the relevance of background risk levels to the case for existential risk mitigation (Section 6.2); renewed importance of population dynamics, including the demographics of digital minds (Section 6.3); and a novel form of the cluelessness challenge to longtermism (Section 6.4). But first, let us begin with some clarificatory remarks (Section 2).

2 Preliminaries

Before beginning, it is important to note three points. First, throughout this discussion I work within a broadly totalist population axiology and take a broadly consequentialist approach to valuing acts that affect existential risk. I do this because such an approach is often thought to be most charitable to the advocate of existential risk mitigation, as well as to separate my argument from the other complementary routes of resistance outlined in Section 1.

Second, the models in this paper treat all existential risks as risks of human extinction. I do not model existential catastrophes which permanently curtail the potential for desirable future development without causing outright extinction. I do this for three reasons. First, many of the estimates that I reconstruct have the same restriction, and I do not want to be accused of fiddling with modeling assumptions. Second, on many views the value

realized after non-extinction catastrophes may be comparatively small, and hence may not significantly affect the value of existential risk mitigation. Finally, non-extinction catastrophes introduce a good deal of modeling complexity that this paper strives to avoid. However, readers are welcome to treat the conclusions of this paper as conclusions about extinction risk, leaving open the possibility of saying something different about existential risk more broadly.

Finally, recent discussions have revealed one way of maintaining a high value for existential risk mitigation against challenges. This is the Time of Perils Hypothesis on which existential risk, though high now, will soon drop to a permanently low level (Ord 2020; Thorstad forthcoming). This paper will not be concerned with the Time of Perils Hypothesis, except for a brief discussion in Section 6.2.

With these clarifications in mind, let us turn to the first mistake in the moral mathematics of existential risk: focusing on cumulative risk.

3 Cumulative risk

To illustrate the importance of existential risk reduction, Nick Bostrom (2013) begins with what he terms a 'conservative' scenario on which humanity remains Earth-bound at a population of one billion people.[1] We remain in this state for another billion years, after which Earth becomes less habitable. In this scenario, the future holds approximately 10^{18} years of human life, or 10^{16} lives if we assume an average lifespan of a hundred years. Bostrom uses this example to make a striking claim:

> Even if we use . . . conservative estimates, which entirely ignor[e] the possibility of space colonization and software minds, we find that the expected loss of an existential catastrophe is greater than the value of 10^{16} human lives. This

[1] In the next two sections, we will see that this is far from a conservative scenario: humanity is likely to go extinct far sooner than a billion years from now (Section 4) and standard demographic models give a significant chance that the human population will fall well below one billion (Section 5).

> implies that the expected value of reducing existential risk by a mere one-millionth of one percentage point is at least a hundred times the value of a million human lives. (Bostrom 2013, pp. 18-19).

Here Bostrom claims that if we assume the Earth can support at least 10^{16} future human lives, a small reduction of 10^{-8} in the probability of existential catastrophe yields, in expectation, at least the value of 10^8 human lives. That is a very large number.

It can seem obvious that Bostrom's claim is correct to a reasonable degree of approximation. After all, if there are 10^{16} future lives that can be lived, then a reduction of 10^{-8} in the chance of human extinction should produce, in expectation, $10^{16} * 10^{-8} = 10^8$ additional future human lives. However, although this claim is not strictly speaking false, it is misleading.

To see why this claim is misleading, we need to introduce two distinctions. The first is merely clarificatory.[2] To say that risk r has been reduced by amount f can be understood in two ways. Typically, what we have in mind is *relative risk reduction*, where r is reduced multiplicatively by a fraction f of its original amount, to $(1 - f)r$. It is in this sense that we speak, for example, of a 20% reduction in risk as a reduction in risk from 100% to 80% or from 10% to 8%. Sometimes, what we have in mind is *absolute risk reduction*, where f is subtracted from the original risk, leaving resulting risk $r - f$. It is in this sense that we speak, for example, of a 20% reduction in risk as a reduction from 100% to 80% or from 30% to 10%. For Bostrom's claim to make sense, Bostrom must be concerned with absolute rather than relative risk reduction. It is in this sense that a reduction of 10^{-8} in risk produces, in expectation, 10^{-8} times the value of an outcome free from catastrophe.

Second and more relevantly, existential risks are repeated risks: they recur throughout many periods of human existence. Repeated risks can be represented in two different ways. On the one hand, we can specify the *period* risk r_P of existential catastrophe occurring in each period over a long interval. For example, we might specify the per-

[2]This isn't quite true, since absolute risk deflates the required reduction by a factor of $1/r$, where r is the starting level of risk. However, when r is large the effect will be moderate.

century risk of catastrophe r_P at 20%, 1%, or 0.01%. On the other hand, we can report the *cumulative risk* r_C of existential catastrophe occurring at least once during the total interval. Over a period of N centuries, the cumulative risk is $r_C = 1 - (1 - r_P)^N$. Bostrom's claim makes sense only if he is concerned with cumulative risks, for it is in this sense that we can treat an absolute risk reduction of 10^{-8} as providing a 10^{-8} chance of saving all of the future lives that the Earth can support.

The problem is that over a long period of human existence, seemingly small reductions in cumulative risk may require astronomically large reductions in period risk. To illustrate the point, note that an absolute reduction of 10^{-8} in cumulative existential risk would bring about a probability of at least 10^{-8} that humanity survives for a billion years. However, the probability of surviving for a billion years, or ten million centuries, depends on the cumulative risk r_P: we survive for ten million centuries with probability $P(S) = (1 - r_P)^{10,000,000}$. For our cumulative survival chance $P(S)$ to exceed the seemingly small probability 10^{-8} requires an extremely low per-century risk of $r_P \approx 1.6 * 10^{-6}$, barely a one-in-a-million risk of existential catastrophe per century.

This discussion reveals two problems with discussing cumulative rather than period risk. First, seemingly small reductions in cumulative risk in fact require astronomical reductions in period risk. In this case, an absolute reduction of one-millionth of a percent in cumulative risk actually requires in the very best case driving per-century risk down to a nearly one-in-a-million chance. By contrast, many longtermists estimate existential risk in this century in the neighborhood of fifteen to twenty percent.[3] This means that in considering a seemingly small absolute drop of 10^{-8} in cumulative risk, Bostrom may actually be considering an unprecedentedly large drop in per-century risk: per-century risk must be driven one hundred thousand times lower than present values. Reframing the matter in terms of per-century rather than cumulative risk helps us to see that there is a clear sense in which the drop in existential risk that Bostrom envisions is not small, but

[3]Ord (2020) puts risk at 16.6%; attendees of the 2008 Global Catastrophic Risks Conference at the Future of Humanity Institute gave a median estimate of 19% (Sandberg and Bostrom 2008); and the Astronomer Royal Martin Rees puts the chance of civilizational collapse at 50% by the end of the century (Rees 2003).

instead very large.[4]

Second, it is misleading to discuss changes in cumulative risk because cumulative risk is often not under the control of current generations. While there may be things that we can do to decrease existential risk in our own century or even in nearby centuries, there may be less we can do to decrease existential risk a thousand or a million centuries from now. And in particular, it is doubtful that we can drop far-future risks by many orders of magnitude through our actions today. For this reason, framing matters in terms of the required reduction in cumulative rather than period risk distorts the fact that humans acting to reduce existential risk are typically in a position to reduce period risk in nearby centuries, but less able to reduce cumulative risk over a long period of time.

For these reasons, our first mistake in the moral mathematics of existential risk is studying cumulative rather than period risks. As we have seen, the turn to cumulative risk misleadingly casts large reductions in period risk as small reductions in cumulative risk, and puts too much emphasis on cumulative risks that humanity today cannot effectively control rather than studying risks in nearby centuries that we can control. We will see in the next section that this first mistake in the moral mathematics of existential risk is compounded when it is combined with a second mistake: ignoring levels of background risk. Let us begin by considering a well-known discussion which makes both mistakes.

4 Background risk

Many authors take biosecurity threats such as bioterrorism and laboratory leaks to be among the primary sources of existential risk for humanity (MacAskill 2022b; Ord 2020; Sandberg and Bostrom 2008). The most-cited discussion of existential risks related to biosecurity ('biorisks') is due to Piers Millett and Andrew Snyder-Beattie (2017). Millett and Snyder-Beattie use a range of models to suggest that efforts to reduce biorisk are cost-effective.

[4]Here we also see that the framing in terms of absolute risks may be misleading, since an absolute risk reduction of 15-20% works out on this model to a relative risk reduction of about five orders of magnitude.

Table 1: MSB cost-effectiveness estimates

Model	**N (biothreats/century)**	**C/NLR (cost/life-year)**
Model 1	0.005 to 0.02	0.125 to 5.00
Model 2	$1.6 * 10^{-6}$ to $8 * 10^{-5}$	31.00 to 1,600
Model 3	$5 * 10^{-5}$ to $1.4 * 10^{-4}$	18.00 to 50.00

The Millett and Snyder-Beattie (MSB) model estimates the cost of an intervention using four parameters. The cost C of a biosecurity intervention is fixed at two hundred and fifty billion dollars, representing a strong societal investment in biosecurity. The number N of biological catastrophes that would be expected to occur each century without intervention is estimated using three different techniques. The number L of human life-years lost in such a catastrophe is estimated using the size of the future human population. Assuming that humanity remains earthbound, with a population of ten billion people for a period of one million years produces an estimate $L = 10^{16}$ of future life-years at stake. Finally, the amount R by which a $250 billion intervention would reduce biorisk is estimated at a relative risk reduction of one percent, from N to $.99N$.

The MSB model combines these separate parameter estimates to model the value of a stylized investment of $250 billion into biosecurity as C/NLR. Estimating lower and upper bounds on the number N of biological catastrophes that occur each century without intervention using three different models yields three cost-effectiveness estimates (Table 1). Because many governments conduct cost-effectiveness analyses that value a life-year at tens of thousands of dollars, Millett and Snyder-Beattie conclude that across models, a stylized biosecurity intervention appears to be cost-effective.

Many complaints could be raised against this model.[5] But for now, I want to raise one new complaint and one old familiar complaint. The new complaint is that the MSB

[5]Here is one complaint: the quantity to be estimated is $E[C/LNR]$, which in the friendliest case simplifies to $C/E[LNR]$ when C is known with certainty. But the MSB model combines separate parameter estimates, and hence calculates $C/(E[L]E[N]E[R])$. This coincides with $C/E[LNR]$ only under the implausible assumption that $Var(LNR) = 0$, implying that LNR is constant. This complaint is closely related to a second complaint: L and N are strongly inversely correlated, since risky futures hold fewer expected lives. For this reason, models such as the MSB model which separately estimate L and N are likely to be inaccurate: L and N must be modeled together.

model ignores background risk. Biological risk is not the only kind of existential risk facing humanity, hence reductions in biological risk are limited in their capacity to drive down overall existential risk, because they cannot affect other background existential risks. Once these other background risks are built into the MSB model, the cost-effectiveness of biosecurity drops in a meaningful and policy-relevant way.

To see the point, assume, following the MSB model, that humanity will maintain a population of 10 billion for the next million years, unless an existential catastrophe occurs, in which case there will be no humans left. Let r be the per-century level of existential risk facing humanity, and assume for simplicity that r takes the same constant value in each century unless we intervene to reduce risk. Letting L be the number of future human life-years, we can express the number of expected future life-years $E[L]$ by multiplying the number of life-years in each century by the probability of surviving to the end of that century without catastrophe:

$$E[L] = 10^{12} \sum_{i=1}^{10,000} (1 - r)^i.$$

What effect do biosecurity interventions have on the expected number of future life-years?

Assuming for simplicity that biological and nonbiological existential catastrophes will not occur in the same century, we can decompose per-century risk as $r = b + n$, where b is the risk of a biological catastrophe and n is the risk of a nonbiological catastrophe. Let X be an action which, following MSB, provides a 1% relative reduction in biological risk, shifting total risk to $r_X = 0.99b + n$. Now, the expected number of future life-years is:

$$E[L|X] = 10^{12} \sum_{i=1}^{10,000} (1 - r_X)^i.$$

In expectation, how many future life-years does X add? It turns out that X adds, in expectation, $10^{12} * \frac{0.01b}{r(r-0.01b)}$ life-years (see Appendix).

Therein lies the rub, for if we assume, as we saw that many longtermists do (Ord 2020;

Table 2: MSB cost-effectiveness estimates against revised model (r = 0.2 and r = 0.01), $/life-year

Model	**N**	**MSB estimate**	**r = 0.2**	**r = 0.01**
Model 1	0.005 to 0.02	0.125 to 5.00	50 to 200	0.25 to 0.50.[6]
Model 2	$1.6 * 10^{-6}$ to $8 * 10^{-5}$	31.00 to 1,600	12,500 to 625,000	30 to 1,500
Model 3	$5 * 10^{-5}$ to $1.4 * 10^{-4}$	18.00 to 50.00	7,100 to 20,000	18 to 50

Rees 2003; Sandberg and Bostrom 2008), that overall existential risk r is is very high, then because r is much greater than b, the denominator dominates and X adds only a much more modest number of life-years in expectation. Table 2 extends Table 1 to report the cost-effectiveness of biosecurity interventions on our improved model of the expected number of future lives, leaving fixed all other elements of the MSB model. I report cost-effectiveness using a pessimistic 20% estimate of per-century risk r, as well as a more optimistic 1% estimate of per-century risk.

Many, but not all of the revised estimates count the stylized biosecurity intervention as cost-effective by leading metrics. Crucially, few of these estimates uncontroversially qualify biosecurity as more effective than leading short-termist interventions, such as global health and anti-poverty. The charity evaluator GiveWell estimates the expected cost of saving a life through the best global health interventions in the range of several thousand dollars (GiveWell 2021), producing a natural short-termist benchmark of about $100 per life-year saved, even if we make the strong concession of ignoring long-term benefits of short-termist interventions.[7] Many of the estimates above fall short of this short-termist benchmark, often by quite a large margin, raising the very real possibility that short-termist interventions may be more cost-effective than biosecurity if we retain the MSB model parameters.

More broadly, this reveals a second mistake in the moral mathematics of existential risk: ignoring background risk. We saw that building levels of background risk r into our

[7]This is a strong concession, since there is broad consensus that anti-poverty and global health efforts contribute to economic development, which in turns drives welfare growth.

Table 3: MSB cost-effectiveness estimates against revised doubly model (r = 0.2 and r = 0.01), $/life-year

Model	**N**	**MSB estimate**	**r = 0.2**	**r = 0.01**
Model 1	0.005 to 0.02	0.125 to 5.00	250 to 1,000	13 to 50.
Model 2	$1.6 * 10^{-6}$ to $8 * 10^{-5}$	31.00 to 1,600	60,000 to 3.1 million	3,000 to 150,000
Model 3	$5 * 10^{-5}$ to $1.4 * 10^{-4}$	18.00 to 50.00	35,000 to 100,000	1,800 to 5,000

discussion of biorisk mitigation tends to substantially drop the expected value of biorisk reduction.[8] For this reason, it is important to consider background risk when modeling the value of existential risk mitigation.

This second mistake in moral mathematics is compounded by our first mistake within the MSB model. We can further reduce the estimated cost-effectiveness of biosecurity interventions by recalling the need to focus on period risk rather than cumulative risk. MSB ask how valuable it would be to reduce biological risk by 1% across all time periods. But in Section 3, I argued that because our actions today are unlikely to reduce risk in distant centuries, it is more revealing to ask how valuable it would be to reduce risk in nearby centuries.

Suppose we reframe the question in these terms: how cost-effective is an act X' which reduces biological risk by 1% in our own century, but leaves risk in future centuries unchanged? It turns out that the expected number of future lives given this intervention is $E[L|X'] = 10^{12} * 0.01b/r$.[9] This quantity is smaller than before, since r is less than one and the denominator has gone from roughly r^2 to precisely r. In particular, because background risk r counts once rather than twice in this expression, the effect of diminishing background risk is substantially reduced in this expression (Table 3).[10]

Now the stylized biosecurity intervention fails standard tests for cost-effectiveness

[8]For a full treatment of this issue, see (Thorstad forthcoming).

[9]To see this, apply the formula for the value of absolute risk reduction from (Thorstad forthcoming), Section 3.1, with v = 10^{12} and $f = 0.01b$.

[10]In Model 1, the case $r = 0.01$ was calculated using an upper bound of $b = 0.01$ to respect the constraint $b \leq r$.

on several models, and falls short of our low-ball short-termist benchmark of $100/life-year across almost all models, often by many orders of magnitude. This is a powerful illustration of how several mistakes in moral mathematics may combine to change the qualitative nature of our model's recommendations.

So far, we have considered two mistakes in moral mathematics: focusing on cumulative risk rather than period risk, and ignoring background risk. We saw how these mistakes combine to inflate estimates of the cost-effectiveness of existential risk mitigation. However, both of these assumptions treated the size of the future human population as fixed and large. This is, as we will see, a third mistake in the moral mathematics of existential risk: ignoring population dynamics.

5 Population dynamics

5.1 Introduction

Longtermists often remind us that the future human population could, in principle, be very large. We saw in Section 3 that Nick Bostrom treats a population of 10^{16} future individuals as a conservative estimate.[11] Dramatizing the point, Will MacAskill (2022b) asks us to view the human population as 'stick figures' of ten billion people each. To date, roughly ten 'stick figures' of humans have existed. However, MacAskill continues, if humanity survives on Earth at a population of ten billion people for another five hundred million years, with an average lifespan of a hundred years, there are five million 'stick figures' of humans yet to come, enough to populate twenty-thousand pages with about 250 stick figures on each page.

A similar argument is made by Hilary Greaves and Will MacAskill (2021), drawing on Newberry (2021). Greaves and MacAskill consider varying estimates of the carrying ca-

[11] Greaves and MacAskill (2021) go further, holding that any reasonable estimate of the expected number of future lives should be at least 10^{24}.

Table 4: Estimates of future lives based on duration and carrying capacity

Scenario	Carrying capacity (Lives/century)	Duration (centuries)	Future lives
Earthbound (Bostrom)	10^9	10^7	10^{16}
Earthbound (MacAskill)	10^{10}	$5 * 10^6$	$5 * 10^{16}$
Earthbound (Greaves/MacAskill)	10^{10}	10^4	10^{14}
Solar System (Greaves/MacAskill)	10^{19}	10^8	10^{27}
Milky Way (Greaves/MacAskill)	10^{25}	10^{11}	10^{36}

pacity of regions that humanity might settle: the number of humans they could support.[12] For example, we might estimate the carrying capacity of Earth at ten billion lives per century; the carrying capacity of the solar system at 10^{19} lives per century; and the carrying capacity of the Milky Way at 10^{25} lives per century. Next, Greaves and MacAskill estimate the duration for which humanity might exist in such a region. Finally, they multiply the duration of human life with the carrying capacity of each region to estimate how many future humans might live there. Table 4 reports the Greaves and MacAskill estimates together with Bostrom (2013) and MacAskill's (2022b) estimates, which are calculated in the same way.

In Section 4, we saw one way to push back against such numbers: model background risk. Let's make the generous modeling assumption that these populations are evenly distributed throughout time.[13] Next, recalculate the expected number of future lives across levels of per-century existential risk r as in Section 4. Now, matters look very different (Table 5). Across all models, the number of expected future lives has substantially dropped, and at high levels of risk it struggles to grow significantly above carrying capacity.

In this section, I want to explore a complementary strategy for pushing back. The

[12]Greaves and MacAskill also consider the number of *digital* lives that these regions can support. I do not consider the prospects for digital lives in this section, both because there are a number of open questions about the possibility, likelihood and value of creating digital lives, and also because the population dynamics of digital populations might be quite different from the population dynamics of human populations.

[13]In reality, on the last two models populations are much larger later on, but ignoring this fact can only help the longtermist.

Table 5: Revised estimates of future lives after incorporating background risk

Scenario	Original estimate	r = 0.2	r = 0.01	r = 0.001
Earthbound (Bostrom)	10^{16}	$4 * 10^{9}$	10^{11}	10^{12}
Earthbound (MacAskill)	$5 * 10^{16}$	$4 * 10^{10}$	10^{12}	10^{13}
Earthbound (Greaves/MacAskill)	10^{14}	$4 * 10^{10}$	10^{12}	10^{13}
Solar System (Greaves/MacAskill)	10^{27}	$4 * 10^{19}$	10^{21}	10^{22}
Milky Way (Greaves/MacAskill)	10^{36}	$4 * 10^{25}$	10^{27}	10^{28}

models above ask a question about feasible populations: how many lives could the universe hold if we managed to stuff it to carrying capacity? But scientific estimates of future population size do not ask this question. The field of demography asks instead how human populations are likely to evolve over time, given their initial state and relevant factors such as cultural norms, economic conditions and technology. Most demographers now think that the future human population is likely to be far smaller than any of the projections above, because it is unlikely to hover permanently near carrying capacity.

Below, I consider the standard story about future population growth (Section 5.2). On this story, it turns out that the future human population may not be significantly larger than the present population. Then I consider a highly optimistic model, which comes largely uncoupled from current scientific projections, to show how even under the most optimistic assumptions, building population dynamics into models tends to substantially lower the value of existential risk mitigation by lowering the expected size of the future human population (Section 5.3).

Throughout, it is worth bearing in mind that the future is uncertain. Although the standard demographic models discussed in Section 5.2 represent our best scientific projections, it is possible that these projections are wrong. While readers are invited to resolve uncertainties using their own background views, two remarks are worth stressing. First, even readers who place nontrivial confidence in optimistic scenarios for future population growth may not place substantial confidence in the views surveyed above, on which population size hovers near carrying capacity. There is a significant gap be-

tween the most optimistic and the most pessimistic population projections, and moving beyond pessimism need not carry us all the way to full optimism. Second, mistakes in the moral mathematics of existential risk gain force in conjunction with one another, and potentially also with complementary normative and descriptive strategies. If it turns out that uncertainty about population dynamics drives the expected size of the future human population a few orders of magnitude lower than estimates based on carrying capacity, then that is reason enough to take population dynamics seriously as part of a broader challenge to the value of existential risk mitigation.

5.2 Standard population models

The models above assume that the worst-case scenario is one in which humanity remains earthbound with a population comparable in size to the current population of eight billion people. However, most demographers think this is unlikely: quite probably, the future human population will decline for some centuries (Basten et al. 2013; Lutz et al. 2014; United Nations 2022) and this decline may well be permanent (Alexandrie and Eden forthcoming; Geruso and Spears forthcoming; Spears et al. 2023).

For most of human history, population growth was dominated by a Malthusian regime: humans wanted to have more children, but were limited by economic factors such as the availability of food and other resources (Malthus 1798). As the economy grew, populations quickly expanded until average incomes approached subsistence level and no more children could be fed. However, rapid industrialization over the past few centuries has brought a different regime. Current technology enables to us support far more than eight billion people, and to support them at a level far above subsistence. Nevertheless, we have not done so. New considerations such as cultural norms, values, and social structures exert increasing influence on fertility rates (Barro and Becker 1989), and these considerations drift apart from the question of how many lives we can in principle support. Factors such as women's empowerment (Sen 1999; United Nations 2022) and labor force participation (Jensen 2012); changing societal messaging about children and family structure

(Jensen and Oster 2009; Kearney and Levine 2015); declining religiosity and the spread of Enlightenment ideals (Lestahege 1992) increasingly push fertility rates downwards.

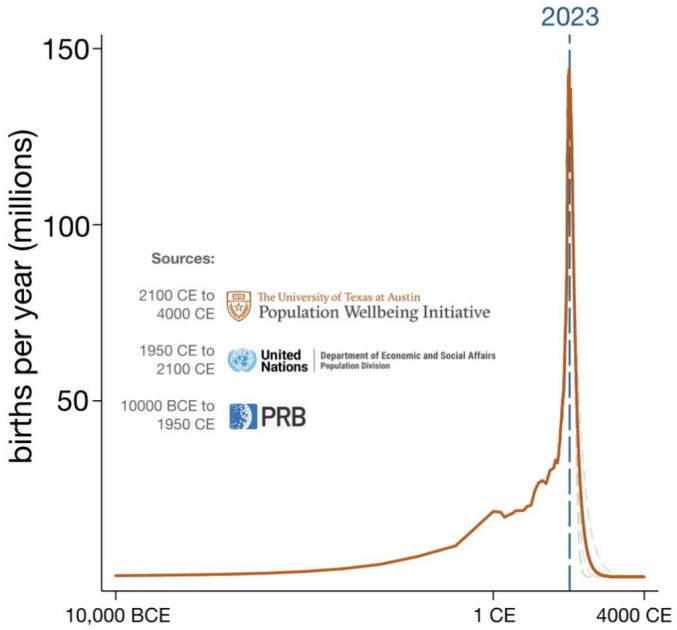
In a post-industrial society, fertility rates have taken a nosedive. Global fertility rates have steadily declined to around 2.3 children per woman, down from over 5 children per woman in 1965 (United Nations 2022). This is troubling because a fertility rate below 2 leads to population decline, and there is no good reason to think that the replacement fertility rate of 2 children per woman is a hard stopping point. Indeed, fertility rates have been well below replacement in most developed countries for decades, and increasingly the same trends are seen elsewhere. For some representative examples, fertility rates in 2022 were 1.1 in South Korea, 1.3 in Italy, 1.4 in Japan, 1.7 in China and Brazil, and 1.8 in the United States (UN Population Fund 2022).

At present, demographers are fairly confident that population growth will halt by 2100 at a peak population no larger than 11 or 12 billion, and will then begin to decline (Lutz et al. 2014; United Nations 2022). Existing policy measures show few signs of reversing this decline (Geruso and Spears forthcoming) and have often brought undesirable side-effects (Connelly 2010). And if demographic trends are not reversed, the longtermist's estimates of future population size may be dramatic overestimates.

How small could the future human population be? Several demographers project sharp declines in human population out until at least 2300 (Basten et al. 2013; Raftery and Sevcikova 2023; Spears et al. 2023). Recently, Mike Geruso and Dean Spears (forthcoming) extended these projections to predict the total number of future people who are likely to be born if current trends continue. Assuming that fertility rates will converge to a level of 1.66 births per woman (Institute for Health Metrics and Evaluation 2020; Lutz et al. 2014), the number of annual births will soon decline to a low level. This striking projection puts the entire future population of humanity at no more than twenty to thirty billion lives, even ignoring background levels of existential risk. Here is an illustration of how this projection behaves over the next two millennia (Figure 1).

Crucially, this story is not dependent on the precise numbers used. Geruso and Spears

Figure 1: Projected number of annual births, from Geruso and Spears (forthcoming).



extend their model to consider equilibrium fertility rates between 1.0 and 1.8 births per woman. Even on the most optimistic fertility rate of 1.8, only about thirty billion humans have yet to be born. As Geruso and Spears put it, we have dropped from MacAskill's projection of five million future 'stick figures', each representing ten billion future lives, to a projection of two to three stick figures.

Other economists and demographers have joined Geruso and Spears in urging that demographic projections could spell trouble for the value of existential risk mitigation (Alexandrie and Eden forthcoming). If there are not nearly so many future lives yet to be lived, then it is relatively less important to ensure that human extinction is averted and relatively more important to live well today. Crucially, it may be very important to avoid medium-sized population shocks, such as wars and famines, in the coming decades because these shocks may have an effect on the future number of human lives within 1-2 orders of magnitude of the effect of outright extinction (Alexandrie and Eden forthcoming).

Although this is the story told by our best current demographic projections, readers

may raise a number of objections to reliance on this story. For example, they may hold that high-fertility subpopulations will inherit an increasing share of the human population (Kaufmann 2010), eventually reversing the fertility decline. But many demographers place low confidence in this scenario (Gitel-Basten et al. 2014; Geruso and Spears forthcoming; Raftery and Sevcikova 2023). For one thing, fertility rates are falling rapidly even in the highest-fertility groups (Arenberg et al. 2022), so there is no guarantee that any group will retain high fertility rates. More to the point, fertility norms are at best incompletely transmissible across generations (Beaujouan and Solaz 2019), so we need not expect the distant descendants of high-fertility groups to share their enthusiasm for reproduction. Finally, it would take many centuries for high-fertility subgroups to gain a sizable population share (Raftery and Sevcikova 2023), giving ample time for standard processes of normative and societal change to modify reproductive practices.

Alternatively, so-called 'techno-optimists' may hold that technology will soon reduce the cost of child-bearing and increase the ease of reproduction strongly enough to reverse fertility trends. For example, increasing automation of domestic tasks and childcare combined with novel fertility-enhancing technologies may make child-rearing a more attractive option for many couples. Many demographers are skeptical about such stories (Leridon 2004; Kubitza and Gehrke 2018): we have been promised for centuries that technology will reverse fertility declines, but as yet this has not come about. Fertility-enhancing technologies have proliferated alongside technologies such as washing machines and disposable diapers that have dramatically increased the ease of child-rearing. This has happened alongside other supportive developments, including rapid economic growth and a near-halving of working hours in many nations (Lee et al. 2007). If all of these effects have been insufficient to halt declining fertility rates, then we should be somewhat less confident that future technological developments will do the trick.

So far, we have seen that on standard demographic models, falling fertility rates lead to dramatic reductions in the projected size of the future population and corresponding reductions in the value of existential risk mitigation efforts. Although readers are invited to

bring a healthy degree of uncertainty to the enterprise of long-run demographic modeling, we saw that two standard arguments for a reversal of current fertility trends may be less than air-tight. Importantly, even if we do accept that high-fertility subgroups or novel technological developments will lead to fertility growth, this will still leave us very far from scenarios in which population size expands quickly and permanently to carrying capacity. In general, it seems likely that even skeptical reflection on standard demographic models will reduce the expected size of the future population far below carrying capacity.

But surprisingly, even this much is not necessary to motivate the importance of attending to population dynamics. Even combining the Malthusian assumption that population size grows quickly to carrying capacity with highly optimistic assumptions about future technology does not eliminate the role of demographic modeling in decreasing the expected size of the future population.

5.3 A techno-optimist, Malthusian model

Let us begin with some technological assumptions that are as optimistic as they come. Suppose that within a millennium, humanity will gain the ability to settle the stars at a breakneck pace. For the rest of human history, we will then set sail in all directions at one-tenth of the speed of light, colonizing the galaxy in an ever-expanding spherical region around Earth. Add to this a Malthusian story on which resources are the primary constraint on population size, so that a spherically-expanding region will be quickly settled as it grows. These are the starting assumptions of a recent model of the value of existential risk mitigation by Christian Tarsney (2023). Let us first see how this model performs before demographic assumptions are included, and then we will see that even optimistic demographic assumptions tend to dramatically reduce the value of existential risk reduction on the model.

Tarsney compares the value of two stylized interventions, each costing a million dollars. The near-termist intervention, N, provides a sure gain of 10,000 quality-adjusted life years (QALYs), comparable to current estimates of the cost-effectiveness of global health

interventions. The longtermist intervention L increases the probability p that humanity will not go extinct within the next millennium.[14] Tarsney asks under what conditions L has higher expected value than N.

As above, Tarsney assumes that after a millennium, humanity will begin settling the stars in all directions at a constant rate s equal to one-tenth of the speed of light. This will continue until humanity becomes extinct, with constant probability r per year, or the universe becomes unfriendly to human life at some distant time t_f.[15] In the meantime, each century we reap value v_e from settling the Earth and v_s from each star settled. Tarsney estimates v_e at 6 billion QALYs, comparable to the value of life on Earth today. Tarnsey estimates that each settled star will yield 5% of this value, or 300 million QALYs. Letting $n(x)$ return the number of stars in a sphere of radius x around Earth, Tarsney estimates the expected value of L as:[16]

$$E[V(L)] = p \int_{t=0}^{t_f} (v_e + v_s n(st)) e^{-rt} dt.$$

Tarsney finds that the expected value of the longtermist intervention L exceeds that of the short-termist intervention N so long as annual risk of extinction r falls below 0.000135, or approximately a 1.34% risk of extinction per century. Although we saw earlier that many authors take r to be higher than this, it is not altogether implausible that r could be much lower than this, in which case existential risk mitigation looks like a promising idea.

One challenge for this model is that it neglects population dynamics. It assumes that

[14]This is a slight simplification of Tarsney's model, which distinguishes between positive and negative 'exogenous nullifying events' which may erase the impact of L. I read negative nullifying events as extinction-level events, following Tarsney's primary interpretation of the model. I assume positive nullifying events are impossible, an assumption friendly to the longtermist.

[15]Tarsney sets t_f to 10^{14} years, although in this model a lower value of t_f is unlikely to change qualitative model behavior.

[16]Tarsney defines $n(x)$ in two pieces, to reflect the fact that stars are denser in the region closer to Earth:

$$n(x) = \begin{cases} (4\pi/3)x^3 d_g & 0 \leq x \leq 130{,}000 \text{ light-years} \\ (4\pi/3)(x^3 d_s + 130{,}000(d_g - d_s)) & x > 130{,}000 \text{ light-years} \end{cases}$$

where the densities of stars per cubic light-year $d_g = 2.9 * 10^{-9}$ and $d_s = 2.5 * 10^{-5}$ are derived from current cosmological estimates.

humanity continues immediately past each planet it settles and on towards the next. Although readers may assign some credence to such scenarios, demographers typically assign higher credence to stories such as the following (Boyle et al. 2013). Early settlers find resources plentiful and economic growth easy. Absent domestic stressors, they have little reason to set sail for new lands. As settlers gain more resources and grow the economy of a new planet, they reproduce at a reasonable rate allowed by new resources. Eventually, however, the planet becomes crowded and economic opportunities lessen. This makes it increasingly attractive for colonists to set sail for new planets, leading to migration.

Let us assume, optimistically, that a band of human colonists takes a thousand years to turn a planet from an early settlement into a mature colony whose population and economy have reached the point where settling new planets is a desirable solution to overcrowding and stagnant economic growth. Let's take the average distance between stars in the Milky Way, about five light-years (Krauss and Starkman 1999), as a good proxy for how far they will have to travel. This is, in many ways, an overestimate, because estimates of the number of habitable planets are much lower than the number of stars (Kasting et al. 1993; Kopparapu 2013; Petigura et al. 2013), and because as Tarsney notes, stars become significantly less dense outside of the Milky Way, but it cannot hurt to be generous in order to fend off small complaints about modeling assumptions. This gives a rough annual speed of interstellar expansion of $s = 5/1,000$ times the speed of light.

Putting this new value for s into Tarsney's model and leaving all other model elements unchanged, we now find that the expected value of the longtermist intervention L exceeds that of the short-termist intervention N only once annual extinction risk r falls below 0.0000145, a per-century risk of approximately 0.145%. This is almost ten times lower than the previously required level of risk, and significantly harder to achieve. Now the case for existential risk mitigation looks much less strong. To appreciate how much the case has been weakened, note that the level of annual risk $r = 0.000135$ which previously sufficed to make the longtermist intervention more L more valuable now counts the short-

termist intervention N as five hundred times more valuable than L.

The lesson of this discussion is that population dynamics matter. On standard population models, even a constant population of ten billion humans lies out of reach, and there is a good chance that most of the humans who have ever lived have already been born. Standard population models exert very strong downward pressure on the value of existential risk mitigation, for example dropping MacAskill's estimated five million 'stick figures' of ten billion humans each down to an estimated 2-3 stick figures. Some speculative models combine the techno-optimist assumption that technology will make rapid population expansion easy with the Malthusian assumption that humans will exploit technology to expand in proportion to the resources they are able to exploit, rather than enriching a smaller population of humans. Even on these models, we saw, the most ambitious reasonable demographic assumptions exert a substantial downward force on the value of existential risk mitigation and the level of background risk needed to make the case for investment in existential risk mitigation.[17]

6 Learning from the mistakes

In this paper, we considered three mistakes in the moral mathematics of existential risk. We saw that correcting each mistake exerts substantial downward pressure on the value of existential risk mitigation, particularly when the mistakes are corrected together. We also saw how identifying these mistakes shifts the focus of debates about the moral importance of existential risk mitigation by identifying morally relevant factors which are often neglected in debates about existential risk mitigation, but which are strong enough that they should be a key focus of future debates.

The first mistake in the moral mathematics of existential risk is focusing on cumulative risk rather than period risk (Section 3). This is a mistake because the cumulative risk

[17]The value of existential risk mitigation may be further reduced when we note that a falling population may bring an end to economic growth (Jones 2022) and may reduce society's ability to protect against future risks (Geruso and Spears forthcoming).

faced by all future generations is largely out of our control. It is also misleading because seemingly small reductions in cumulative risk often require very large reductions in period risk, not only in our own century but also in all future centuries. For example, we saw that Bostrom's (2003) proposed 'small' reduction in cumulative risk by a millionth of a percent over the course of a billion years actually requires driving per-century risk to roughly one-in-a-million, and annual risk to roughly one-in-a-hundred-million. This is, as we saw, about five orders of magnitude lower than many longtermists take risk to be today.

The second mistake in the moral mathematics of existential risk is ignoring background levels of existential risk (Section 4). This is a mistake because the value of reducing any single risk, such as biosecurity hazards, diminishes substantially if there are other risks which will not be reduced by our actions. We considered a leading estimate of the cost-effectiveness of biosecurity interventions due to Piers Millett and Andrew Snyder-Beattie (2017) and saw how incorporating background risk takes several orders of magnitude off the cost-effectiveness of biosecurity interventions, making leading short-termist interventions comparably effective to biosecurity interventions on many models. We then saw how the first mistake combines with the second: once we focus on risk reductions in our own century, rather than across all centuries at once, we see that leading short-termist interventions are often substantially more cost-effective than biosecurity spending on the Millett and Snyder-Beattie model.

The third mistake in the moral mathematics of existential risk is ignoring population dynamics (Section 5). This mistake confuses the technical question of how many lives a region of space could in principle support with the demographic question of how many lives that region is likely to support, given the ways in which human populations evolve. We saw that on standard population models, representing our best scientific understanding of how the human population is likely to evolve, leading longtermist interventions of the future population are dramatic overestimates: MacAskill's (2022b) five million 'stick figures' of ten billion unborn humans each may well be replaced by only two

or three stick figures. We also considered a highly optimistic model on which technological progress reintroduces a Malthusian regime of rapid interstellar expansion (Tarsney 2022). We saw that even under these optimistic assumptions, introducing plausible constraints on population dynamics substantially reduces the value of existential risk mitigation.

It is important to pay attention to these mistakes in moral mathematics not only because they reduce the value of existential risk mitigation, but also because they help us to see debates about existential risk mitigation in a new light. Here are four concrete sets of normative lessons from this discussion that I hope will structure future debates.

6.1 Cumulative risk and intergenerational coordination

Reaping large benefits from existential risk mitigation means driving down the cumulative risk faced by humanity over long stretches of human civilization. This introduces an intergenerational coordination problem: how can each generation ensure that future generations will do their duty and drive down risk? Even a few irresponsible generations can dramatically influence cumulative risk: for example, if we mean to drive per-century risk down to one-in-a-million, and we think that the current generation is running at least a one-in-ten risk of existential catastrophe, then we will need to convince at least a hundred thousand future generations to drive risk down to zero in order to offset our misdeeds, or convince even more generations than this to bear a smaller sacrifice.

Solving this intergenerational coordination problem is hard for four reasons. First, as we saw, the levels of risk we aim to achieve are quite low and may require substantial sacrifice, if they are even feasible at all. Second, because every generation bears only a tiny fraction of the cost of existential catastrophe, solving the intergenerational coordination problem requires instilling an unusual degree of concern for externalities borne by future generations. Third, this level of concern is especially hard to instill given that people are often impatient (Frederick et al. 2002) and exhibit only limited degrees of altruism (Fehr and Fischbacher 2003). Finally, enforcement is difficult because each generation has limited means to monitor and punish the selfishness of future generations, so there is

pressure for each generation to defect from a collectively optimal solution.

Recasting existential risk mitigation as an intergenerational coordination problem helps us to see the importance of two questions. First, is it feasible to achieve a high degree of coordination around low levels of existential risk? And second, are there ethical limitations on the means that can be used to bind future generations to comply with desired levels of risk, especially given that they had no direct voice in present agreements? Each question opens productive avenues for future debate.

6.2 Background risk, dialectical flips, and the Time of Perils

We saw in Section 4 that if background levels of existential risk are raised, then the value of mitigating any particular risk, such as biosecurity threats, is dramatically reduced. This happens because a world in which biosecurity risks are reduced would nonetheless be a risky world, and hence a world significantly more vulnerable to future catastrophe. David Thorstad (forthcoming) generalizes the point to show that in most plausible models, raising the overall level of existential risk tends to lower, rather than raise, the value of reducing all risks at once. Thorstad shows that pessimistic assumptions about the background level of existential risk often slice many orders of magnitude off the value of a fixed relative or absolute risk reduction, and in particular that the high levels of background risk many longtermists take humanity to face today make it very hard to assign astronomical value to existential risk mitigation efforts.

One surprising implication of this discussion that Thorstad emphasizes is a dialectical flip in debates about existential risk. Naively, we might think that a good way to convince philosophers of the importance of existential risk mitigation is to argue that levels of risk are very high, and a good way to resist the importance of existential risk mitigation is to argue that levels of risk are very low. But counterintuitively, unless more is said, precisely the opposite is true: higher levels of background existential risk tend to dramatically lower the importance of existential risk mitigation, and lower levels of background existential

risk tend to raise the importance of existential risk mitigation.[18] This means that, oddly, many parties to debates about existential risk may have been arguing on behalf of their opponents.

A second important implication that Thorstad recognizes is that if we begin from high levels of background risk, the only clearly viable strategy for arguing that existential risk mitigation is astronomically important is to adopt the Time of Perils Hypothesis on which risk will soon fall by many orders of magnitude, and stay low for the rest of human history (Ord 2020). The Time of Perils Hypothesis is a striking claim, and it is not always appreciated how much of the case for existential risk mitigation may depend on it. However, both Thorstad and others have offered reasons to doubt the Time of Perils Hypothesis (MacAskill 2022a), and these reasons would also cut against the value of existential risk mitigation.[19]

6.3 Population dynamics, demographic interventions, and digital minds

Section 5 reminds us of the importance of shifting from the technical question of how many lives a given region could support to the demographic question of how many lives the region is likely to support given the dynamics of human populations. This discussion raises two important implications. First, interventions aimed at increasing the size of the future human population can now become at least as important as efforts to mitigate existential risk (Alexandrie and Eden forthcoming), although the precise details are sensitive to considerations of population axiology.[20] For example, if we think that there are likely to be at most 20-30 billion future humans absent intervention, but that in principle there could be many more humans than this, then it may be orders of magnitude more important to do what we can to expand the size of the future human population

[18]This discussion could be complicated by competing factors. For example, perhaps lowering the background level of existential risk reduces the tractability of existential risk mitigation.

[19]Note that even under more optimistic assumptions about background risk, denying the Time of Perils Hypothesis still tends to reduce the value of existential risk mitigation.

[20]For example, on the neutrality view many efforts to increase future population size will have little value.

than it is to prevent the otherwise modest future population from suffering catastrophe.

Second, a focus on population dynamics increases the importance of scenarios involving digital minds. Many authors have noted one advantage of focusing on the development of digital minds: we could, in principle, support a larger population of digital minds than human minds (Bostrom 2013; Greaves and MacAskill 2021). However, a second advantage of digital minds may be at least as important: digital populations programmed to value expansion may be far more likely than humans to expand to a meaningful proportion of their maximum possible size. This means that views which take seriously the value of creating large populations of future digital minds may want to focus on ensuring the development and safety of digital populations, even at the expense of the safety and well-being of future human populations. I leave it to the reader to decide what to make of this implication.

6.4 Cluelessness and model uncertainty

One of the best-known challenges to longtermism is the *cluelessness problem*: it is very hard to know how our acts will affect the future (Greaves 2016; Lenman 2000; Mogensen 2021). For example, perhaps by driving down the street I will run over the erstwhile founder of a world government. But perhaps instead I will run over her chief opponent. Or perhaps the government was going to be a bad one anyways.

Many longtermists have held out hope that existential risk mitigation could solve the cluelessness problem.[21] Because existential risks can be identified and confronted today, we may not be clueless about the nature of current risks or the acts that could reduce them. As long as we are also reasonably confident that the future will be worth preserving, we may then be reasonably confident that acts aimed at preventing current risks are good acts, even if we are unsure about the precise nature of the future they will lead to.

The discussion in this paper suggests that cluelessness may have its revenge. We saw that many recent discussions of the value of existential risk mitigation take insufficient

[21] For discussion see MacAskill (2022b), Thorstad (forthcoming) and Rini (2022).

account of three morally relevant factors that dramatically change the value of existential risk mitigation when properly accounted for. One way to interpret this discussion would be as suggesting that existential risk mitigation is less important than we took it to be. But a different interpretation is suggested by the fact that if we missed at least three morally relevant factors in the past, there are likely to be other relevant factors that have not been considered. Due to the complexity of valuing existential risk mitigation, it is very hard to construct models which incorporate all, or even a small subset of the morally relevant considerations.

If that is right, then there may be no escape from cluelessness through attention to existential risk mitigation. Even under the controversial assumptions that humanity faces high levels of risk today, that we are not clueless about what can be done to mitigate current risks, and that the future will be worth living, we may nonetheless retain substantial cluelessness in the form of model uncertainty about whether the most relevant moral considerations have been represented, and represented well, in discussions of the value of existential risk mitigation. How high should our level of model uncertainty be? I leave this question to readers to judge. But if future discussions reveal a number of other morally salient considerations which have been excluded from discussions of existential risk mitigation, then we should be relatively less confident that existential risk mitigation provides an escape from cluelessness.

Appendix: MSB model with background risk

In the MSB model with background risk r, the expected number of future lives is:

$$E[L] = 10^{12} \sum_{i=1}^{10,000} (1-r)^i.$$

Decompose per-century risk as $r = b + n$ where b represents biological risk and n represents nonbiological risk. Let X be an action which, following MSB, provides a 1% relative reduction in biological risk, shifting total risk to $r_X = 0.99b + n$. Now, the expected number

of future life-years is:

$$E[L|X] = 10^{12} \sum_{i=1}^{10,000} (1 - r_X)^i.$$

This is a geometric series, hence the number of lives added by X is:

$$\begin{aligned} E[L|X] - E[L] =& 10^{12}\left(\frac{1 - (r - 0.01b)}{r - 0.01b} - \frac{1 - r}{r}\right) \\ =& 10^{12} \frac{0.01b}{r(r - 0.01b)}. \end{aligned}$$

References

Alexandrie, Gustav and Eden, Maya. forthcoming. "Is extinction risk mitigation uniquely cost-effective? Not in standard population models." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Essays on longtermism*, forthcoming. Oxford University Press.

Arenberg, Samuel, Kuruc, Kevin, Franz, Nathan, Vyas, Sangita, Lawson, Nicholas, LoPalo, Melissa, Budolfson, Mark, Geruso, Mike, and Spears, Dean. 2022. "Intergenerational transmission is not sufficient for positive long-term population growth." *Demography* 59:2003–12.

Barro, Robert and Becker, Gary. 1989. "Fertility choice in a model of economic growth." *Econometrica* 57:481–501.

Basten, Stuart, Lutz, Wolfgang, and Scherbov, Sergei. 2013. "Very long range global population scenarios to 2300 and the implications of sustained low fertility." *Demographic Research* 28:1145–66.

Beaujouan, Eva and Solaz, Anne. 2019. "Is the family size of parents and children still related? Revisiting the cross-generational relationship over the last century." *Demography* 56:595–619.

Bostrom, Nick. 2003. "Astronomical waste." *Utilitas* 15:308–14.

—. 2013. "Existential risk prevention as a global priority." *Global Policy* 4:15–31.

Bostrom, Nick and Ćirković, Milan (eds.). 2011. *Global catastrophic risks*. Oxford University Press.

Boyle, Paul, Halfacree, Keith, and Robinson, Vaughan. 2013. *Exploring contemporary migration*. Routledge.

Buchak, Lara. forthcoming. "How should risk and ambiguity affect our charitable giving?" *Utilitas* forthcoming.

Connelly, Matthew. 2010. *Fatal misconception: The struggle to control world population*. Harvard University Press.

Cordelli, Chiara. 2016. "Reparative justice and the moral limits of discretionary philanthropy." In Rob Reich, Chiara Cordelli, and Lucy Bernholz (eds.), *Philanthropy in democratic societies: History, institutions, values*, 244–66. Chicago University Press.

Curran, Emma. 2022. "Longtermism, aggregation, and catastrophic risk." GPI Working Paper 18-2022, https://globalprioritiesinstitute.org/longtermism-aggregation-and-catastrophic-risk-emma-j-curran/.

Fehr, Ernst and Fischbacher, Urs. 2003. "The nature of human altruism." *Nature* 425:785–91.

Frederick, Shane, Lowenstein, George, and O'Donoghue, Ted. 2002. "Time discounting and time preference: A critical review." *Journal of Economic Literature* 40:351–401.

Frick, Johann. 2017. "On the survival of humanity." *Canadian Journal of Philosophy* 47:344–67.

Geruso, Mike and Spears, Dean. forthcoming. "With a whimper: Depopulation and longtermism." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Essays on longtermism*. Oxford University Press.

Gitel-Basten, Stuart, Tomá, Sobotka, and Krytof, Zeman. 2014. "Future fertility in low fertility countries." In Wolfgang Lutz, William Butz, and Samir KC (eds.), *World population and human capital in the twenty-first century*, 39–146. Oxford University Press.

GiveWell. 2021. "GiveWell's Cost-Effectiveness Analyses." https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models.

Greaves, Hilary. 2016. "Cluelessness." *Proceedings of the Aristotelian Society* 116:311–39.

Greaves, Hilary and MacAskill, William. 2021. "The case for strong longtermism." Global Priorities Institute Working Paper 5-2021, https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/.

Heikkinen, Karri. 2022. "Strong longtermism and the challenge from anti-aggregative moral views." Global Priorities Institute Working Paper 5-2022, https://globalprioritiesinstitute.org/karri-heikkinen-strong-longtermism-and-the-challenge-from-anti-aggregative-moral-views/.

Institute for Health Metrics and Evaluation. 2020. "Population forecasting." https://vizhub.healthdata.org/population-forecast/.

Jensen, Robert. 2012. "Do labor market opportunities affect young women's work and family decisions? Experimental evidence from India." *Quarterly Journal of Economics* 127:753–92.

Jensen, Robert and Oster, Emily. 2009. "The power of TV: Cable television and women's status in India." *Quarterly Journal of Economics* 124:1057–94.

Jones, Charles. 2022. "The end of economic growth? Unintended consequences of a declining population." *American Economic Review* 112:3489–527.

Kasting, James, Whitmire, Daniel, and Reynolds, Ray. 1993. "Habitable zones around main sequence stars." *Icarus* 101:108–28.

Kaufmann, Eric. 2010. *Shall the religious inherit the earth? Demography and politics in the twenty-first century*. Profile Books.

Kearney, Melissa and Levine, Peilip. 2015. "Media influences on social outcomes: The impact of MTV's 16 and Pregnant on teen childbearing." *American Economic Review* 105:3597–3632.

Kopparapu, Ravi Kumar. 2013. "A revised estimate of the occurrence rate of terrestrial planets in the habitable zones around Kepler M-Dwarfs." *Astrophysics Journal Letters* 767:L8.

Krauss, Lawrence and Starkman, Glenn. 1999. *Teaching about cosmology*. AIP Publishing.

Kubitza, Christoph and Gehrke, Esther. 2018. "Why does a labor-saving technology decrease fertility rates? Evidence from the oil palm boom in Indonesia." EFForTS Discussion Paper Series No. 22, https://www.econstor.eu/handle/10419/179250.

Lee, Sangheon, McCann, Deirdre, and Messenger, Jon C. 2007. *Working time around the world: Trends in working hours, laws and policies in a global comparative perspective*. Routledge.

Lenman, James. 2000. "Consequentialism and cluelessness." *Philosophy and Public Affairs* 29:342–70.

Leridon, Henri. 2004. "Can assisted reproduction technology compensate for the natural decline in fertility with age? A model assessment." *Human Reproduction* 19:1548–53.

Lestahege, Ron. 1992. "Beyond economic reductionism: The transformation of the reproductive regimes in France and Belgium in the 18th and 19th Centuries." In Calvin Goldscheider (ed.), *Fertility transitions, family structure, and population policy*, 1–44. Westview.

Lloyd, Harry. 2021. "Time discounting, consistency and special obligations: a defence of Robust Temporalism." Global Priorities Institute Work-

ing Paper 11-2021, https://globalprioritiesinstitute.org/time-discounting-consistency-and-special-obligations-a-defence-of-robust-temporalism-harry-r-lloyd-yale-university/.

Lutz, Wolfgang, Butz, William P., and KC, Samir (eds.). 2014. *World population and human capital in the twenty-first century*. Oxford University Press.

MacAskill, William. 2022a. "Are the living at the hinge of history?" In Jeff McMahan, Tim Campbell, Jame Goodrich, and Ketan Ramakrishnan (eds.), *Ethics and existence: The legacy of Derek Parfit*, 331–57. Oxford University Press.

—. 2022b. *What we owe the future*. Basic books.

Malthus, Thomas. 1798. *An essay on the principle of population as it affects the future improvement of society, with remarks on the speculations of Mr. Goodwin, M. Condorcet, and other writers*. J. Johnson in St Paul's Church-yard.

Millett, Piers and Snyder-Beattie, Andrew. 2017. "Existential risk and cost-effective biosecurity." *Health Security* 15:373–84.

Mogensen, Andreas. 2021. "Maximal cluelessness." *Philosophical Quarterly* 71:141–62.

—. 2022. "The only ethical argument for positive δ? Partiality and pure time preference." *Philosophical Studies* 179:2731–50.

Monton, Bradley. 2019. "How to avoid maximizing expected utility." *Philosophers' Imprint* 19:1–25.

Naverson, Jan. 1973. "Moral problems of population." *The Monist* 57:62–86.

Newberry, Toby. 2021. "How many lives does the future hold?" Technical report, Global Priorities Institute, https://globalprioritiesinstitute.org/wp-content/uploads/Toby-Newberry_How-many-lives-does-the-future-hold.pdf.

Ord, Toby. 2020. *The precipice*. Bloomsbury.

Petigura, Erik, Howard, Andrew, and Marcy, Geoffrey. 2013. "Prevalence of Earth-size planets orbiting Sun-like stars." *Proceedings of the National Academy of Sciences* 110:19273–8.

Pettigrew, Richard. 2022. "Effective altruism, risk, and human extinction." Global Priorities Institute Working Paper 2-2022, https://globalprioritiesinstitute.org/effective-altruism-risk-and-human-extinction-richard-pettigrew-university-of-bristol/.

Raftery, Adrian and Sevcikova, Hana. 2023. "Probabilistic population forecasting: Short to very long-term." *International Journal of Forecasting* 39:73–97.

Rees, Martin. 2003. *Our final hour*. Basic books.

Rini, Regina. 2022. "An effective altruist? A philosopher's guide to the long-term threats to humanity." *Times Literary Supplement* .

Sandberg, Anders and Bostrom, Nick. 2008. "Global catastrophic risks survey." Technical Report 2008-1, Future of Humanity Institute, https://www.global-catastrophic-risks.com/docs/2008-1.pdf.

Sen, Amartya. 1999. *Development as freedom*. Oxford University Press.

Smith, Nicholas. 2014. "Is evaluative compositionality a requirement of rationality?" *Mind* 123:457–502.

Spears, Dean, Vyas, Sangita, Weston, Gage, and Geruso, Mike. 2023. "Long-term population projections: Scenarios of low or rebounding fertility." Population Wellbeing Initiative working paper.

Tarsney, Christian. 2022. "The epistemic challenge to longtermism." Global Priorities Institute Working Paper 3-2022, https://globalprioritiesinstitute.org/christian-tarsney-the-epistemic-challenge-to-longtermism/.

Tarsney, Christian and Wilkinson, Hayden. 2023. "Longtermism in an infinite world." Global Priorities Institute Working Paper 4-2023,

https://globalprioritiesinstitute.org/longtermism-in-an-infinite-world-christian-j-tarsney-and-hayden-wilkinson/.

Thorstad, David. forthcoming. "High risk, low reward: A challenge to the astronomical value of existential risk mitigation." *Philosophy and Public Affairs* forthcoming.

UN Population Fund. 2022. "World Population Dashboard." https://www.unfpa.org/data/world-population-dashboard.

United Nations. 2022. *World population prospects 2022*. United Nations Department of Economics and Social Affairs.

Unruh, Charlotte. forthcoming. "Constraining longtermism? A non-consequentialist objection to longtermism." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Essays on longtermism*. Oxford University Press.