Is Existential Risk Mitigation Uniquely Cost-Effective? Not in Standard Population Models

Gustav Alexandrie (Global Priorities Institute, University of Oxford) and Maya Eden (Brandeis University)

Global Priorities Institute | July 2023

GPI Working Paper No . 5-2023

Is Existential Risk Mitigation Uniquely Cost-Effective? Not in Standard Population Models

Gustav Alexandrie* and Maya Eden*†

July 4, 2023

Abstract

What socially beneficial causes should philanthropists prioritize if they give equal ethical weight to the welfare of current and future generations? Many have argued that, because human extinction would result in a permanent loss of all future generations, extinction risk mitigation should be the top priority given this impartial stance. Using standard models of population dynamics, we challenge this conclusion. We first introduce a theoretical framework for quantifying undiscounted cost-effectiveness over the long term. We then show that standard population models imply that there are interventions other than extinction risk mitigation that can produce persistent social benefits. In fact, these social benefits are large enough to render the associated interventions at least as cost-effective as extinction risk mitigation.

Keywords: human extinction, existential risk, fertility, population growth, global catastrophic risks

JEL Classification: D61, D64, Q54, Q56

---

*Global Priorities Institute, University of Oxford.
†Economics Department, Brandeis University.

We want to thank Loren Fryxell, Luis Mota Freitas, Michael Geruso, Rossa O'Keefe-O'Donovan, Caroline Falkman Olsson, Kevin Kuruc, Charlotte Siegmann, Dean Spears, Luca Stroppa, Benjamin Tereick, and Phil Trammell for helpful feedback on previous versions of this paper.

1 Introduction

In the coming century, humanity may face global catastrophic risks stemming from climate change, nuclear war, pandemics, and emerging technologies such as artificial intelligence (Häggström 2016; Ord 2020). Many interventions for reducing these risks are likely to be cost-effective by the light of standard cost-benefit analysis (Posner 2004; Shulman and Thornley, forthcoming). However, it has often been argued that, under a zero rate of pure time preference[1], special priority should be given to the subset of these interventions that most effectively reduce the risk of human extinction. In a widely cited passage, Parfit (1984: 453) introduces this line of argument by comparing three possible outcomes:

(i) No catastrophe occurs.

(ii) A catastrophe kills 99% of the existing world population.

(iii) A catastrophe kills 100%.

Insofar as human life is valuable, (i) is clearly socially better than (ii), which in turn is better than (iii).[2] But which of these two differences is greater in terms of welfare loss? Counterintuitively, Parfit and many others have argued that, although the welfare difference between (i) and (ii) is greater if only the current generation is considered, the welfare difference between (ii) and (iii) is greater if all generations are considered equally. The motivation for this is that, while any global catastrophe would lead to an immense welfare loss for the current generation, human extinction

---

[1] Adopting a zero rate of pure time preference amounts to not discounting future welfare. Note that this is fully consistent with discounting future *consumption* based on the expected rate of economic growth and the diminishing marginal utility of consumption. Surveys of the arguments for and against adopting a zero rate of pure time preference are found in Dasgupta (2008), Greaves (2017), and Groom et al. (2022).

[2] For views to the contrary, see e.g., Benatar (2013) and Pettigrew (2022).

would *additionally* lead to an even greater welfare loss by irreversibly preventing all subsequent generations from coming into existence.[3] Therefore, in Parfit's view, "[w]hat matters *most* is how we respond to various risks to the survival of humanity" (Parfit 2017: 436, emphasis added). This line of thought has been invoked in cost-effectiveness analyses of interventions that reduce the risk of human extinction posed by asteroids (Matheny 2007), climate change (Ng 2016), and pandemics (Millett and Snyder-Beattie 2017). We refer to it as *the long-run argument for prioritizing extinction risk mitigation* (or simply, 'the long-run argument').

An important assumption underlying the long-run argument for prioritizing extinction risk mitigation over other types of risk mitigation is that the welfare effects of human extinction would be permanent, whereas the welfare effects of a non-extinction catastrophes would not. More precisely, the argument assumes that, if a non-extinction catastrophe were to occur, humanity would have a good chance of eventually recovering. However, the likelihood of such recovery depends on people's fertility decisions, which in turn depend on economic and social factors. Understanding these factors is necessary for determining whether extinction would indeed be uniquely consequential in the long run, or whether some non-extinction catastrophes would have comparably persistent effects on long-run population and welfare levels (cf. Ord ms-a).

In this paper, we explore how shocks to the size of the current population might affect long-run population levels, and what this implies for philanthropic priority setting. We start by introducing a theoretical framework for quantifying the undiscounted cost-effectiveness of risk reduction efforts. A heuristic implied by this frame-

---

[3] Parfit writes, "Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second" (Parfit 1984: 453-454).

work is that the undiscounted cost-effectiveness of reducing the risk of a negative population shock is proportional to the ratio of lives lost in the long run (in percentage terms) to lives lost in the short run (in percentage terms).

In the remainder of the paper, we assess the implications of various population models for the relationship between decreases in current population levels and long-run population levels. First, we discuss shocks that reduce the current population level, but that leave all other factors of production unaltered. We show that, for such shocks, the assumption that population levels eventually recover after any non-extinction shock is implied by the Malthusian model of fertility (Malthus 1798). Importantly, however, this assumption is *not* implied by models that take fertility choices to be primarily determined by social norms. Nor is it implied by the Barro-Becker model (Becker and Barro 1988; Barro and Becker 1989), which is the workhorse model for studying the economic determinants of modern fertility dynamics. Indeed, in our calibration of the Barro-Becker model, non-extinction shocks to current population levels can result in permanent drops in long-run population levels that are disproportionately larger than the size of the initial shock.

We then proceed by analyzing events that reduce both population size and other factors of production proportionally by the same amount. Given constant returns to scale technology, such events leave economic determinants of fertility choices unaffected and therefore result in a permanent, proportional reduction in the size of the global population. Interventions that save lives and increase the capital stock in equal proportion therefore have permanent effects in standard economic fertility models. Our undiscounted cost-effectiveness framework suggests such interventions could be as cost-effective as extinction risk mitigation. Moreover, a back-of-the-envelope calculation suggests that these interventions may be even more cost-effective than extinction risk mitigation provided that the determinants of population levels remain sufficiently stable far enough into the future. While these cost-effectiveness estimates should be interpreted with considerable caution, they nonetheless suggest that interventions other than extinction risk mitigation could have significant impact on long-run social welfare.

2 Outlining the argument

Let us for simplicity restrict our attention to the subset of interventions whose social impact primarily stems from their effects on the number of people or life years that are brought into existence (as opposed to their effects on people's quality of life at any given time). We refer to this as the set of *population-affecting* interventions. Note that population-affecting interventions include both lifesaving interventions (e.g., antimalarial bednet distribution or extinction risk mitigation) and non-lifesaving interventions (e.g., changing fertility norms or sustainably increasing the supply of natural resources) that may affect long-run future population levels.

A stylized version of the long-run argument for prioritizing extinction risk mitigation over other population-affecting interventions can be stated as follows:

(P1) The social value of a population-affecting intervention is approximately proportional to how much it increases expected long-run population levels.

(P2) Additional philanthropic spending on extinction risk mitigation increases expected long-run population levels more than additional philanthropic spending on any other population-affecting intervention.

(C) Therefore, additional philanthropic spending on extinction risk mitigation is more socially valuable than additional philanthropic spending on any other population-affecting intervention.

The first premise, (P1), can be supported by the following two claims:

> **Generalized totalism.** *Social value increases linearly in the number of good lives*[4] *that are brought into existence (by the same amount regardless of when they are brought into existence).*

> **Astronomical stakes.** *In expectation, the vast majority of all current and future lives are going to be lives that are lived in the far future, and these lives are in expectation going to be good.*

The reasoning is simple: if the social value increases linearly in the number of good lives, but the set of lives in the far future is vastly larger in expectation than the set of current lives, then the social value of any population-affecting intervention must be largely determined by its effects on long-run population levels. Although there are strong arguments in favor of *Generalized totalism* and *Astronomical stakes*, there are also important counterarguments.[5] However, for the remainder of this paper, we grant that (P1) holds.

The second premise, (P2), can be supported by the following two empirical hypotheses:

---

[4] 'Good lives' here simply refers to lives that contribute positively to social welfare.

[5] Generalized totalism is implied by additively-separable social welfare criteria such as total utilitarianism and total prioritarianism (see e.g., Blackorby et al. 1995; Spears and Zuber 2021, for arguments in favor of these views). Moreover, as Tarsney and Thomas (2020) show, even non-additive axiologies, such as average utilitarianism, rank-discounted utilitarianism and variable value views, converge in practice to the recommendations of additive axiologies if there is a large enough 'background population' that is unaffected by choices. Arguments for Astronomical stakes are discussed in Bostrom (2003; 2013), Beckstead (2013), Ord (2020), and MacAskill (2022), whereas arguments against are discussed in Thorstad (2022), which partly extends models developed by Adamczewski (ms) and Ord (2020).

> **Recovery.** *Any change in population that does not result in human extinction would only have a transitory effect and would therefore not increase long-run population levels.*

> **Priority of saving lives.** *The population-affecting intervention that most cost-effectively increases long-run population levels is a lifesaving intervention.*

The idea is again simple: if population levels always recover after non-extinction changes and if the most cost-effective way of increasing long-run population levels is a lifesaving intervention, then extinction risk mitigation must be the most cost-effective way of increasing long-run population levels. This establishes (P2). In the rest of this paper, we analyze whether *Recovery* and *Priority of saving lives* are consistent with standard population models.

We consider three different population models. The first model is a social determinants model of fertility choices. According to this model, families target an ideal family size that is determined by social factors, primarily related to desirable family dynamics. We argue that, in this model, *Recovery* is unlikely to hold.

Another model that we consider is the Barro-Becker model (Becker and Barro 1988; Barro and Becker 1989), which emphasizes the role of economic factors in fertility choices. In this model, changes in population affect the macroeconomic conditions in ways that may ultimately affect fertility rates. As deaths or transitory changes in fertility rates may have permanent effects on population levels in this model, *Recovery* does not hold.

A third model that we consider is the Malthusian model (Malthus 1798). According to the Malthusian model, population levels are constrained by the availability of natural resources. Of the models that we consider, this is the only one that unequivocally supports *Recovery*, which makes it the most likely candidate for supporting the long-run argument for prioritizing extinction risk mitigation. However, we show that this model does not necessarily support *Priority of saving lives*: in the Malthusian model, interventions that permanently increase the supply of natural resources can permanently increase steady state population levels.

Table 1: Illustration of which models imply *Recovery* and/or *Priority of saving lives*.

Model	Does the model imply that *Recovery* holds?	Does the model imply that *Priority of saving lives* holds?
Social determinants model	No	
Barro-Becker model	No	
Malthusian model	Yes	No

3 Undiscounted cost-effectiveness

3.1 A framework for quantifying undiscounted cost-effectiveness

In this section, we introduce a framework for quantifying the cost-effectiveness of different interventions from the perspective of a longtermist decision-maker that gives equal ethical weight to each generation. The key assumption is that policymakers behave myopically, which is suboptimal from the longtermist's perspective. As a result, the undiscounted cost-effectiveness of an intervention is related to the ratio of its long-term benefits (in percentage terms) and its short-term benefits (in percentage terms).

The policymakers' problem. Suppose that the world's policymakers maximize the expected value of random variable U. In the baseline scenario, the value of U is some (good) value, U_0. However, there are n other possible bad events that could occur. Event i occurs with probability p_i and results in value $U_i < U_0$ for the policymakers. The expected value of U is therefore given by

$$\left(1 - \sum_{i=1}^{n} p_i\right) U_0 + \sum_{i=1}^{n} p_i U_i.$$

The probability of each event i is endogenous, as it depends on the resources that the policymakers devote to averting it. This implies the existence of a function, C_i, such that

$$p_i = C_i(m_i)$$

where m_i is the amount of resources devoted to averting event i. We assume that C_i is twice differentiable, strictly decreasing ($C' < 0$) and strictly convex ($C'' > 0$). This reflects that the marginal reduction in the probability of i from an additional unit of resources devoted to averting i is diminishing, as the best opportunities for risk mitigation are successively exhausted.

The policymakers' optimization problem is thus given by

$$\max_{\{m_i\}_i^n} \left(1 - \sum_{i=1}^{n} C_i(m_i)\right) U_0 + \sum_{i=1}^{n} C_i(m_i) U_i$$

$$\text{s.t.} \ \sum_{i=1}^{n} m_i = m$$

where m is an exogenously given amount of resources that the policymakers devote to averting bad events.

Assuming an interior solution in which some of the policymakers' resources are devoted to the mitigation of each risk, the first-order conditions of this optimization problem imply the existence of some $\lambda > 0$ such that

$$C_i'(m_i^*)(U_i - U_0) = \lambda \text{ for all events } i, \tag{1}$$

where m_i^* is the amount of spending to reduce the risk of event i that is optimal from the policymakers' perspective. Equation (1) states that, when policymakers allocate their risk mitigation spending optimally, the policymakers' marginal benefit of reducing the risk of event i, represented by the LHS of (1), is the same for all events i. The economic intuition behind this is that optimizing policymakers always prioritize spending on those events for which risk mitigation provides the highest marginal benefit, which drives down the marginal benefit of further spending until the marginal benefits of all risk mitigation efforts are equalized.

The longtermist's problem. Consider a longtermist who cares more about future generations than the policymakers do. Rather than maximizing the expected value of U, the longtermist wants to maximize the expected value of some W, which is given by

$$\left(1 - \sum_{i=1}^{n} p_i\right) W_0 + \sum_{i=1}^{n} p_i W_i.$$

Assuming that the longtermist only has a small amount of resources, their marginal benefit of reducing the risk of event i is given by $C_i'(m_i^*)(W_i - W_0)$. Since the equilibrium condition (1) can be restated as $C_i'(m_i^*) = \lambda/(U_i - U_0)$, one can substitute for $C_i'(m_i^*)$ to get the following expression for the longtermist's marginal benefit of reducing event i:

$$C_i'(m_i^*)(W_i - W_0) = \lambda \frac{W_i - W_0}{U_i - U_0}.$$

The longtermist's marginal benefit from averting event i is therefore proportional to the ratio $(W_i - W_0)/(U_i - U_0)$.[6] This ratio, which we refer to as the *long-term value ratio*, increases proportionally in the degree to which mitigating the risk of an event i is cost-effective from the longtermist's perspective. Henceforth, we will use 'cost-effectiveness' to refer to 'cost-effectiveness from the longtermist's perspective'.

In what follows, we interpret U as the expected number of current lives relative to the baseline U_0. So, for example, $U_i = 0.75$ captures an event that reduces the current population by 25% relative to the baseline U_0. Similarly, we interpret W as the expected total number of current and future lives relative to the baseline W_0.[7] The statement $W_i = 0.75$ thus captures an event that reduces the sum of current and future population levels by 25% relative to the baseline W_0. Note that this interpretation implies the normalization that $W_0 = U_0 = 1$ and $W_j = U_j = 0$ for any near-term extinction event j.[8]

Insofar as humanity is expected to last for a long time, the current population constitutes only a relatively small fraction of the lives that the longtermist cares about. The long-term value ratio of spending to reduce the risk of an event can thus be heuristically interpreted as the ratio of the lives lost in the long run (in percentage terms) to the lives lost in the short run (in percentage terms) if the event were to occur.

---

[6] An important complication that we ignore in this paper is that the policymakers may take (their expectations of) the longtermist's funding decisions into account when making their own funding decisions (see Trammell 2021 for an analysis of public good provision when funders have heterogeneous time preferences). These concerns may be less pressing if the longtermist's actions are instead conceived of as advocacy work to convince policymakers to reallocate their funds.

[7] This simplification seems reasonable under *Generalized totalism* (see section 2 and footnote 5 for more details).

[8] Since U and W are unique up to affine transformations, one can add constants to each utility function to ensure that $U_j = W_j = 0$ for any extinction event j, and subsequently scale the utility function by some positive constant to achieve $U_0 = W_0 = 1$.

3.2 The long-run argument for extinction risk mitigation

The framework presented in the previous subsection can be used to formalize a stylized version of the long-run argument for prioritizing extinction risk reduction. The first thing to note is that the normalization ensuring that $W_0 = U_0 = 1$ and $W_j = U_j = 0$ (for any extinction event j) implies that the long-term value ratio of reducing the risk of near-term extinction is normalized to one, that is, $(W_i - W_0)/(U_i - U_0) = 1$.

As noted in section 2, an important assumption underlying the long-run argument for prioritizing extinction risk mitigation is that of *Recovery*. This is the assumption that, as long as humanity does not go extinct, long-run welfare and population levels would eventually recover after a shock. Under this assumption, for any non-extinction event i, the short-run welfare loss from i would be *proportionally* worse than the corresponding long-run welfare loss, that is:

$$\frac{U_i - U_0}{U_0} < \frac{W_i - W_0}{W_0} \tag{2}$$

for all non-extinction shocks i. Since $(U_i - U_0) < 0$ and $U_0 = W_0 = 1$, inequality (2) can be rearranged to say that $(W_i - W_0)/(U_i - U_0) < 1$ for all non-extinction shocks i. The long-term value ratio for efforts that reduce non-extinction risks is therefore strictly less than 1 under the recovery assumption.[9] Thus, given the recovery assumption, it is more cost-effective to reduce the risk of human extinction than to reduce other risks. The next section explores whether the recovery assumption holds in standard population models.

---

[9] How *much* lower than 1 is the long-term value ratio of reducing non-extinction risks given the recovery assumption? The answer depends on the rate of population recovery and the number of generations that come into existence after the recovery. As the fraction of generations that come into existence after the recovery tends to 1, the long-term value ratio of reducing non-extinction risks goes towards 0.

4 Shocks to population levels

4.1 Long-run effects in three population models

The number of people who will exist in the future depends on the fertility decisions of their predecessors.[10] These decisions are, in turn, the result of economic and social factors. The economic factors include individual wealth and factor prices. Wealth and income determine the amount of resources that people can devote to child-rearing, as well as the standard of living that they can afford each of their children. Wages capture a component of the costs of raising children, which often requires a reduction in work hours.

To understand how population shocks change fertility decisions, it is useful to understand how they affect the economic environment. If 50% of the population suddenly disappeared, there would be roughly 50% fewer workers. This would mean that each surviving worker could produce output using twice as much capital. For example, each farmer would have twice as much land; each factory worker would have twice as many machines, etc. As a result, standard economic theory predicts that there would be increases in the marginal product of labor and (therefore also) in wage rates. Average wealth would also increase, as the ownership of the economy's capital stock would be distributed among fewer people. These predictions appear to be broadly in line with the historical evidence indicating that the Black Death – the proportionally largest population shock in European history – led to a rise in living standards for ordinary people in late medieval Europe (Jedwab et al. 2022).[11]

---

[10] Population size depends on both fertility and mortality. Our paper is largely framed in terms of fertility, but many of the insights generalize to settings where changes in mortality are taken into account. Readers that are interested in a more focused discussion about mortality (in particular, delaying senescence) are referred to Kuruc and Manley (forthcoming) for a useful discussion.

[11] As Jedwab et al., (2022) point out, however, it should be noted that there is some disagreement among economic historians about "the degree to which the post–Black Death era was a 'golden age'

How would these economic changes affect fertility decisions? One possibility is that they would have no effect whatsoever. There is some debate in the academic literature about the importance of economic factors for fertility decisions (see e.g., De Silva and Tenreyro 2020). Some argue that fertility rates are primarily determined by cultural factors, such as social norms for the ideal family size. We call this the *social determinants model*. In this model, a population shock may have a persistent, proportional effect on long-run population levels, as the changes in economic conditions leave fertility choices unaltered. A 50% drop in population would result in population levels that are lower by 50% indefinitely (or at least until there is a change in the underlying social determinants of fertility). The social determinants model therefore suggests that the long-term value ratio of reducing the risk of catastrophes of any size is equal to 1, and so extinction risk mitigation is neither more nor less cost-effective than the mitigation of smaller catastrophes.

There are, however, channels through which economic factors could plausibly affect survivors' fertility decisions. The *Malthusian model* is perhaps the most well-known model of the economic determinants of fertility (Malthus 1798; see Becker 1988 for a more modern account of the model). This model emphasizes the income effect: as people's income and wealth increase, they can afford to have more children. It also emphasizes that production is constrained by the (fixed) quantity of natural resources. (Note that, in line with Becker (1988), we use the term 'Malthusian' in a broad sense to describe any model in which population is limited by a binding natural resource constraint. Importantly, this is compatible with any level of average consumption in the steady state, depending on people's fertility preferences.)[12]

---

for workers" (p. 150), and "the extent to which these developments were driven by demographics" (p. 150).

[12] In contrast, Malthus (1798:40) held the narrower view that steady state consumption must be at the level of subsistence, as "[t]he passion between the sexes has appeared in every age to be so nearly the same that it may always be considered, in algebraic language, as a given quantity".

The Malthusian model supports the recovery assumption that underlies the long-run argument for prioritizing extinction risk mitigation. To see this, consider a Malthusian economy with replacement fertility, and imagine a sudden negative shock to the size of the current population. Such a shock would leave more natural resources to go around, leading to higher wealth and income per person. Because of the income effect, this would in turn lead to above-replacement fertility that would remain until the population level recovers to its original size. Therefore, the Malthusian model implies that reducing the risk of extinction has a higher long-term value ratio than reducing the risk of smaller population shocks, whose effects are temporary. (That said, as we will argue in the next section, the Malthusian model also highlights the possibility of other interventions that may be as cost-effective as extinction risk mitigation.)

Another way in which changes in economic factors may affect fertility is through the substitution effect. Because labor shortages lead to higher wages, people have an incentive to work more. One way to have more time to work is to have fewer children. As a result, people may decide to have fewer children after a shock that reduces the size of the population. If this were to happen, the shock would be amplified: for example, a 50% drop in population may lead to long-run population levels that are even less than 50% of what they would have been otherwise. In the case of such shocks, the long-term value ratio would be greater than 1. Given our theoretical framework, reducing the probability of such shocks would therefore be more cost-effective than extinction risk mitigation.

To assess the relative strengths of the income and substitution effects, we present a calibration of the *Barro-Becker model* (Becker and Barro 1988; Barro and Becker 1989). The Barro-Becker model is the workhorse model for studying economic determinants of modern fertility dynamics. In this model, all capital is reproducible, so long-run population levels are not constrained by fixed natural resources. Since people are assumed to get utility both from consumption and from having children, the model allows for both the income effect and the substitution effect. The model and our calibration of it, which uses standard parameter values, are detailed in the appendix.

The results of our calibration of the Barro-Becker model are illustrated in Figure 1. The blue continuous line in Figure 1A plots the relationship between the size of the initial population shock and the resulting drop in steady state population levels, as implied by our calibration. For initial population shocks that are relatively small ($<13\%$), the eventual drop in steady state population levels is proportionally smaller than the initial shock. However, for (non-extinction) initial population shocks that are relatively large ($>13\%$), the reverse is true. In other words, our calibration implies that the relative strength of the substitution effect compared to the income effect increases in the size of the initial shock.

The blue continuous line in Figure 1B translates this relationship into undiscounted cost-effectiveness by plotting the long-term value ratio associated with reducing the risk of a shock against the size of that shock.[13] The long-term value ratio is below 1 when population levels recover from the initial shocks due to the income effect, and above 1 when the initial shock is amplified due to the substitution effect. Interestingly, our calibration implies risks that result in roughly a 35% drop in the size of the initial population are those with the highest long-term value ratio. We take these results to suggest that there could indeed be risk mitigation efforts for which the long-term value ratio is greater than 1. However, given the uncertainty associated with the model and its parametrization, we caution readers from drawing any conclusions stronger than this based on our calibration.

---

[13] To make this translation we assume that $(U_i - U_0)$ is the proportional drop in initial population size, and that $(W_i - W_0)$ is the proportional drop in steady state population levels. Thus, Figure 1A plots $(W_i - W_0)$ against $(U_i - U_0)$, whereas Figure 1B plots $(W_i - W_0)/(U_i - U_0)$ against $(U_i - U_0)$.

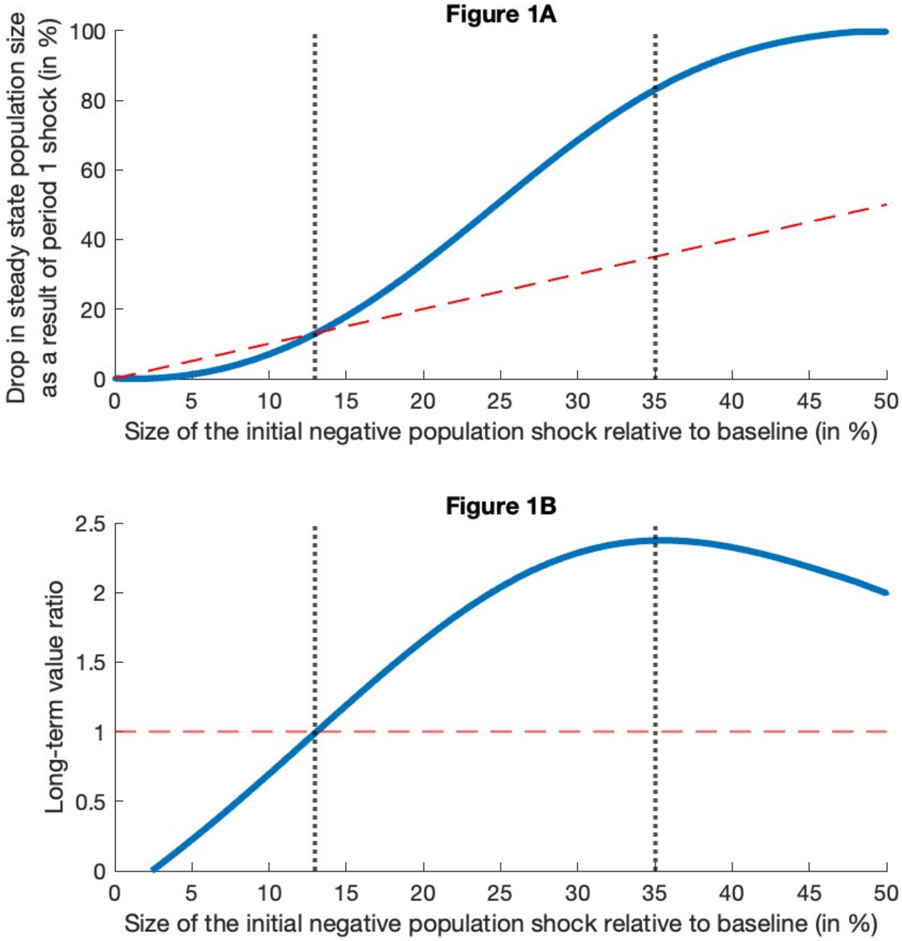

Figure 1: The blue continuous lines plot the drop in steady state population levels (Figure 1A) and the long-term value ratio (Figure 1B) against the size of the initial population shock, as implied by our calibration of the Barro-Becker model. The red dashed lines represent the case where the percentage drop in the initial population level is the same as the percentage drop in steady state population levels.

4.2 How plausible is the Malthusian model?

So far, we have illustrated that the recovery assumption that underlies the long-run argument for prioritizing extinction risk mitigation is supported by the Malthusian model, but not by the social determinants model nor by the Barro-Becker model.

In particular, the social determinants model suggests that extinction risk mitigation is neither more nor less cost-effective than the mitigation of smaller risks, and our calibration of the Barro-Becker model suggests that extinction risk mitigation might be less cost-effective than some other risk mitigation efforts. To evaluate the long-run argument, it is therefore of particular interest to further assess the plausibility of the Malthusian model, which does support the recovery assumption.

The first thing to note is that the Malthusian model is broadly considered irrelevant for explaining modern fertility dynamics. Capital accumulation and technological progress generated by industrialization have vastly increased the efficiency by which natural resources are utilized to the point that they no longer place binding constraints on population levels. Moreover, contrary to the predictions of the Malthusian model, fertility has fallen substantially in modern economies since the industrial revolution while wealth and income per capita have grown.

The case for nonetheless considering the Malthusian model is that Malthusian population dynamics may reemerge in the long run. First, evolutionary pressures for higher fertility might increase long-run population growth to the extent that natural resource constraints become binding once more (cf. Bostrom 2004; Collins and Page 2019).[14] Second, the development of artificial intelligence might result in rapid prolonged growth of machine labor and reproducible capital that eventually hits binding resource constraints such as energy or land (cf. Hanson 2016: 162-166; Korinek and Stiglitz 2018: 383-386). In either of these cases, non-extinction shocks to population would only have temporary effects on long-run population levels, so the long-term value ratio associated with reducing such shocks would be lower than for extinction

---

[14] In contrast, Arenberg et al. (forthcoming) argue that "empirical facts and models of heritability do not provide reason to conclude that positive population growth is bound to continue via the dynamics of a higher-fertility type making up an ever-increasing share of the global population".

risk mitigation.[15] That said, it should be noted that fertility rates are currently below replacement in many high and middle income countries, suggesting that the long-run trend could be towards population decline rather than population increase (Basten et al. 2013; Geruso and Spears, forthcoming). Moreover, in the next section, we argue that the Malthusian model suggests that there might exist interventions other than extinction risk mitigation with a long-term value ratio of 1.

5 Shocks to all factors of production

5.1 Theoretical considerations

The previous section focused on what population models imply about the relationship between long-run population levels and shocks that reduce population while leaving other factors of production unaltered. In the Malthusian model, such shocks have no long-run effect at all, but in other models they may have proportional or even disproportional long-run effects. However, there is no reason to restrict attention to population shocks that leave other factors of production unaltered. Many shocks that affect population size also affect other factors of production. For example, wars result in human casualties, but also in the destruction of factories and cities. Similarly, climate change is likely to result in a large loss of lives, but also in a loss of natural resources. By considering the possibility of shocks that affect all factors of production, we can arrive at the more robust conclusion that, at least in our cost-effectiveness framework, there are interventions that are as cost-effective as extinction risk mitigation.

---

[15] It could also be argued that, even if one thinks that Malthusian future scenarios are unlikely, these scenarios are also disproportionately important from a long-term perspective (because some of these are the scenarios that contain the most social value) and could therefore nonetheless dominate expected social welfare calculations.

The key insight underlying this conclusion is that the economic factors affecting fertility are invariant to the scale of the economy. Population models typically assume that parents' decisions depend on their own wealth and income, but not on how many other people there are. Given constant returns to scale technology, per capita wealth and per capita income are not determined by the scale of the economy, but by the ratio of capital to labor. This feature implies that changing the scale of the economy, i.e., proportionally changing all factors of production, has no effect on fertility decisions. Consequently, these models imply that any shock that proportionately changes labor and capital would have a permanent effect on population levels.

Consider, for instance, a scenario in which a nuclear war kills 50% of the population *and* destroys 50% of the capital stock. Since the capital-labor ratio would be unaltered, constant returns to scale production technology implies that wages and rental rates would also be unaltered.[16] Economic fertility models would then typically imply that people choose to have the same number of kids as they would otherwise have had, which implies that the population size would remain permanently 50% lower. It follows that the long-term value ratio associated with reducing the risk of such shocks is 1. Therefore, our theoretical framework suggests that the undiscounted cost-effectiveness of mitigating these risks is as high as that of mitigating the risk of human extinction.

One might question whether this hypothetical possibility is empirically relevant. For example, perhaps there are no available interventions for reducing the likelihood of shocks that would proportionally reduce all factors of production. Alternatively, perhaps, for whatever reason, policymakers behave less myopically when it comes to shocks that proportionately affect all factors of production. To address these con-

---

[16] An important caveat that we ignore here is that population levels may well have important effects for the rate of technological progress, as emphasized by e.g., Jones (2022).

cerns, we propose a concrete intervention and provide a back-of-the-envelope calculation indicating that, if the Barro-Becker model (in which all capital is reproducible) holds indefinitely, our proposed intervention is more cost-effective than extinction risk mitigation.

5.2 Back-of-the-envelope calculation

Consider an intervention that saves lives and proportionally increases the stock of reproducible capital.[17] Further, assume that reproducible capital can substitute for natural resources in production, so that the Barro-Becker assumption of constant returns to scale is plausible in the long run.

Distributing bed nets in malaria-prone regions in low-income countries is considered a highly cost-effective way of saving lives. According to a recent estimate by GiveWell, it costs around $5,000 to save a life by distributing bed nets (GiveWell 2022).

To maintain a constant capital-labor ratio, this intervention must be accompanied by a proportional change in the capital stock. In other words, the value of the global capital stock must be increased by the current value of the capital stock per person. Importantly, the returns to this capital must accrue to the people whose lives were saved, as their future fertility decisions depend not only on their wages but also on their wealth. An implementation of our intervention would be a combination of distributing bed nets in a malaria-prone region while simultaneously transferring wealth to that region, either as direct transfers or through investment in infrastructure. Note that the wealth transfer would have to consist of resources that would otherwise have

---

[17] Note that standard economic growth models with exogenous population imply that increases in population lead to increases in interest rates, which in turn means that more investments are made, which eventually pushes the capital-labor ratio back to the steady state. These dynamics work differently in the Barro-Becker model, which is why we instead envision an intervention that both saves lives and increases the capital stock.

been consumed rather than invested.

To calculate the required capital investment, note that, according to the World Bank data for year 2021, global GDP per capita is around \$12,000. About two-thirds of this is attributable to labor income, suggesting that capital income per capita is around \$4,000. The standard no-arbitrage condition for investment implies that the marginal product of capital is equal to the interest rate plus the depreciation rate, $r + \delta$. A reasonable parameterization is $r = \delta = 0.05$, so that $r + \delta = 0.1$. Because capital income is equal to the marginal product of capital multiplied by the capital stock, i.e., $\$4{,}000 = 0.1 \cdot k$, the per-capita capital stock, k, is given by:

$$k = \$40{,}000.$$

This suggests that the combined intervention of saving a life and increasing the capital stock to offset the decline in the capital-labor ratio would cost around \$45,000. It is notable that the bulk of the cost is the capital investment component rather than the lifesaving component.

This estimate suggests that, with \$100, it is possible to save $\$100/\$45,000 \approx 0.0022$ of a life while maintaining the capital-labor ratio constant. Given a current world population of around 8 billion, this constitutes a permanent, proportional increase in the population of about $0.0022/(8 \cdot 10^9)$. Using Greaves and MacAskill's (2021) estimate that the expected number of future lives is around 10^{24}, it follows that the total number of lives saved by spending \$100 on our proposed intervention is

$$\frac{0.0022 \cdot 10^{24}}{8 \cdot 10^9} = 2.75 \cdot 10^{11}.$$

While estimates such as the one above should not be interpreted literally (see Karnofsky 2011), it is nonetheless worth noting that the estimated returns of our proposed intervention are substantially higher than the returns that Greaves and MacAskill (2021) estimate for extinction risk mitigation. According to their estimates, the expected number of lives saved from spending \$100 on asteroid detection is 300,000 and the expected number of lives saved from spending \$100 on biosecurity is $2 \cdot 10^8$. Hence, astonishingly, compared to their biosecurity estimate, the back-of-the-envelope calculation above suggests that our proposed intervention saves around 1,000 times more lives for the same amount of money.

There are, of course, many extremely simplifying assumptions that go into the back-of-the-envelope calculation above. Importantly, it assumes that the Barro-Becker model holds indefinitely, which seems questionable given that future technology may allow for very different modes of reproduction. Technological developments could for instance potentially have substantial effects on fertility decisions and population growth by facilitating sex selection (Kolk and Jebari 2022), cloning (Saint-Paul 2003), and perhaps even mind-uploading (Hanson 2016). Moreover, combining life-saving with capital investment admittedly amounts to an unconventional and perhaps politically impractical intervention.

Our aim here is not to argue that our estimate is reliable or that our proposed intervention is among the most effective ways of improving long-run welfare. Our aim is rather to illustrate, using an empirically grounded example, that there may indeed be interventions other than extinction risk mitigation that are cost-effective in virtue of having a long-run effect on the size of the global population.

5.3 Increasing the stock of natural resources

Although the back-of-the-envelope calculation above assumes that reproducible capital can substitute for natural resources in production, a conceptually similar calculation could in principle be performed in the case of the Malthusian model where reproducible capital and natural resources are not substitutes. In the Malthusian model, permanently increasing the supply of natural resources by some proportion would (via the income effect) increase long-run population levels by the same proportion. For example, if the quantity of arable land constrains long-run population levels, then our cost-effectiveness framework implies that preventing permanent destruction of arable land would increase short-run and long-run social value by the same proportion, and therefore have a long-term value ratio of 1 – the same as extinction risk mitigation.

Similarly, if we anticipate humanity (or its descendants) to eventually become an intergalactic civilization, there is a resource constraint in the form of the number of reachable galaxies. For each year that intergalactic space expansion is delayed, $(2 \cdot 10^{-8})\%$ of the reachable universe is permanently lost due to the exponentially accelerating expansion of the universe (Ord ms-b: 23; also cf. Armstrong and Sandberg 2013). The Malthusian model therefore suggests that speeding up intergalactic space expansion would have a permanent effect on population levels in the far future. If there are interventions that would accelerate intergalactic space expansion without having any short-run benefits, the long-term value ratio of these interventions could be greater than that of extinction risk mitigation.[18] However, it is worth noting that previous literature addressing this question generally finds that extinction risk mitigation is much more cost-effective than speeding up space expansion (Bostrom 2003, Ord ms-b).

---

[18] Note that this conclusion only holds under the rather strong assumption that the myopic policymakers spend some positive amount of resources on these space expansion interventions (perhaps because they care about the far future a little bit). Without this assumption, the policymakers' problem has a corner solution, which violates the conditions we used to derive the long-term value ratio in section 3.1.

6 Conclusions

Assessing the long-run argument for prioritizing extinction risk mitigation is important for philanthropic priority setting. Drawing on standard economic fertility models, this paper poses a challenge to this argument. As illustrated in section 5, such models typically imply that any shocks that proportionally decrease all factors of production have proportional, permanent effects on long-run population levels. Therefore, in our theoretical cost-effectiveness framework, the undiscounted cost-effectiveness of mitigating such shocks is comparable to that of extinction risk mitigation. Moreover, a back-of-the-envelope calculation, using the Barro-Becker model and plausible empirical estimates, implies that our proposed intervention, which combines bed net distribution with wealth transfers, is more cost-effective than extinction risk mitigation (provided that the Barro-Becker model holds indefinitely). Although these cost-effectiveness estimates are mainly intended to serve as helpful illustrations and should therefore be interpreted with considerable caution, they nonetheless suggest that interventions other than extinction risk mitigation could have significant impact on long-run social welfare.

In addition, our analysis of pure population shocks in section 4 highlights the possibility that the most cost-effective interventions might be those that mitigate large, non-extinction catastrophes rather than those targeted at extinction risk mitigation. The reason for this is that some reasonable fertility models have nontrivial long-run dynamics: a large, non-extinction shock to population may be amplified in the long-run. However, more work is needed to assess the likelihood of such amplification as well as possible ways to mitigate shocks of this kind.

Our challenge to the long-run argument for prioritizing extinction risk mitigation is thus not that extinction risk mitigation is less cost-effective than the argument purports, but rather that there may exist other interventions that are equally or perhaps even more cost-effective. In particular, we point out that (a) the argument seems to rely on the assumption that humanity would eventually recover after any shock to population, and (b) the recovery assumption is violated by standard population models with the important exception of the Malthusian model.

Our challenge is particularly pressing if fertility rates are expected to remain low in the long run or if the exogenous rate of human extinction is expected to be high. In both of these scenarios, long-run population levels are unlikely to be governed by Malthusian dynamics. Conversely, the long-run argument for prioritizing extinction risk mitigation seems more resilient to our challenge if evolutionary or technological factors are expected to result in large future population levels limited by binding natural resource constraints. However, as argued in subsection 5.3, there might be interventions that have higher undiscounted cost-effectiveness than extinction risk mitigation even in these Malthusian scenarios.

Our discussion also provides insights about the potential long-term effects of different *types* of global catastrophes. In particular, it suggests that the extent to which a catastrophe destroys capital is an important factor for assessing recovery dynamics. Asteroids or wars, which result in both deaths and in the destruction of capital, are likely to have very different long-run population effects compared to pandemics, which could result in the same numbers of deaths while leaving the capital stock largely intact. Our results point to the possibility that the former type of catastrophe may lead to a proportional loss in long-run population levels, whereas the latter type of catastrophe may result in either a disproportionately large or a disproportionately small long-run population effect. This suggests that assessing the potential of catastrophic events to destroy reproducible and natural resources – in addition to their potential to cause fatalities directly – may be of special significance for long-term welfare.

References

Adamczewski, T. (ms), *The expected value of the long-term future.* Unpublished manuscript.

Arenberg, S. *et al.* (forthcoming), 'Heritable Fertility Is Not Sufficient to Prevent Long-run Population Decline', in *Demography.*

Armstrong, S., and Sandberg, A. (2013), 'Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox', in *Acta Astronautica* 89:1-13.

Barro, R. J., and Becker, G. S. (1989), 'Fertility Choice in a Model of Economic Growth', in *Econometrica* 57/2: 481-501.

Basten, S., Lutz, W., and Scherbov, S. (2013), 'Very Long Range Global Population Scenarios to 2300 and the Implications of Sustained Low Fertility', in *Demographic Research* 28: 1145-1166.

Becker, G. S. (1988), 'Family Economics and Macro Behavior', in *The American Economic Review* 78/1: 1-13.

Becker, G. S., and Barro, R. J. (1988), 'A Reformulation of the Economic Theory of Fertility', in *The Quarterly Journal of Economics* 103/1: 1-25.

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future.* Doctoral dissertation. Rutgers The State University of New Jersey, School of Graduate Studies.

Blackorby, C., Bossert, W., and Donaldson, D. (1995), 'Intertemporal Population Ethics: Critical-level Utilitarian Principles', in *Econometrica* 63/6: 1303-1320.

Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', in *Utilitas* 15/3: 308-314.

Bostrom, N. (2004), 'The Future of Human Evolution', in Bostrom, N., Ettinger, R. C. W., and Tandy C. (eds.), *Death and Anti-death: Two Hundred Years after Kant, Fifty Years After Turing* (Ria University Press), 339-371.

Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', in *Global Policy* 4/1: 15-31.

Collins, J., and Page, L. (2019), 'The Heritability of Fertility Makes World Population Stabilization Unlikely in the Foreseeable Future', in *Evolution and Human Behavior* 40/1: 105-111.

Cordoba, J. C. (2015), 'Children, Dynastic Altruism and the Wealth of Nations', in *Review of Economic Dynamics* 18/4: 774-791.

Dasgupta, P. (2008), 'Discounting Climate Change', in *Journal of Risk and Uncertainty* 37/2: 141-169.

De Silva, T., and Tenreyro, S. (2020), 'The Fall in Global Fertility: A Quantitative Model', in *American Economic Journal: Macroeconomics* 12/3: 77-109.

Geruso, M. and Spears, D. (forthcoming), 'With a Whimper: Depopulation and Longtermism', in Barrett, J., Thorstad, D., and Greaves, H. (eds.), *Essays on Longtermism* (Oxford University Press).

GiveWell. (2022), *Our Top Charities*. Link: https://www.givewell.org/charities/top-charities (Accessed 31 December 2022).

Greaves, H. (2017), 'Discounting for Public Policy: A Survey', in *Economics & Philosophy* 33/3: 391-439.

Greaves, H., and MacAskill, W. (2021), *The Case for Strong Longtermism*, GPI Working Paper No. 5-2021.

Groom, B., Drupp, M. A., Freeman, M. C., & Nesje, F. (2022), 'The Future, Now: A Review of Social Discounting', in *Annual Review of Resource Economics* 14:467-491.

Häggström, O. (2016), *Here Be Dragons: Science, Technology and the Future of Humanity.* Oxford University Press.

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

Jedwab, R., Johnson, N. D., and Koyama, M. (2022), 'The Economic Impact of the Black Death', in *Journal of Economic Literature* 60/1: 132-78.

Jones, C. I. (2022), 'The End of Economic Growth? Unintended Consequences of a Declining Population', in *American Economic Review* 112/11: 3489-3527.

Karnofsky, Holden. 2011. 'Why We Can't Take Expected Value Estimates Literally (Even When They're Unbiased)', *The GiveWell Blog.* Link: https://blog.givewell.org

/2011/08/18/why-we-cant-takeexpected-value-estimates-literally-even-when-theyre-unbiased/ (Accessed 31 December 2022).

Kolk, M., and Jebari, K. (2022), ’Sex Selection for Daughters: Demographic Consequences of Female-Biased Sex Ratios’, in *Population Research and Policy Review* 41: 1-21.

Korinek, A., and Stiglitz, J. E. (2018), ’Artificial Intelligence and its Implications for Income Distribution and Unemployment’, in Agrawal A., Gans J., and Goldfarb A. (eds.), *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press), 349-390.

Kuruc, K. and Manley, D. (forthcoming), ’The Ethics, Economics, and Demographics of Delaying Senescence’, in Barrett, J., Thorstad, D., and Greaves, H. (eds.), *Essays on Longtermism* (Oxford University Press).

MacAskill, W. (2022), *What We Owe the Future* (Basic Books, 2022).

Malthus, T. (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).

Matheny, J. G. (2007), ’Reducing the Risk of Human Extinction’, in *Risk Analysis: An International Journal* 27/5: 1335-1344.

Millett, P., and Snyder-Beattie, A. (2017), ’Existential Risk and Cost-Effective Biosecurity’, in *Health Security* 15/4: 373-383.

Ng, Y. K. (2016), ’The Importance of Global Extinction in Climate Change Policy’, in *Global Policy* 7/3: 315-322.

Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Hachette Books).

Ord, T. (ms-a), The Value of Very Long-reaching Effects. Unpublished manuscript.

Ord, T. (ms-b), ‘The Edges of Our Universe’. *arXiv*:2104.01191. Link: https://arxiv.org/pdf/2104.01191.pdf (Accessed 31 December 2022).

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

Parfit, D. (2017), *On What Matters: Volume 3* (Oxford University Press).

Posner, R. A. (2004), *Catastrophe: Risk and Response* (Oxford University Press).

Saint-Paul, G. (2003), 'Economic Aspects of Human Cloning and Reprogenetics', in *Economic Policy*, 18/36: 73-122.

Schulman, C. and Thornley, E. (forthcoming), 'How Much Should Governments Pay to Prevent Catastrophes? Longtermism's Limited Role', in Barrett, J., Thorstad, D., and Greaves, H. (eds.), *Essays on Longtermism* (Oxford University Press).

Spears, D., and Zuber, S. (2021), *Foundations of Utilitarianism Under Risk and Variable Population.* IZA Discussion Paper No. 14515, Available at SSRN: https://ssrn.com/abstract=3879363 (Accessed 31 December 2022).

Tarsney, C., and Thomas, T. (2020), *Non-additive Axiologies in Large Worlds.* GPI Working Paper No. 9-2020.

Thorstad, D. (2022), *Existential risk pessimism and the time of perils*, GPI Working Paper No. 1-2022.

Trammell, P. (2021), *Dynamic Public Good Provision under Time Preference Heterogeneity: Theory and Applications to Philanthropy.* GPI Working Paper No. 9-2021.

Appendix

This appendix introduces the Barro-Becker model in more mathematical detail (see Becker and Barro 1988; Barro and Becker 1989, for further discussion) and describes how we calibrated the model to generate Figure 1. Time is discrete and infinite, and each generation is alive in one time period only. The utility of each person alive in period t is

$$u_t = c_t^{\sigma} + \alpha n_t^{1-\epsilon} u_{t+1}$$

where $c_t > 0$ is the consumption of a person in generation t; $\sigma \in (0, 1)$ captures the marginal utility of consumption, $n_t \geq 0$ is the number of children; $\alpha > 0$ is a parameter that governs how much people value having children; $(1 - \epsilon) \in (0, 1)$

captures the marginal utility of having children; and u_{t+1} is the expected utility of each child.

People allocate their labor incomes, w_t, and their capital incomes, $(1 + r_t)k_t$, between their own consumption, child rearing expenses, and saving for the benefit of their children. The budget constraints are thus given by

$$w_t + (1 + r_t)k_t = c_t + n_t(a(1 + g)^t + bw_t) + n_t k_{t+1}.$$

The cost of raising children is given by $a(1+g)^t + bw_t$. The first component of the cost, $a(1 + g)^t$, is a cost in terms of goods, which is assumed to grow at the rate of technological progress. The second component of the cost, bw_t, is a time cost: each child requires sacrificing a fraction $b \in (0, 1)$ of the individual's work time.

Output is produced using constant returns to scale technology in capital and labor. Technological progress is constant and labor augmenting. Output at time t is given by

$$Y_t = A(N_t k_t)^{\zeta}((1 + g)^t N_t(1 - bn_t))^{1-\zeta}$$

where g is the rate of technological progress, A is the baseline productivity level, and N_t is the population size in period t, and ζ is the capital intensity parameter. In this model, the interest rate, r_t, and the wage rate, w_t, are determined based on the marginal products of capital and labor, respectively.

In our calibration, we focus on a steady state without technological progress $g = 0$, reflecting the hypothesis that, in the long run, the stock of knowledge will converge, as new ideas will get increasingly harder to find, and substantial resources will have to be devoted towards maintaining the stock of knowledge. The length of a period is taken to be 25 years, roughly corresponding to the age of fertility. The depreciation rate, δ, is chosen based on an annual depreciation rate of 5% a year, roughly in line with global averages. The capital intensity parameter, ζ, is chosen to roughly match the long-run capital income share.

The preference parameters, σ and ϵ, and child-rearing cost parameters, a and b, are taken directly from the calibration in Cordoba (2015). The parameter α is calibrated to generate a steady state with constant population. The productivity parameter, A, is normalized to match average wages, which are specified in 2010 dollars (this is not an important normalization). Given these parameter values, the Barro-Becker model implies the relationship between initial shock size and steady state population drop implied by Figure 1.

Table 2: Calibration parameters

Parameter	Description	Value
δ	Capital depreciation rate	0.72
σ	Intergenerational substitutability parameter	0.3
ϵ	Diminishing returns to number of children	0.288
a	Material cost of childrearing	11.4
b	Time cost of childrearing	0.16
α	Intergenerational altruism parameter	0.09
ζ	Capital intensity of production	1/3