Mistakes in the moral mathematics of existential risk

David Thorstad

1 Introduction

I Many authors give very high estimates for the value of existential risk mitigation:

  i "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, p. 18).

  ii Estimated cost of saving a life through SpaceGuard survey was $0.14/life (Greaves and MacAskill 2021).

II The aim of this paper is to discuss three mistakes that arise in calculating the value of existential risk mitigation:

  i Focusing on cumulative risk over per-unit risk (§2).

  ii Ignoring background risk (§3).

  iii Ignoring population dynamics (§4).

III These mistakes have two effects on debates:

  i Debates are **mislocated** because they ignore crucial parameters.

  ii The value of existential risk mitigation is **overstated**.

IV Correcting these mistakes will reveal important new directions for future work (§5).

2 Mistake 1: Focusing on cumulative risk

I **Bostrom:**

  i Suppose humanity would last for a billion years at a population of one billion people.

  ii Then the future holds 10^{18} life-years, or about 10^{16} lives at current lifespans.

  iii **Claimed implication:** An intervention which reduces risk by just one millionth of one percent would be as valuable as an intervention that saved one hundred million present people.

> 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 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, p. 18).

II **Why this seems tempting:** A prospect which saves 10^{16} lives with probability 10^{-8} saves in expectation $10^{16} * 10^{-8} = 10^{8}$ lives.

III **To see what's gone wrong here:** We need two distinctions.

IV **Distinction 1: Absolute versus relative risk**

Table 1: Varieties of risk reduction

	Gloss	Definition	Example of 10% reduction
Absolute reduction	Subtract fixed amount	r to r-f	80% ->70% 20% ->10%
Relative reduction	Cut away fixed fraction	r to (1-f)r	80% ->72% 20% ->18%

Bostrom is concerned wiith an *absolute* risk reduction of 10^{-8}.

V **Distinction 2: Cumulative versus per-unit risk**

I Over a long period (say, 1 billion years) we face **cumulative risk** r_C of going extinct, and survive with probability $1 - r_C$.

II Each interval in that period (say, 1 century) we face **per-unit risk** r_U of going extinct.

III **Relationship between cumulative and per-unit risk:** $r_C = 1 - (1 - r_U)^N$ over a period of length N (here N is ten million).

IV **Important observation:** Bostrom is concerned with cumulative risk.

VI **Why Bostrom's claim is misleading:**

i An absolute reduction of 10^{-8} in cumulative risk would drive $r_C \leq 1 - 10^{-8}$.

ii This requires $r_U \approx 1.6 * 10^{-6}$ or lower each century.

iii **Making the reduction look small:** We only need an absolute reduction of cumulative risk by 10^{-8}.

iv **Making the reduction look large:** We need to drive per-century risk to about one in a million, whereas many effective altruists think it's now at least one in ten.

VII **Why focus on per-unit risk:**

i **Numbers are more perspicuous:** Seemingly 'small' changes in cumulative risk require large changes in per-unit risk.

ii **Focus on what you can affect:** It's hard to act on r_C. More plausibly, our acts change r_U.

3 Mistake 2: Ignoring background risk

3.1 Introducing the MSB model

I **Millett and Snyder-Beattie (2017):** Argue that biosecurity interventions aimed at preventing existential catastrophe are cost-effective by standard cost-effectiveness metrics.

II **MSB model:** Separately estimate four parameters (Table 2):

Table 2: MSB model parameters

Parameter	Interpretation	Estimation
C	Cost of intervention	\$250 billion
N	Number of biohazards/century (without intervention)	3 models
L	# of Life-years saved by stopping a catastrophe = (??) size of future population given no catastrophes = 10 billion people for one million years	10^{16} life-years
R	%Reduction in relative risk from intervention (of biological catastrophe)	1%

III Number of biohazards/century N is given range estimates across three models (Table 3):

Table 3: MSB cost-effectiveness estimates

Model	N (biothreats/century)	C/NLR (cost/life-year, USD)
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

IV **Apparent conclusion:** Biorisk reduction is robustly cost-effective across models.

V **A complaint to set aside:** L,N are (highly) probabilistically dependent, so they must be estimated together.

3.2 A new mistake: Ignoring background risk

I **Introducing background risk:** We can split per-century extinction risk r into its biorisk component b and non-biorisk component n as

$$r = b + n. \tag{1}$$

II **Reducing biorisk:** Intervention X provides 1% relative reduction in biorisk *across all centuries*, so that risk is now:

$$r_X = 0.99b + n. \tag{2}$$

III **Expected number of future life-years:**

$$E[L|X] = 10^{12} \sum_{i=1}^{10,000} (1 - r_X)^i. \tag{3}$$

IV **Expected *additional* number of future life-years:** (see appendix for derivation):

$$E[L|X] - E[L] = 10^{12} * \frac{0.01b}{r(r - 0.01b)}. \tag{4}$$

V **Problem:** Many EAs think r is high, so that $r >> b$, driving down (6).

Table 4: 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.[1]
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

VI **Salient comparison:** GiveWell puts best short-termist interventions at around $5,000/life, so around $50-200/life-year, even ignoring other effects.

3.3 An old mistake: Focusing on cumulative risk reduction

I **Observation:** Our intervention reduces near-term risk (say, this century), not cumulative risk.

II **Reducing biorisk:** Intervention X' provides 1% relative reduction in biorisk *in this century*, so that risk this century drops to:

$$r_{X'} = 0.99b + n. \tag{5}$$

III **Expected *additional* number of future life-years:**[2]

$$E[L|X'] - E[L] = 10^{12} * 0.01b/r. \tag{6}$$

Table 5: MSB cost-effectiveness estimates against doubly 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	250 to 1,000	13 to 50.[3]
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

IV **Salient comparison:** GiveWell puts best short-termist interventions at around $\$5,000$/life, so around $\$50 - 200$/*life-year*, even ignoring other effects.

4 Population dynamics

> Future people count. **There could be a lot of them.** We can make their lives go better. (MacAskill 2022).

4.1 Population size and carrying capacity

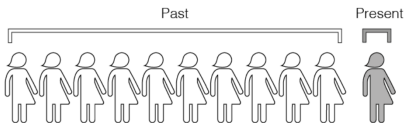
I **Motivating insight:** Many authors estimate the future human population by multiplying (a) the **carrying capacity** of a given region, (b) the **duration** of human life in that region.

Table 6: Estimates of future lives based on duration and carrying capacity

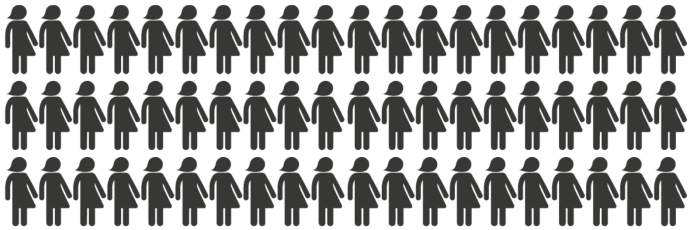
Scenario	**Carrying capacity (Lives/century)**	**Duration (centuries)**	**Future lives**
Earthbound (Bostrom)	10^9	10^7	10^{16}
Earthbound (MacAskill)	10^{10}	$5 * 10^6$	$5 * 10^{16}$
Earthbound (Greaves/MacAskill)	10^{10}	10^4	10^{14}
Solar System (Greaves/MacAskill)	10^{19}	10^8	10^{27}
Milky Way (Greaves/MacAskill)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

II **These estimates are very large:** MacAskill (2022) says we've only had ten 'stick figures' of humanity but have five million 'stick figures' to come.

[2]To see this, apply to formula for the value of absolute risk reduction from (Thorstad 2022), Section 3.1, with v = 10^{12} and $f = 0.01b$.



Next, we'll represent the future. Let's just consider the scenario where we stay at current population levels, and live on Earth for five hundred million years. These are all the future people:



.

III **One way to push back:** Incorporate background risk (correct second mistake):

Table 7: 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}

IV **Another way to push back:** Consider population dynamics (they're increasingly divorced from carrying capacity).

4.2 Standard demographic models

I **Broad consensus among demographers:**

i Human population dynamics are increasingly non-Malthusian: factors beyond resource constraints drive population dynamics (Barro and Becker 1989).

ii World population growth will be near-zero, probably declining by 2100 at a population of around 10-12 billion people (United Nations 2022).

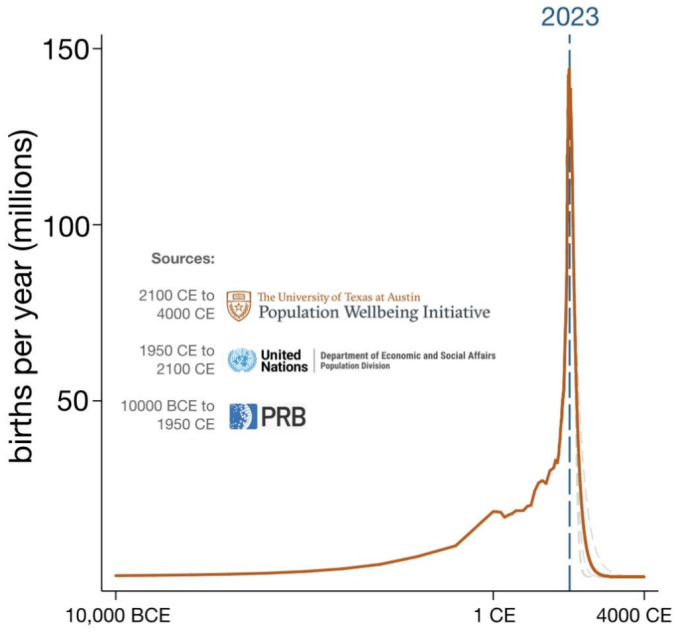
iii If we are comfortable making longer projections, it is likely that fertility will be (substantially) below replacement until at least 2300 (Basten et al. 2013; Raftery and Sevcikova 2023; Geruso and Spears forthcoming; Spears et al. 2023).

iv We have no especially good reason to think that fertility will ever recover (Geruso and Spears forthcoming; Lutz et al. 2014).

II **If this is right:** The future human population could be very small indeed.

III **An example:** A study out of the Population Wellbeing Initiative (Geruso and Spears forthcoming) projects the future human population, given fertility rates converging to 1.66 births/woman, at about 30 billion unborn people (!!).

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



And this prediction is robust to other levels of sub-replacement fertility:

Figure 2: Robustness of Geruso/Spears projections, from Geruso and Spears (forthcoming)

hypothetical asymptotic fertility:	1.8	1.66	1.5	1.2	1.0
(example 2023 country or region, according to UN)	South America	US	Europe	East Asia	South Korea
% of 2023 world population in countries at or below this fertility rate	43%	38%	25%	19%	1%
% of all human lives which would have been already born	77%	82%	85%	86%	86%
% of all human lives which would remain yet to be born	23%	18%	15%	14%	14%
number of future stick figures (future births ÷ 10 billion, rounded)	3	3	2	2	2

4.3 Nonstandard demographic models

I **Objection:** What if humanity *does* permanently re-enter a Malthusian regime?

II **Response 1 (offstage):** This is not a mainstream scientific view, and might require some uncomfortable sacrifices.

III **Response 2 (onstage):** Demographic modeling still drags down the value of existential risk mitigation.

IV **A techno-optimist, Malthusian model:** Tarsney (2022).

i After 1,000 years, humanity settles stars in all directions at a breakneck pace $\mathbf{s} = 0.1c$, for c the speed of light.

ii This continues until

i Extinction, with constant probability **r**/year.

ii The universe becomes uninhabitable at distant time $\mathbf{t_f}$.[4]

iii Each year, we reap value:

i $\mathbf{v_e}$ from earth settlement, set to 6 billion QALYs.

ii $\mathbf{v_s}$ from each settled star, set to 300 million QALYs.

iv Compare two interventions, each costing a million dollars:

i **N**, a neartermist intervention providing 10,000 QALYs.

ii **L**, a longtermist intervention reducing the (absolute) risk of extinction in the next millennium by $p = 2 * 10^{-14}$.

v Then where $n(x)$ returns the number of stars in a radius of x light-years around earth, L is valued at:

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

vi This gives $E[L] > E[N]$ when $r < 0.000135$, putting per-century risk around 1.34%.

vii **Inserting population dynamics:** An example (optimistic story):

i It takes 1,000 years for a band of human settlers to grow a planet into a mature colony.

ii Then resource pressures cause young settlers to set forth (in all directions?) towards the nearest star system.

iii Average star is about 5 light years away, bounding the speed of interstellar expansion at about $5c/1,000$.

viii **Re-doing the model:** Plugging $s = 0.005c$ into Tarsney's model, we have now that $E[L] > E[N]$ when $r < 0.0000145$, a per-century risk of about 0.145%.

ix **Comparison to previous model:** At $r = 0.000135$, $E[N] > 500E[L]$.

5 Implications

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

I **Observation from Section 3:** Raising levels of background risk substantially reduces the value of a fixed (relative or absolute) risk reduction.

[4]Note, t_f is largely irrelevant to the comparison between N, L across all reasonable values for t_f.

II **An interesting dialectical flip** (Thorstad 2022) Unless we say more (which maybe we should):

   i **Longtermists should hope risk is low:** Lowering risk estimates raises the value of risk mitigation efforts.

   ii **Short-termists should hope risk is high:** Raising risk estimates lowers the value of risk mitigation efforts.

III **One way out of this flip:** Time of Perils Hypothesis.

5.2 Population dynamics, demographic interventions, and digital minds

I **Observation from Section 4:** Population dynamics are very important, and could drag down the size of the future population.

II **Implication 1:** Efforts to increase the size of the future population can be at least as important as efforts to mitigate existential risk (Eden and Alexandrie forthcoming; Geruso and Spears forthcoming).

III **Implication 2:** We should think seriously about the population dynamics of digital minds. Could these be better?

5.3 Cumulative risk and intergenerational coordination

I Driving down cumulative risk requires **intergenerational coordination** to drive down per-century risk.

II This is a tricky coordination problem for many reasons:

   i Target levels of risk are very low.

   ii Each generation bears only a small fraction of the cost of extinction.

   iii People are often impatient and display limited degrees of altruism.

   iv Hard to enforce sanctions against past/future noncompliance.

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:

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

The number of lives added by X is:

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

References

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.

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

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

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.

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/.

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

Millett, Piers and Snyder-Beattie, Andrew. 2017. "Existential risk and cost-effective biosecurity." *Health Security* 15:373–384.

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

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/.

Thorstad, David. 2022. "Existential risk pessimism and the time of perils." Global Priorities Institute Working Paper 1-2022.

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