

Inequality and AI Existential Risk

Enrico D. Turri*

October 17, 2026

Abstract

AI has the potential to raise living standards dramatically while at the same time creating a risk of human extinction. This paper argues that inequality affects how much existential risk a society accepts: under majoritarian voting, inequality increases society's willingness to take an existential gamble if the value of life is increasing in consumption. Calibrating the model to the US, present-day inequality leads the median voter to accept an extinction probability of 5.7%, against the 4.9% of a society with perfect equality, a one sixth increase. How gains are allocated matters for who favours AI development: when gains are shared proportionally, it is the poor who favour AI, since their lower consumption makes their lives less valuable relative to the gains AI promises. If instead AI increases inequality upon survival, the decisive voter becomes more conservative if she is cut out of the gains, while the rich favour AI. Existential risk also changes the political economy of redistribution: since the rich have more valuable lives to lose, all agents above a finite consumption level support a redistribution that makes the median better off and lowers the risk society takes. In addition, when AI can reduce existential risk from other threats, inequality leads to AI under-development. Redistribution can thus be an important tool in managing existential risk.

Keywords: artificial intelligence, AI, existential risk, value of life, inequality, median voter, redistribution.

JEL Codes: D63, D72, D81, J17, O33.

*London School of Economics. Email: e.turri@lse.ac.uk. I'd like to thank Ethan Ilzetzki, Silvana Tenreyro, and Pietro Garibaldi for their continuous encouragement and support throughout my PhD studies.

1 Introduction

“Now we’re creating alien beings – meaning we’re not the only intelligent beings around. [...] I’m willing to be selfish and choose human beings.”

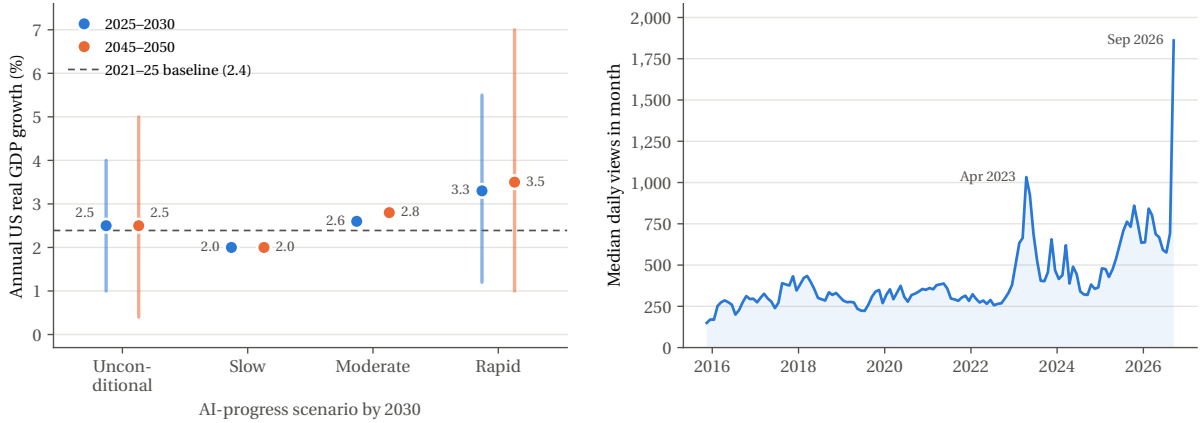
— Geoffrey Hinton, interview with *The Guardian*, 13 October 2026

Artificial intelligence (AI) might be the most consequential technological breakthrough in the history of humankind, and has already revolutionised fields as disparate as chemistry, biology, pure mathematics, coding, and text and image generation. It also has the potential to raise living standards dramatically: economists’ most optimistic forecasts have AI raising GDP growth to 7% a year by 2045–2050, more than three times its long-run average of 2% (Figure 1a; Karger et al., 2026). Yet, developing such a transformative technology could result in human extinction, whether through superintelligent systems escaping human control or by giving actors with limited resources cheap weapons of mass extermination, such as engineered viruses. Concerns about these risks have been raised by prominent computer scientists, as well as by former employees of frontier AI labs (Cecco, 2026; Hajdari, 2026), and have entered the mainstream: on a typical day, the Wikipedia article on existential risk from AI was viewed around 300 times before 2023, more than 1,000 times in April 2023 and 1,860 times in September 2026, about six times as often as before 2023 (Figure 1b). Developing AI is thus a gamble, trading higher living standards against a risk of extinction. This raises the question: how does a society decide whether to develop AI and how much to risk, and what affects the eventual decision?

This paper argues that how much existential risk a society accepts is a political choice, and that inequality affects the outcome. The value of being alive depends on how much each individual consumes, but consumption is distributed unequally; this implies that for the same gains from AI people differ in how much they stand to lose, and in turn they also differ in the risk they are willing to accept. A society must nonetheless take a single decision, and this requires aggregating different preferences: whether to develop AI is a problem of political economy, and should be studied using its methods.

The paper begins by showing that richer agents tolerate less existential risk. I develop a model of AI development along the lines of Jones (2024), in which AI can permanently raise consumption growth, but developing it implies a one-shot positive probability of extinction. When considering whether to develop AI, each agent weighs the gain from AI against the life she puts at risk. If the value of life rises with consumption, as long-run and health-spending estimates indicate (Hall and Jones, 2007; Murphy and Topel, 2006), a richer agent has more to lose and accepts less risk for the same proportional gain in consumption. In an unequal society each agent therefore wants a different level of risk, the rich less and the poor more, and the more unequal the society, the more these preferences differ.

Figure 1: Expected gains from AI and existential risk



(a) GDP growth forecasts by AI scenario

(b) AI existential risk Wikipedia article views

(a) Median forecasts of annualized US real GDP growth over 2025–2030 and 2045–2050 by economists surveyed between October 2025 and February 2026 (66 and 64 respondents), unconditionally and conditional on slow, moderate or rapid progress in AI capabilities by 2030. Bars span the median 10th and 90th percentile forecasts, elicited for the unconditional and rapid cases only; the dashed line is the 2021–2025 baseline reported in the source. Data from Karger et al. (2026), Tables 19–20. (b) Median daily views in each month of the English Wikipedia article on existential risk from artificial intelligence, summing over all the titles the article has had, November 2015–September 2026, from the Wikimedia pageviews API.

The main result of the paper is that majoritarian voting makes an unequal society accept more risk than it would under perfect equality, and that redistribution lowers the risk the society takes. Because tolerated risk falls with consumption, the median consumer is decisive: whatever she accepts, everyone poorer than her accepts too. However, empirically consumption distributions are skewed to the right, so the median sits below the mean. This implies that the risk tolerance under majoritarian voting is higher than under perfect equality. The resulting pro-AI coalition at the highest tolerated risk is made up of the poorer half of society, while agents richer than the median oppose development. Redistribution that raises the median reduces risk taking, so that redistribution becomes a tool to manage existential risk.

The model allows us to estimate how inequality in present-day societies affects the willingness to develop AI and risk extinction. I calibrate the model economy using the US consumption distribution. With a relative risk aversion of two, the decisive median accepts a 5.68% probability of extinction against the 4.88% accepted under perfect equality, a one sixth increase. Over-risking of the median relative to the mean increases in the curvature of utility, but is almost invariant to the size of the growth prize.¹

¹Following Jones (2024), the baseline calibration assumes that AI can raise GDP per capita growth from 2% to 10% a year, an optimistic estimate. If AI can deliver a smaller increase, the risk society is willing to take falls, but

The model also delivers additional insights.

- Supermajority rules are a partial substitute for redistribution: they replace the median with a richer decisive voter, and a threshold at the quantile of the mean implements the risk a society with perfect equality would take, without any transfer;
- Inequality affects risk taking beyond majoritarian voting: a utilitarian planner who cannot redistribute may take even more risk than the median consumer, because concern for the poor makes the consumption gains more valuable;
- If AI also widens inequality, the decisive voter becomes more cautious, since she is cut out of the gains: the pro-AI coalition can evolve from the poorer half of society to an interior band of the distribution, and over-risking can turn into under-risking;
- Existential risk reshapes the politics of redistribution: the rich, who have the most life to lose, may back it as self-protection, and support for it can be U-shaped in consumption, with the richest siding with the poorest;
- Where AI can also protect against other existential threats, climate change, asteroid impacts, pandemics, the rich gain an additional motive to develop it, and an unequal society under-develops it: the median voter, poorer than the average, pays less for protection than she would under perfect equality;

AI may also lengthen life, a benefit that Jones (2024) shows to matter greatly, since mortality improvements are in the same units as extinction risk. I show that when longevity gains are common across all individuals, they reduce over-risk. A longevity gain is worth the same extinction risk to every agent, whatever her consumption: both the gain and the risk are measured in years of life, so the value of a year of life cancels. If AI also lengthens life, it raises everyone's tolerated risk by the same amount; this in turn shrinks the wedge that inequality opens. At the baseline calibration, halving mortality lifts the tolerated risk of both the median and the mean to around 29% and cuts over-risking from a sixth to about 2%. At the same time, if AI lengthened the lives of the rich more than those of the poor, the shape of the pro-AI coalition could change, with the rich favouring it.

Related literature. This paper contributes to three strands of literature. The first is the literature on transformative technologies and life risk. Jones (2016b) shows that, as societies grow richer, it is optimal to redirect innovation toward saving lives; Jones (2024) brings this logic to AI, trading its growth prize against the risk of extinction, developing the theoretical framework on not the proportional over-risking due to inequality.

which this paper builds; Acemoglu and Lensman (2024) study when transformative technologies with large, partly irreversible costs should be adopted. A literature on existential risk and growth links the hazard to technology development itself, so that a society may pass through a “time of perils”, before safer times arrive (Aschenbrenner, 2020; Trammell and Aschenbrenner, 2026). Martin and Pindyck (2015) use the value of a statistical life to value the gains from averting catastrophes, and show that when a society faces multiple threats, it might be optimal to avert none. What all these papers share is a framework with a representative consumer. This paper instead makes society a distribution of agents who must reach a collective decision, and studies how inequality affects how much extinction risk it accepts.

The second strand studies the value of life and mortality risk across people. The mechanism of the paper runs through the value of life (Rosen, 1988), whose elasticity to consumption long-run and health-spending estimates place above one (Hall and Jones, 2007; Murphy and Topel, 2006). How the value of life should enter social choices over mortality is examined by Adler et al. (2014) and Fleurbaey (2010), building on Harsanyi (1955), with the preferences of Bommier et al. (2015) and Bommier et al. (2017) offering an alternative framework; I draw on these for the welfare comparisons and the constrained-planner benchmarks.

The third strand is the literature on the political economy of redistribution and risk. The over-risking result is a median-voter theorem (Meltzer and Richard, 1981; Persson and Tabellini, 2000); the finding that existential risk can turn the rich into supporters of redistribution speaks to the Meltzer–Richard tradition and to the insurance-based reversal (Moene and Wallerstein, 2001).

Finally, an empirical literature documents that inequality and low relative standing raise individual risk-taking (Payne et al., 2017; Kuziemko et al., 2014); the channel proposed by this paper is distinct, since the source of heterogeneity is the value of life rather than behaviour, but points in the same direction at the level of collective choice.

Outline. The paper proceeds as follows. Section 2 presents the model and shows that tolerated existential risk falls with consumption whenever the value of life rises with it. Section 3 introduces inequality, derives the over-risking result under majoritarian voting and under a planner who cannot redistribute, quantifies it for the US, and studies supermajority rules, the case in which AI itself increases inequality, and the politics of redistribution under existential risk. Section 4 studies AI as a life-saving technology, through longer lives and through protection against other existential risks. Section 5 concludes. All proofs are in Appendix A.

2 The value of life and existential risk tolerance

This section describes a framework to model AI development under existential risk along the lines of Jones (2024), and studies how tolerated risk changes with the level of *status quo* (i.e. no AI development) consumption.

The economy is populated by expected-utility agents who derive utility from being alive and consuming, through a flow utility u , differentiable with $u' > 0$. An agent who dies receives a utility normalised to zero, so that she prefers life to death at consumption c if $u(c) > 0$.² Agents discount the future at rate ρ and face a constant mortality hazard μ : an agent survives a further a years with probability $e^{-\mu a}$ and, as in the perpetual-youth model of Blanchard (1985), her remaining lifetime looks the same at every age. Consumption grows at rate g , so an agent who consumes c today consumes ce^{ga} after a years. Her expected remaining lifetime utility is

$$U(c, g) \equiv \int_0^\infty e^{-(\rho+\mu)a} u(ce^{ga}) da. \quad (1)$$

Without AI development, at time zero agents consume c_0 and consumption grows at rate g_0 . Assume that $U(c_0, g_0) > 0$ so that in the *status quo* life is preferred to death.

Agents face a choice: develop AI or not. Developing AI raises growth from g_0 to $g_{ai} > g_0$, but implies a one-time extinction probability $\delta \in [0, 1]$: with probability $1 - \delta$ consumption grows at g_{ai} from then on, and with probability δ every agent dies. An agent accepts the gamble if and only if the expected utility of accepting exceeds the utility of declining, $(1 - \delta) U(c_0, g_{ai}) \geq U(c_0, g_0)$. Since agents are identical they agree, and AI is developed if and only if the risk is low enough,

$$\delta \leq \delta^* \equiv 1 - \frac{U(c_0, g_0)}{U(c_0, g_{ai})}, \quad (2)$$

where the cut-off δ^* lies in $[0, 1]$ because $0 < U(c_0, g_0) \leq U(c_0, g_{ai})$, the AI path being richer at every date.

Unlike most conditions in economics, the cut-off is a ratio of utility levels, not of marginal utilities. Agents weigh a lifetime against the outside option of death, whose utility is zero, and what matters is how much a lifetime is worth, not how much one more unit of consumption is.

How does the cut-off change with *status quo* consumption? The answer depends on whether agents value being alive more when they consume. The value of a year of life, in years

²Can death ever be preferable to life for an agent who is alive at equilibrium? This is a somewhat philosophical question that deserves further exploration in this context. A well known fact is that when utility from consumption and life are additively separable, isoelastic utilities can only satisfy one of the following two: *i*) life is preferable to death at any consumption level, or *ii*) utility flow from consumption is bounded from above and the value of life is increasing in standards of living.

of consumption, at consumption level c is

$$v(c) \equiv \frac{u(c)}{u'(c)c}.$$

The definition follows Rosen (1988): $u(c)$ is the utility of living a year consuming c ; dividing by the marginal utility $u'(c)$ expresses it in units of consumption; dividing again by c makes it a pure number, the number of years of consumption a year of life is worth. Along a growth path the relevant object is the value of the remaining life as a multiple of the consumption stream,

$$v(c, g) \equiv \frac{U(c, g)}{\partial U(c, g) / \partial \ln c} = \frac{\int_0^\infty e^{-(\rho+\mu)a} u(c e^{ga}) da}{\int_0^\infty e^{-(\rho+\mu)a} u'(c e^{ga}) c e^{ga} da}, \quad (3)$$

lifetime utility divided by the utility of raising the whole path by one per cent. It is a weighted average of the yearly value $v(c e^{ga})$ along the path, with the marginal utility of consumption at each date as the weight, and at $g = 0$ it reduces to the Rosen definition, $v(c, 0) = v(c)$.

Tolerated risk falls with consumption whenever the value of life rises with it. The next proposition makes the statement precise for any flow utility.

Proposition 1. *Let $g_{ai} > g_0 > 0$, let $U(c, g)$ and $\partial U(c, g) / \partial \ln c$ be finite on both paths, $U(c, g_0) > 0$, and let v be eventually monotone or bounded on $[c, \infty)$. Then:*

- (i) $\text{sign } \frac{d\delta^*}{dc} = \text{sign } [v(c, g_0) - v(c, g_{ai})];$
- (ii) *if $v' \geq 0$ above c , with strict inequality on a set of positive measure, δ^* is strictly decreasing in c ; if $v' \leq 0$ (strictly somewhere), strictly increasing; if v is constant, δ^* does not depend on c .*

Whether the value of life rises with consumption is an empirical question, and the answer lies in the elasticity of the value of a statistical life to consumption. The value of a statistical life is $VSL(c) = u(c) / u'(c) = c v(c)$, so its consumption elasticity is $\varepsilon_{VSL} = 1 + d \ln v / d \ln c$, and v rises with consumption if and only if $\varepsilon_{VSL} > 1$. Long-run and health-spending estimates put the elasticity above one (Hall and Jones, 2007; Murphy and Topel, 2006); within-country estimates from labour markets tend to put it below. The paper sits with the former, and specialises utility to constant relative risk aversion with a curvature that makes the value of life increasing.

In the remainder of the paper, utility has constant relative risk aversion, with an additive constant that captures the value of being alive,

$$u(c) = \begin{cases} \bar{u} + \frac{c^{1-\gamma}}{1-\gamma} & \gamma \neq 1 \\ \bar{u} + \log c & \gamma = 1 \end{cases}$$

for which the value of life in consumption units is

$$v(c) = \begin{cases} \bar{u}c^{\gamma-1} + \frac{1}{1-\gamma} & \gamma \neq 1 \\ \bar{u} + \log c & \gamma = 1 \end{cases}$$

Assumption 1 (CRRA utility with an increasing value of life). Utility is CRRA with $\gamma \geq 1$ and $\bar{u} > 0$, so that $v'(c) > 0$ for all c .

The restriction $\gamma \geq 1$ is the same Jones (2024) imposes. With $\bar{u} > 0$ it makes the value of life rise with consumption, $v'(c) = (\gamma - 1)\bar{u}c^{\gamma-2}$ for $\gamma > 1$ and $v'(c) = 1/c$ for $\gamma = 1$, and it makes lifetime utility finite on every path, since $\rho + \mu + (\gamma - 1)g \geq \rho + \mu > 0$.³ A richer society then accepts less extinction risk from AI: Proposition 1(ii) gives $d\delta^*/dc < 0$, under Assumption 1 and indeed for any flow utility whose value of life rises with consumption.⁴ This gradient is the result on which the rest of the paper builds. Along a growth path the value of life is, by (3),

$$v(c, g) = \begin{cases} \bar{u}c^{\gamma-1} \frac{\rho + \mu + (\gamma - 1)g}{\rho + \mu} + \frac{1}{1-\gamma} & \gamma \neq 1 \\ \bar{u} + \log c + \frac{g}{\rho + \mu} & \gamma = 1 \end{cases} \quad (4)$$

the same function of c as v with the constant $\bar{u}$ scaled up by growth, and it satisfies $c \partial_c v(c, g) = 1 + (\gamma - 1)v(c, g)$, as v does. The tolerated risk cut-off then has a closed form, derived in Appendix A,

$$\delta^* = \frac{g_{ai} - g_0}{(\rho + \mu + (\gamma - 1)g_0) v(c, g_{ai})} : \quad (5)$$

tolerated risk is the growth gain, times the horizon over which the status-quo path is enjoyed, divided by the value of life on the AI path. The cut-off decreases in c as the value of life $v(c, g)$ increases in consumption under $\gamma \geq 1$.

Appendix A also writes the cut-off and its slope in terms of the yearly value of life $v(c)$, equations (10) and (11). As $c \rightarrow \infty$, $\delta^* \rightarrow 0$ with an elasticity that converges to $-(\gamma - 1)$: at $\gamma = 2$ tolerated risk (asymptotically) halves with every doubling of consumption; with log utility the elasticity tends to zero. Higher consumption lowers risk-taking, and it does so faster when the curvature of utility is higher.

³More in general, any utility function for which relative risk aversion $\gamma(c) \equiv -c u''(c)/u'(c)$ is greater than or equal to one exhibits increasing value of life, since $c v'(c) = 1 + v(c)[\gamma(c) - 1]$.

⁴The proposition takes $g_0 > 0$. If $g_0 = 0$, $v(c, 0) = v(c)$ lies below $v(c, g_{ai})$, a weighted average of v above c , and part (i) gives the same sign.

3 Voting on AI development with inequality

Section 2 studied AI development under existential risk with perfect equality, where every agent coincides with the average one. There is then no political process: all agents agree on whether to develop AI. However, the analysis showed that tolerated risk changes with *status quo* consumption, and we know that in reality consumption is unequally distributed. This implies that the decision on whether to develop AI is a political process in which preferences are aggregated. This section studies how inequality affects the choice of whether to develop AI once it is taken by majoritarian voting.

Unequal societies and the median consumer. Consider a society populated by a measure one of agents, indexed by $i \in [0, 1]$, who share the utility function u . Each agent consumes $c_i = \omega_i \bar{c}$, where $\int_0^1 \omega_i di = 1$ so that aggregate and average consumption equal $\bar{c}$. Let c_{med} be median consumption. While the distribution of consumption is unequal, AI gains are shared proportionally: conditional on survival, every agent's consumption grows at g_{ai} instead of g_0 forever, so that relative positions are preserved at every date. Developing AI is still a gamble that involves extinction: if AI backfires, all agents die.

Asked whether to develop AI, each agent expresses a cut-off $\delta_i^* = \delta^*(c_i) = 1 - U(c_i, g_0) / U(c_i, g_{\text{ai}})$, which is the maximum extinction probability she is willing to accept. Since the value of life is increasing in consumption, $\delta^*(c_i)$ is decreasing in c_i (or ω_i) by Proposition 1, so richer agents value life more and are willing to risk less.

The simplest democratic procedure puts AI development to a vote: AI is developed if and only if a majority votes in favour, that is, if and only if its risk δ is below the median of the individual cut-offs,

$$\delta \leq \text{med}_i \delta_i^* .$$

Since the individual cut-off falls with consumption, the median consumer is decisive: whatever she accepts, everyone poorer accepts too, and whatever she rejects, everyone richer rejects too. The decision is again a threshold rule, with the median's cut-off as the threshold,

$$\delta \leq \delta^*(c_{\text{med}}) .$$

When consumption is unequally distributed but AI benefits all agents proportionally, majoritarian voting develops AI if and only if the probability of extinction is below the median consumer's cut-off. Inequality therefore matters for whether AI is developed, and so does redistribution carried out before the vote: whatever raises median consumption lowers the risk society accepts. The median is also the only feature of the distribution that matters in this setting. Empirically, income and wealth distributions are well approximated by a log-normal with

a Pareto tail at the top; but since the median voter is decisive, the choice is robust to the shape of inequality at the top.

There is a natural benchmark to which the outcome of the vote can be compared. A utilitarian planner who can redistribute, given CRRA utility, chooses perfect equality, $c_i = \bar{c}$, and accepts at most $\delta^*(\bar{c})$. Consumption and wealth distributions are skewed to the right, so $c_{\text{med}} < \bar{c}$ and $\delta^*(c_{\text{med}}) > \delta^*(\bar{c})$: the risk society takes under perfect equality is lower than the risk it takes under majoritarian voting.

Proposition 2. *For any right-skewed distribution, the median voter tolerates strictly more extinction risk than the redistributive first best. The over-risking depends on the distribution only through the ratio of median to mean consumption, and rises as that ratio falls.*

The logic is immediate: the decisive voter is poorer than the average, her value of life is lower, and she gambles where a planner evaluating the gamble at the mean would not. The gap depends on the distance between median and mean. Hence redistribution, by moving the median, can affect how much existential risk a society accepts.

A constrained utilitarian planner. It is natural to ask whether inequality affecting how much risk a society takes is a result specific to majoritarian voting. It turns out it is not: a similar, and potentially stronger, distortion arises for a benevolent planner who lacks the instruments to redistribute consumption. A utilitarian planner who cannot redistribute also follows a cut-off rule, evaluated at the *equally distributed equivalent* consumption (Atkinson, 1970), the level that, consumed by all agents, delivers the same welfare,

$$U(c_{EDE}, g) = \int_0^1 U(c_i, g) di ,$$

on either path; under CRRA utility this is the Atkinson equivalent $c_{EDE} = (\int_0^1 c_i^{1-\gamma} di)^{1/(1-\gamma)}$, and AI is developed by the planner if and only if

$$\delta \leq \delta^*(c_{EDE}) .$$

By Jensen's inequality, $c_{EDE} < \bar{c}$ for any concave utility, so the constrained planner accepts more risk than the redistributive first best. Whether c_{EDE} is larger or smaller than c_{med} depends in general on the shape of the consumption distribution and on the curvature of utility; however, it may well be that the equally distributed equivalent is lower than the median. For a log-normal distribution⁵ with dispersion σ of log consumption and with CRRA utility,

⁵This distribution fits the distributions of wealth and consumption in the data well, apart from the top, and is used in the calibration exercise below.

$c_{EDE} = \bar{c} e^{-\gamma\sigma^2/2}$; hence, for $\gamma > 1$, $c_{EDE} < c_{\text{med}} = \bar{c} e^{-\sigma^2/2}$, and the planner who cannot redistribute takes more risk than the median voter, who takes more than the redistributive first best, $\delta^*(c_{EDE}) > \delta^*(c_{\text{med}}) > \delta^*(\bar{c})$.

Proposition 3. *A benevolent utilitarian planner who cannot redistribute takes the gamble and develops AI using a threshold rule involving the equally distributed equivalent consumption.*

The constrained planner generally over-risks compared to the redistributive first best, and under a log-normal consumption distribution and CRRA utility over-risks even more than the median consumer.

Redistribution is thus a tool of existential-risk policy for planners and voters alike.

US calibration. How large is the wedge at empirically relevant inequality? I use the calibration of Jones (2024, Section II.B): $\rho = 1\%$, $\mu = 1\%$ (a life expectancy of 100 years), growth of $g_0 = 2\%$ a year without AI, the long-run average of the US over the last 150 years (Jones, 2016a), and of $g_{\text{ai}} = 10\%$ a year with it, his fast-growth scenario; $\bar{u}$ is pinned by a value of a life-year of six times per-capita consumption at the mean, $\nu(\bar{c}) = 6$, the ratio of a \$250,000 life-year (a \$10 million VSL spread over some 40 years) to about \$40,000 of per-capita consumption (Jones, 2024, p. 578). Along the status-quo path the value of life is then $\nu(\bar{c}, g_0) = 13$ at $\gamma = 2$, and on the AI path $\nu(\bar{c}, g_{\text{ai}}) = 41$, so by (5) the cut-off at the mean is $\delta^*(\bar{c}) = 0.08 \times 25/41 = 4.88\%$: an agent at mean consumption accepts a one-in-twenty chance of extinction for growth of 10% a year. For tractability the distribution is log-normal with dispersion σ , which gives closed forms, $c_{\text{med}}/\bar{c} = e^{-\sigma^2/2}$ and a Gini coefficient $G(\sigma) = 2\Phi(\sigma/\sqrt{2}) - 1$, where Φ is the CDF of the standard normal; the over-risking results themselves depend on the distribution only through $c_{\text{med}}/\bar{c}$. There are alternative empirical targets for inequality, in consumption, income or wealth. To be conservative, I consider inequality between the consumption Gini, around 0.30 (Fisher et al., 2015), and the income Gini after taxes and transfers, about 0.42 (Congressional Budget Office, 2023).

Table 1 reports the cut-offs of the mean, the median and the equally-distributed equivalent, and the over-risking of the median consumer relative to the first best, in per cent of the first-best risk,

$$\mathcal{O} = \frac{\delta^*(c_{\text{med}}) - \delta^*(\bar{c})}{\delta^*(\bar{c})} \times 100,$$

for different values of the Gini coefficient, at γ of 1.01, 2 and 3 as in Jones (2024). Over-risking rises with the curvature of utility. At $\gamma = 2$ and the consumption Gini, the median accepts 5.68% against the first best's 4.88%, an over-risking of 16.5%; at the income Gini the wedge is 37%. The wedge is almost invariant to the size of the growth prize: as $g_{\text{ai}} \rightarrow \infty$, the singularity of Jones (2024), the over-risking ratio converges to $(\bar{c}/c_{\text{med}})^{\gamma-1} - 1$, which is 16.0% at $\gamma = 2$ and 34.6%

at $\gamma = 3$ for the consumption Gini, against 16.5% and 34.9% at $g_{ai} = 10\%$.⁶ Figure 2 illustrates the mechanism at $\gamma = 2$ and the consumption Gini: the cut-off falls with consumption, so the median consumer, to the left of the mean, tolerates more risk than the first best, and the shaded gap between the two cut-offs is the over-risking.

Table 1: Tolerated risk and over-risking by curvature and inequality

γ	Gini	tolerated risk δ^* (%)			over-risking $\mathcal{O}$ (%)	
		mean	median	EDE	median	EDE
1.01	0.30	35.05	35.57	35.57	1.5	1.5
1.01	0.34	35.05	35.73	35.74	1.9	2.0
1.01	0.42	35.05	36.14	36.15	3.1	3.1
2	0.30	4.88	5.68	6.62	16.5	35.7
2	0.34	4.88	5.95	7.27	22.0	49.0
2	0.42	4.88	6.68	9.19	37.0	88.4
3	0.30	1.88	2.53	4.62	34.9	146.2
3	0.34	1.88	2.77	6.09	47.8	224.4
3	0.42	1.88	3.49	12.25	85.6	552.3

Notes. Cut-offs of the mean, the redistributive first best; of the median, majoritarian voting; and of the equally-distributed equivalent, the planner who cannot redistribute, with the over-risking of the last two relative to the first best, $\mathcal{O} = \delta^*(\cdot)/\delta^*(\bar{c}) - 1$, in per cent. AI raises growth from $g_0 = 2\%$ to $g_{ai} = 10\%$ a year; $\rho = \mu = 1\%$, $v(\bar{c}) = 6$, log-normal consumption.

The mechanism does not depend on log-normality: Table 2 reports the same over-risking as a function only of the median/mean consumption ratio, so any right-skewed distribution with the same ratio delivers the same wedge.

Table 2: Over-risking and the ratio of median to mean consumption

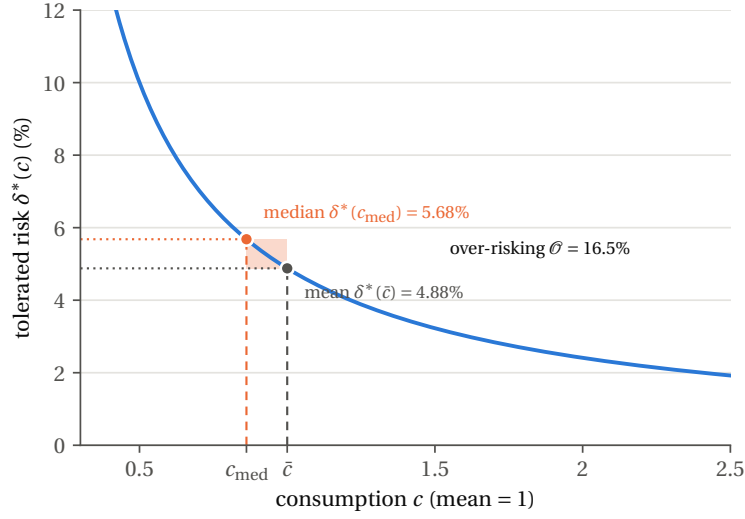
γ	0.95	0.90	$c_{med}/\bar{c} = 0.862$	0.80	0.75	0.70
1.01	0.5	1.0	1.5	2.2	2.9	3.6
2	5.4	11.4	16.5	25.8	34.4	44.4
3	10.9	23.7	34.9	56.9	78.8	105.6

Notes. Over-risking $\mathcal{O}$ of the median, in per cent of the first-best risk, as a function of the ratio of median to mean consumption only; 0.862 is the ratio of a log-normal with Gini 0.30. Any distribution with the same ratio gives the same wedge. $g_{ai} = 10\%$, $\rho = \mu = 1\%$, $g_0 = 2\%$, $v(\bar{c}) = 6$.

From the simulations, a clear trade-off emerges between how much risk society accepts and the gap between the majoritarian and first best cut-offs, governed by utility curvature, as Fig-

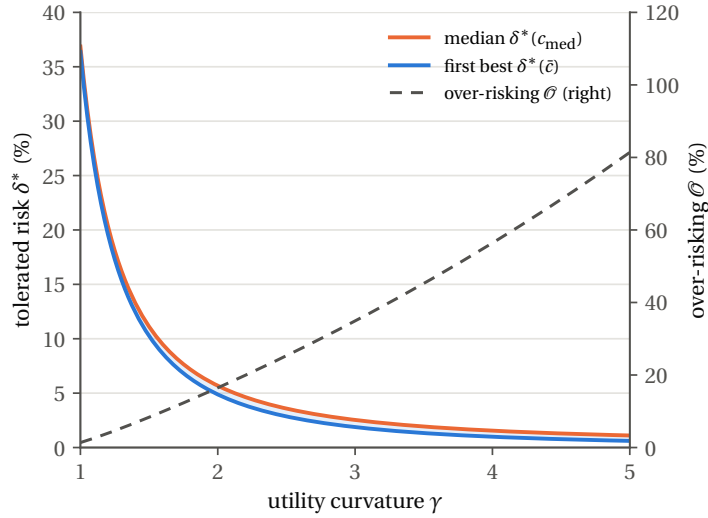
⁶As $g_{ai} \rightarrow \infty$ in (10), $\delta^*(c) \rightarrow \frac{\rho+\mu}{\rho+\mu+(\gamma-1)g_0} \cdot \frac{1}{(\gamma-1)v(c)+1}$, the individual counterpart of eq. (7) in Jones (2024). Since $(\gamma-1)v(c)+1 = (\gamma-1)\bar{u}c^{\gamma-1}$, the constant $\bar{u}$ cancels in the ratio of two cut-offs, which equals $(\bar{c}/c_{med})^{\gamma-1}$: everyone's prize is the same bound $\bar{u}/(\rho+\mu)$, and inequality bites through what each agent has to lose.

Figure 2: The over-risking mechanism



Tolerated risk $\delta^*(c)$ as a function of consumption, with mean consumption normalised to one. The cut-off falls with consumption, so the median consumer, who is poorer than the mean, tolerates more risk than the redistributive first best, 5.68% against 4.88%; the shaded gap is the over-risking, 16.5% of the first-best risk. Calibration: $\gamma = 2$, $v(\bar{c}) = 6$, $\rho = \mu = 1\%$, $g_0 = 2\%$, $g_{ai} = 10\%$, log-normal consumption with Gini 0.30.

Figure 3: The curvature trade-off



Tolerated risk of the median consumer and of the redistributive first best, left axis, and over-risking, right axis and dashed line, as functions of the curvature of utility γ . Both cut-offs fall as γ rises, so society grows more cautious, but the first best falls faster and the over-risking increases, from 1.5% at $\gamma = 1.01$ to 34.9% at $\gamma = 3$. Calibration: $v(\bar{c}) = 6$, $\rho = \mu = 1\%$, $g_0 = 2\%$, $g_{ai} = 10\%$, log-normal consumption with Gini 0.30.

ure 3 shows. As γ rises the value of life steepens, so both the decisive median and the mean grow more cautious in absolute levels; but the mean's tolerated risk falls faster than the median's, so over-risking $\mathcal{O}$ increases. In sum, at $\gamma \approx 1$ society gambles freely and inequality barely distorts the choice; at higher γ , maximum tolerated risk is lower but the decisive median tolerates far more risk than the redistributive optimum, so redistribution can have a stronger effect on risk taking.

Alternative voting rules. Simple majority is only one voting procedure to aggregate preferences, and for high-stakes choices a qualified majority is often proposed instead. Suppose AI is developed only if a fraction $q \in [0, 1]$ of the population votes for it. Because the individual cut-off δ^* is decreasing in consumption, the agents in favour at risk δ are the poorer ones, a population fraction $F(c)$ with $\delta^*(c) \geq \delta$. The threshold is then the cut-off of the voter at the q -quantile of consumption.

Proposition 4. *Under a rule that develops AI only if a fraction $q \in [0, 1]$ votes in favour, AI is developed if and only if $\delta \leq \delta^*(c_q)$, where c_q is the q -quantile of consumption. The tolerated risk $\delta^*(c_q)$ is decreasing in q , with $q = \frac{1}{2}$ recovering the median rule; a higher threshold makes the decisive voter richer and society more cautious.*

A qualified majority is therefore a substitute for redistribution: both raise the consumption of the decisive voter, redistribution by lifting c_{med} toward $\bar{c}$, a supermajority by replacing the median with a richer c_q , and both lower the existential risk society accepts. Setting $q = F(\bar{c})$, the share of the population below the mean, makes $c_q = \bar{c}$ and implements the first-best risk $\delta^*(\bar{c})$ without any transfer; with a right-skewed distribution this requires $q > 1/2$, and at the calibration $q = 0.61$.

Giving each agent votes in proportion to her consumption works in the same way. AI is then developed if and only if the agents in favour, the poorer ones, hold at least half of total consumption, so the decisive voter is the agent below whom half of consumption lies, and she is richer than the median whenever consumption is unequal.⁷ For the log-normal she consumes $\bar{c} e^{\sigma^2/2}$, more than the mean, so weighting votes by consumption goes past the first best: at the calibration the decisive voter sits at the 0.71 quantile and tolerates 4.19%, against the first best's 4.88% and the median's 5.68%.

⁷The agents in favour at risk δ are those with $\delta^*(c_i) \geq \delta$, a population share $F(c)$ holding the consumption share $\int_{\{i: c_i \leq c\}} c_i \, di / \bar{c}$. AI is developed if and only if this share is at least one half, that is, if and only if $F(c) \geq q$ with q the population share that holds the bottom half of consumption: a supermajority rule. Under inequality the bottom p of the population holds less than a share p of consumption, because its average consumption is below $\bar{c}$, so $q > 1/2$. For the log-normal the bottom p holds the share $\Phi(\Phi^{-1}(p) - \sigma)$, so $q = \Phi(\sigma)$ and $c_q = c_{\text{med}} e^{\sigma \Phi^{-1}(q)} = c_{\text{med}} e^{\sigma^2} = \bar{c} e^{\sigma^2/2}$; at Gini 0.30, $\sigma = 0.545$, $q = 0.71$ and $c_q = 1.16 \bar{c}$.

Ex ante, ex post and other welfare criteria. What if welfare is not utilitarian? Moving beyond utilitarianism raises a further issue: whether the welfare being maximised is evaluated before the extinction gamble is realised, ex ante, or after, ex post. An expected-welfare-maximising utilitarian planner is the only case in which the distinction does not arise, because welfare is linear in utilities. Adding up each agent’s expected utility before the gamble is realised, ex ante, or summing the utilities realised in each state and averaging across states, ex post, give the same value: developing AI is worth $\int (1 - \delta)U(c_i, g_{ai}) di = (1 - \delta) \int U(c_i, g_{ai}) di$ either way, extinction contributing zero welfare in both cases, so the planner uses the unique cut-off $\delta^*(c_{EDE})$. Whether welfare is computed ex ante or ex post matters, however, for criteria that are not linear in utilities.

A Rawlsian planner who maximises the utility of the worst-off and cannot redistribute is a case in point. Ex ante, she is only concerned with the poorest consumer, whose tolerated risk $\delta^*(c_{min})$ exceeds even the equally-distributed equivalent’s. Ex post, she considers the worst-off position once the gamble is realised, and in any gamble with positive extinction risk that position is death, so she never develops AI as long as the poorest has a positive value of life in the status quo. The Rawlsian planner thus spans the whole range, from over-risking to outright rejection, and the gap between the two readings is the ex ante/ex post divide in the aggregation of mortality risk (Adler et al., 2014; Fleurbaey, 2010). Prioritarian criteria, which weight the worse-off more but are not concerned with them alone, fall in between: at the calibration, a planner’s ex ante over-risking rises with her inequality aversion, from the utilitarian benchmark $\delta^*(c_{EDE})$ toward the Rawlsian $\delta^*(c_{min})$.⁸

Inequality-increasing AI. Until now I modelled AI as a “proportional” gamble: conditional on survival, the consumption distribution is the same with and without AI, and the growth rate of consumption is the same for all agents. A distinctive feature of AI, however, is that the technology carrying the risk may also reshape the distributions of consumption, income and wealth, through skill- and capital-biased gains. This feature can be tractably modelled in this framework by allowing, at adoption, the consumption distribution to become more unequal, and then letting the growth rate of consumption be the same for all consumers.⁹ When the distribu-

⁸A standard way of parametrising inequality aversion with lifetime utility is to apply the Atkinson transform to lifetime utility itself, so that welfare is $\int U(c_i, g)^{1-\varepsilon}/(1-\varepsilon) di$, with $\int \log U(c_i, g) di$ when $\varepsilon = 1$ (Adler, 2012, ch. 5); Adler et al. (2021) apply it to lifetime utility under mortality risk. An aversion $\varepsilon = 0$ gives the utilitarian criterion, and $\varepsilon \rightarrow \infty$ the Rawlsian one. Ex ante, AI is developed if and only if $\int [(1 - \delta)U(c_i, g_{ai})]^{1-\varepsilon} di / (1 - \varepsilon) \geq \int U(c_i, g_0)^{1-\varepsilon} di / (1 - \varepsilon)$, so the cut-off is $1 - [\int U(c_i, g_0)^{1-\varepsilon} di / \int U(c_i, g_{ai})^{1-\varepsilon} di]^{1/(1-\varepsilon)}$, one minus the ratio of the power means of order $1 - \varepsilon$ of status-quo and AI-path lifetime utilities. At $\gamma = 2$ and Gini 0.30 it rises from 6.6% at $\varepsilon = 0$ to 7.0% at $\varepsilon = 2$ and 9.6% at $\varepsilon = 5$, and tends to the Rawlsian cut-off as $\varepsilon \rightarrow \infty$.

⁹An alternative would be to allow growth rates themselves to differ between individuals, but since we can compute utility flows in closed form, in the setting of this paper the two modelling strategies are equivalent.

tion is log-normal, we can capture increasing inequality by increasing its dispersion to $\sigma' > \sigma$, preserving ranks and leaving the mean level unchanged; thereafter all agents grow at g_{ai} .

In this setting, the widening leaves the median consumer's status quo unaffected, $c_{med} = \bar{c} e^{-\sigma^2/2}$, but lowers the level from which she grows after AI is developed to $\bar{c} e^{-\sigma'^2/2}$. Her tolerated risk, $1 - U(c_{med}, g_0)/U(\bar{c} e^{-\sigma'^2/2}, g_{ai})$, therefore falls as the gains concentrate. She is strictly harmed, and rejects AI at any risk, if and only if

$$\sigma'^2 > \sigma^2 + \frac{2}{\gamma - 1} \ln \frac{\rho + \mu + (\gamma - 1)g_{ai}}{\rho + \mu + (\gamma - 1)g_0} \quad (\gamma > 1), \quad \sigma'^2 > \sigma^2 + \frac{2(g_{ai} - g_0)}{\rho + \mu} \quad (\gamma = 1). \quad (6)$$

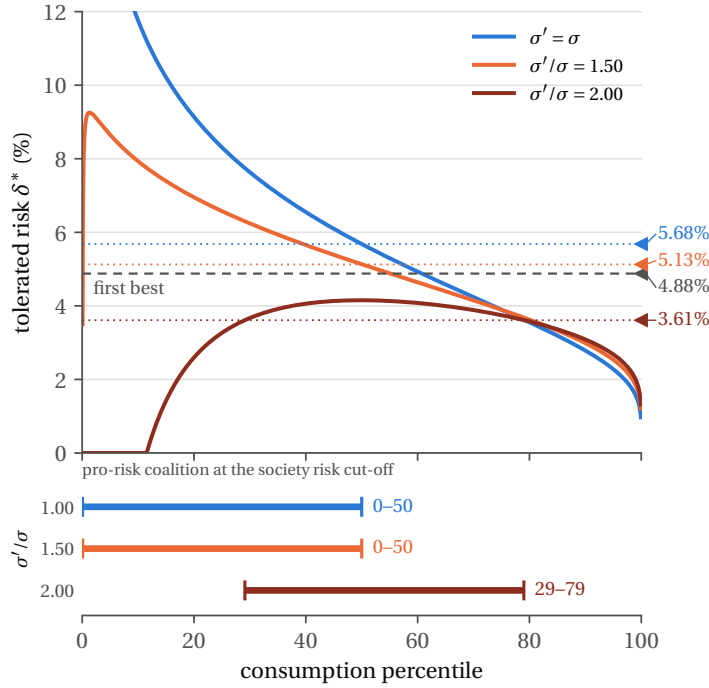
At $\gamma = 2$ and an initial Gini of 0.30, she is strictly harmed beyond a post-gamble Gini $G' \approx 0.74$, where $\sigma'/\sigma \approx 2.90$. Growth of 10% a year is worth to her a level gain of $(\rho + \mu + g_{ai})/(\rho + \mu + g_0) = 3$, so it takes a strong concentration of the gains to cut her out of them and make her reject AI at arbitrarily low risk.

At the same time, the first best is unaffected, since the planner redistributes the additional inequality away, so the median's behaviour translates directly into over- and under-risking: her over-risking falls with the widening and turns into under-risking above a finite dispersion, at the calibration a post-gamble Gini $G' \approx 0.48$, where $\sigma'/\sigma \approx 1.66$. The median consumer over-risks for a moderate widening and becomes a check beyond it. The two faces of inequality therefore enter with opposite signs: the median votes on her own status quo against her own prize. Inequality in the endowment lowers her status quo and makes her bolder; inequality in the gains lowers her prize and makes her more cautious. The striking implication is that a sufficiently regressive AI windfall turns the excluded median into a brake, not because she is wiser, but because she has been cut out of the upside.

The result for the median, however, is not yet a result for society. An AI that widens inequality undermines the argument that made the median consumer decisive, that the cut-off falls with consumption and hence with rank: richer agents still have a higher value of life, but they now also gain more in proportion, as the level from which they grow rises relative to their status quo. For a moderate widening the cut-off δ^* remains monotone in consumption; for larger widenings it becomes hump-shaped, with a single peak that climbs the distribution as the gains concentrate (Lemma 1 in Appendix A).

Without monotonicity, the maximum probability of extinction society accepts is the median of the individual thresholds, $\text{med}_i \delta_i^* = \text{med}_i \delta^*(c_i)$, and AI is developed whenever the risk falls below it. Under a strong enough widening this is the median of a hump-shaped profile, and it no longer coincides with the cut-off of the median consumer. Since the profile has at most one peak, the pro-risk coalition at $\text{med}_i \delta^*(c_i)$ has both tails opposed: the richest, because their value of life is already very high, and, once the widening is severe enough, a tail of the poor for

Figure 4: The cut-off under inequality-increasing AI



Tolerated extinction risk δ^* of each agent against her consumption percentile when AI widens the distribution at adoption to a dispersion σ' , preserving ranks and the mean level, with growth at g_{ai} thereafter; the dashed line is the redistributive first best $\delta^*(\bar{c})$, and the dotted lines and the right axis mark the risk each society tolerates, the median of its cut-offs, next to the first best. A mild widening leaves the cut-off decreasing; from $\sigma'/\sigma \approx 1.40$ it is hump-shaped. Society's cut-off is the median of the cut-offs, which falls below the first best near $\sigma'/\sigma \approx 1.66$, where society shifts from over- to under-risking. The bars below mark the pro-risk coalition in each scenario, the agents whose cut-off is at or above society's cut-off: the bottom half of the distribution up to $\sigma'/\sigma \approx 1.5$, and an interior band beyond, as the poorest are cut out of the upside. Calibration: $\gamma = 2$, $\nu(\bar{c}) = 6$, $\rho = \mu = 1\%$, $g_0 = 2\%$, $g_{ai} = 10\%$, log-normal consumption with Gini 0.30.

whom the prize does not justify the risk.

Figure 4 provides an illustration. At $\sigma' = \sigma$ the cut-off adopted by society is that of the median consumer who, at $\gamma = 2$ and $G = 0.30$, over-risks by the 16.5% of Table 1; over-risking is continuous and decreasing in σ' : it survives moderate widening but turns into under-risking once the gains are concentrated enough (beyond $\sigma'/\sigma \approx 1.66$, a post-gamble Gini $G' \approx 0.48$). The profile of cut-offs becomes hump-shaped from $\sigma'/\sigma \approx 1.40$, but the hump sits in the poor tail, and the median of the cut-offs coincides with the median consumer's until about $\sigma'/\sigma \approx 1.5$; at $\sigma'/\sigma = 2$ the decisive cut-off is 3.61%, against the median consumer's 4.15% and the first best's 4.88%. Beyond that point society risks less than a planner would, because the deciding majority has been cut out of the upside of developing AI. The same distributional force that makes an unequal society over-risk for a proportional gamble makes it under-risk for a regressive one, and in both cases the wedge from the first best can be closed by redistribution, which here would

restore the broad gains that keep poor voters in the pro-AI coalition.

The next proposition collects the results on inequality-increasing AI.

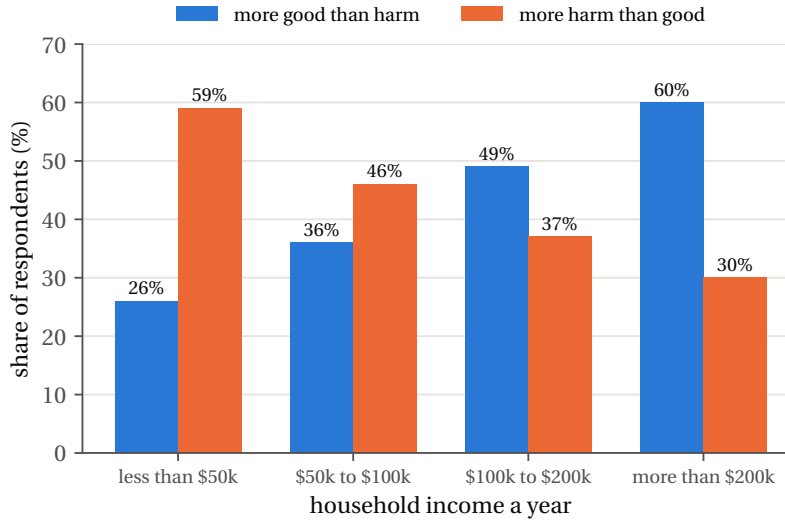
Proposition 5. *Under the rank-preserving widening to $\sigma' > \sigma$ with the mean level unchanged, and Assumption 1 with $\gamma > 1$ in (ii) and (iii):*

- (i) the redistributive first best $\delta^*(\bar{c})$ does not change, while the median consumer's tolerated risk is strictly decreasing in σ' ; her over-risking falls and turns into under-risking above a finite dispersion, and she rejects AI at any risk if and only if (6) holds;*
- (ii) the risk society takes, the median of the individual cut-offs, never exceeds the median consumer's tolerated risk, and equals it whenever every agent poorer than the median tolerates at least as much risk as she does; it is then strictly decreasing in σ' , and society under-risks for every σ' beyond the dispersion at which the median consumer's tolerated risk falls below the first best;*
- (iii) the pro-risk coalition at the risk society takes is half the population and always contains the median consumer; it is the bottom half of the distribution whenever every agent poorer than the median tolerates at least as much risk as she does, and otherwise an interval of ranks that excludes both the richest and the poorest.*

Inequality in the gains can also rationalise a pattern in surveys of opinion. In a national poll of April 2025, 60% of Americans with household incomes above \$200,000 a year expected AI to do more good than harm in their day-to-day life, while 59% of those with incomes below \$50,000 expected more harm than good (Quinnipiac University Poll, 2025). Proposition 1 alone predicts the opposite ranking, since the rich have more life to lose. An AI that concentrates its gains reverses it: the poor, cut out of the upside, come to oppose AI at risks that the better off, whose level rises with the widening, still accept. Figure 5 shows the full gradient: the share expecting more good than harm rises with income in every bracket, from 26% to 60%, and the share expecting more harm falls from 59% to 30%.

Some of the results in this section rest on the log-normality of the consumption distribution before and after AI is developed, and in particular the shape of the pro-risk coalition. A widening need not keep the distribution log-normal, and the Pareto tail at the top, harmless when only the median matters, could matter here: if the top gains from AI even more than in proportion, it may join the coalition in favour of AI. The top may also favour AI for a reason outside this section: if AI also reduces other existential risks, even the richest, whose lives are the most valuable, tolerate it for the protection it buys, as Section 4.2 shows.

Figure 5: Views of AI in day-to-day life by household income



Shares of US adults who expect AI to do more good than harm, or more harm than good, in their day-to-day life, by household income a year; the remainder did not offer an opinion. Quinnipiac University national poll of 1,562 adults, fielded 3 to 7 April 2025, margin of error 2.5 percentage points (Quinnipiac University Poll, 2025).

Redistribution as a political choice. Redistribution has so far appeared as a planner’s instrument that closes the wedge between the decisive voter and the first best. But it is itself a collective choice, and existential risk reshapes its politics. The standard result (Meltzer and Richard, 1981) is that the rich resist redistribution, since a voter above the mean is a net contributor. Existential risk pushes the other way: because the value of life rises with consumption, the rich have the most to lose from extinction, and redistribution toward the median lowers the risk that the decisive voter would otherwise impose on all. The rich may then back redistribution as self-protection rather than altruism.

A simple extension capturing this intuition in the model with constant inequality is the following. Index redistribution by $\tau \in [0, 1]$, compressing each consumption toward the mean as $c_i(\tau) = (1 - \tau) c_i + \tau \bar{c}$, so that $\tau = 0$ is the status quo and $\tau = 1$ full equalisation. Absent any other friction, any consumer with $c_i < \bar{c}$ prefers $\tau = 1$, any consumer with $c_i > \bar{c}$ prefers $\tau = 0$ and a consumer who consumes the average value is indifferent across levels of redistribution. For a right-skewed distribution, $c_{\text{med}} < \bar{c}$, complete redistribution, $\tau = 1$, is then enacted, and the tail at the top opposes it. Consider now the existential risk from AI. Crucially, redistribution in this form preserves rankings within the distribution, so the median consumer is decisive at any τ . The following proposition explores who is willing to vote in favour of a positive degree of redistribution when AI is developed at the risk the decisive median voter tolerates.

Proposition 6. Assume that AI is developed at risk $\delta^*(c_{\text{med}})$. A voter of consumption c_i supports a marginal redistribution from the status quo if and only if

$$v(c_i, g_{\text{ai}}) \frac{|\delta^{*'}(c_{\text{med}})| (\bar{c} - c_{\text{med}})}{1 - \delta^*(c_{\text{med}})} > 1 - \frac{\bar{c}}{c_i} : \quad (7)$$

the value of her life, times the proportional gain in survival that the redistribution brings, exceeds the share of consumption she gives up.

Support can be U-shaped in consumption: every net receiver favours full redistribution and votes for $\tau = 1$, since the right-hand side of (7) is negative for her; an intermediate band of the moderately rich may oppose it, and, because the share of consumption given up is bounded above by one while the value of life grows without bound under CRRA with $\gamma > 1$, all agents above a finite threshold favour it again. The richest therefore side with the poorest, reversing the Meltzer and Richard result at the top.

The proportional gain in survival is small: at the baseline calibration, with $\gamma = 2$ and a consumption Gini of 0.30, a full equalisation raises the survival probability by about one per cent to first order, so a voter supports a marginal redistribution when one per cent of the value of her life exceeds the share of consumption she gives up, which is below one. Whether the opposing band exists depends on the size of the growth gains delivered by AI. With growth of 10% a year the value of life on the AI path is $v(c_i, g_{\text{ai}}) = 42c_i - 1$, and (7) holds for every c_i : all voters, rich and poor, support a marginal redistribution, because a lifetime of fast growth is worth protecting even for the moderately rich. With a smaller prize, growth of 4% a year,¹⁰ the survival gain falls to about half a per cent against $v(c_i, g_{\text{ai}}) = 21c_i - 1$: voters between 1.1 and 8.3 times mean consumption oppose redistribution, and support returns above that, in the Pareto top that the log-normal omits, where the value of life rises without bound like $c_i^{\gamma-1}$. The over-turning of Meltzer and Richard at the top is therefore robust; the band of opposition in the middle depends on the consumption gains AI delivers.

4 AI as a life-saving technology

So far AI has raised consumption growth at a one-time risk of extinction, and its benefit has run into the diminishing marginal utility of consumption. However, AI may have another upside: saving lives. Jones (2024) shows that if AI lengthens life, much larger extinction risks become worth bearing, because longevity gains are in the same units as extinction risk. This section studies how two life-saving channels change the politics of the existential gamble on AI development. Section 4.1 lets AI lengthen life and asks whether longevity changes who favours AI

¹⁰This is still twice the long run growth rate in the US.

and how much an unequal society over-risks. Section 4.2 lets AI bear on other existential risks, which threaten humanity whether or not AI is built, and asks when AI is worth developing as a defence against them and how inequality distorts that choice.

4.1 AI that lengthens life

Suppose that AI also lengthens life: mortality falls with AI from μ_0 to $\mu_{ai} \leq \mu_0$. In this section, I write the mortality parameter explicitly in utility, so that $U(c, g, \mu) \equiv \int_0^\infty e^{-(\rho+\mu)a} u(c e^{g a}) da$.

The mortality-adjusted risk cut-off for an individual is then

$$\delta^*(c) = 1 - \frac{U(c, g_0, \mu_0)}{U(c, g_{ai}, \mu_{ai})}.$$

Consider first a pure longevity gain, $g_0 = g_{ai} = 0$. With constant consumption, lifetime utility is $U(c, 0, \mu) = u(c)/(\rho + \mu)$, so for every agent and every flow utility with $u(c) > 0$,

$$\delta^*(c) = 1 - \frac{u(c)/(\rho + \mu_0)}{u(c)/(\rho + \mu_{ai})} = \frac{\mu_0 - \mu_{ai}}{\rho + \mu_0}.$$

A longer life is bought with extinction risk, and both are measured in years of life, so $u(c)$ cancels. This implies that the maximum tolerated risk is the same for the poor and the rich. The value of a year of life governs trades of consumption against life, and this gamble involves none. With positive consumption growth, the two channels combine. The cut-off factors into them,

$$\delta^*(c) = 1 - \frac{U(c, g_0, \mu_0)}{U(c, g_{ai}, \mu_0)} \cdot \frac{U(c, g_{ai}, \mu_0)}{U(c, g_{ai}, \mu_{ai})}, \quad (8)$$

the growth channel at the old mortality times the longevity channel at the new growth rate. Under Assumption 1 with $\gamma > 1$, the longevity factor is strictly decreasing in c when $g_{ai} > 0$, and the cut-off has a closed form, both derived in Appendix A.

Proposition 7. (i) Let $\gamma > 1$, $\mu_{ai} < \mu_0$, $g_{ai} > g_0$, $c_{med} < \bar{c}$, and let the median prefer life in the status quo to death, $U(c_{med}, g_0, \mu_0) > 0$. As $c \rightarrow \infty$, $\delta^*(c) \rightarrow (\mu_0 - \mu_{ai})/(\rho + \mu_0) > 0$ from above, so a risk at or below this floor is accepted by every agent. The median's over-risking ratio $\delta^*(c_{med})/\delta^*(\bar{c}) - 1$ is strictly smaller with the longevity gain than without it, and it is strictly decreasing in the mortality gain $\mu_0 - \mu_{ai}$ as long as the gain does not exceed $(\gamma - 1)(g_{ai} - g_0)$.

(ii) Let $u' > 0$, $g_0, g_{ai} > 0$, lifetime utilities finite with $U(c, g_0, \mu_0) > 0$, and v nondecreasing on $[c, \infty)$ and strictly increasing on a set of positive measure. Then δ^* is strictly decreasing in c if and only if $g_{ai}/(\rho + \mu_{ai}) > g_0/(\rho + \mu_0)$, strictly increasing if and only if the inequality is

reversed, and constant if and only if the two ratios are equal, in which case $\delta^*(c) = (\mu_0 - \mu_{ai})/(\rho + \mu_0)$ for every agent and every u .

If AI also lengthens life, the gain is worth the same extinction risk to everyone: a longer life bought with the risk of a shorter one is a trade within a single unit, years of life, so the value of a year of life cancels. The gain enters each cut-off as a common factor, lifts every cut-off above a common floor, one quarter when AI halves mortality, and leaves the ranking of agents unchanged. Inequality acts on the consumption part of the gamble only, and the longevity gain reduces the weight of that part, so the median's over-risking falls. Table 3 quantifies these effects. Halving mortality, as in Jones (2024), raises the cut-off of the mean at $\gamma = 2$ from 4.88% to 29.03% and that of the median from 5.68% to 29.69%, and cuts over-risking from 16.5% to 2.3%, and at $\gamma = 3$ from 34.9% to 2.0%. At the calibration the over-risking ratio falls monotonically in the longevity gain.

Table 3: AI that also lengthens life

γ	μ_{ai} (%)	floor (%)	$\delta^*(\bar{c})$ (%)	$\delta^*(c_{med})$ (%)	$\mathcal{O}_{med}$ (%)
1.01	1.0	0.0	35.05	35.57	1.5
1.01	0.5	25.0	57.20	57.62	0.7
2	1.0	0.0	4.88	5.68	16.5
2	0.5	25.0	29.03	29.69	2.3
3	1.0	0.0	1.88	2.53	34.9
3	0.5	25.0	26.53	27.06	2.0

Notes. Cut-offs with $g_{ai} = 10\%$ and the mortality μ_{ai} shown, Gini 0.30; the floor is $(\mu_0 - \mu_{ai})/(\rho + \mu_0)$. Halving mortality raises every cut-off by about the floor and cuts the median's over-risking from 16.5% to 2.3% at $\gamma = 2$.

Longevity gains leave the pro-AI coalition unchanged: it is the poor who favour developing it. One may expect the opposite: a year of life is worth more to the rich, so a longer life should turn them into the supporters of AI. This intuition overlooks that the extinction risk is also measured in years of life, so that the value of a year cancels, and that with a common mortality gain nothing in the trade distinguishes the rich from the poor. The intuition could hold only if AI lengthened life more for the rich than for the poor.

Who favours AI is thus decided by two signs: the slope of the value of life, and the change in growth per unit of the effective discount rate, $g/(\rho + \mu)$. The value of life governs the level of the cut-off, and the ratio decides whether AI buys a richer life or a longer one.

4.2 AI and other existential risks

The analysis so far weighed the AI risk against a safe status quo. However, other catastrophes, including but not limited to asteroid impacts, pandemics, climate change, and alien invasions, could end humanity whether or not AI is built. At the same time, AI might be used to prevent them.¹¹ This subsection adds a background extinction hazard and asks when AI is worth developing once it bears on other risks, and how inequality distorts its development as a defence against them.

The net-survival rule. Society faces a background extinction hazard with or without AI. Without AI it survives it with probability $1 - \delta_0$. Developing AI adds its own hazard δ_{AI} , but also moves the background hazard to $\delta'_0 < \delta_0$, and raises growth to g_{ai} ; with independent competing risks, survival is $(1 - \delta_{AI})(1 - \delta'_0)$. Consider first a society with no inequality, in which all agents consume c_0 today and grow at g_0 , or at g_{ai} if they survive AI development.

With the baseline cut-off $\delta^* = 1 - U(c_0, g_0)/U(c_0, g_{ai})$ of (2), AI development is worthwhile if and only if $(1 - \delta_{AI})(1 - \delta'_0) U(c_0, g_{ai}) \geq (1 - \delta_0) U(c_0, g_0)$, that is, if and only if

$$\delta_{AI} \leq \delta_{AI}^* = 1 - \frac{1 - \delta_0}{1 - \delta'_0} (1 - \delta^*) .$$

The threshold is decreasing in consumption but it does not vanish as $c_0 \rightarrow \infty$: since $\delta^* \rightarrow 0$ by (5), it converges to a positive defensive floor, $\delta_{AI}^* \rightarrow (\delta_0 - \delta'_0)/(1 - \delta'_0)$.

To first order,

$$\delta_{AI}^* \approx \delta^* + \delta_0 - \delta'_0 :$$

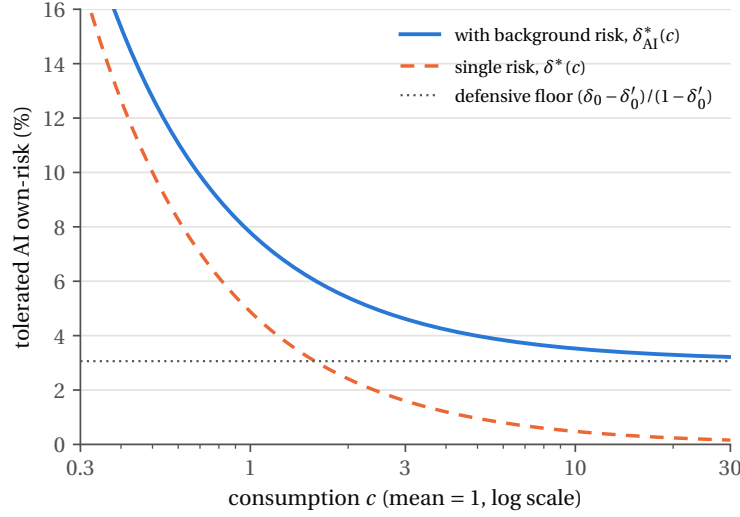
tolerated AI own-risk is the growth-justified baseline δ^* plus a one-for-one credit for the background risk AI removes. AI can therefore be worth developing even with no growth prize, purely because it cuts the background hazard, and an AI that improves net safety is always developed. As a society grows rich the growth motive for AI, δ^* , vanishes, but the defensive motive persists, so a rich society tolerates AI only for defence. Figure 6 illustrates.

Developing AI for defensive purposes only. A society could instead develop AI only to reduce other existential risks, or confine its use to them, which would arguably reduce the risk AI itself creates. Such a restricted AI is costly, and the framework above cannot isolate the cost, since AI is assumed to raise growth at the same time.

To isolate the defensive mechanism, suppose that AI can be developed purely to cut the

¹¹It may also worsen some of these risks, for example climate change through power and water consumption. For the remainder of this subsection such increases in risk are thought to be folded into the baseline AI risk.

Figure 6: The defensive floor



Tolerated AI own-risk against consumption with a background risk that AI can cut, and in the single-risk benchmark, dashed. With the background risk the cut-off falls with consumption but converges to a positive floor, dotted, instead of zero: a rich society tolerates AI only for defence. Calibration: $\gamma = 2$, $\delta_0 = 0.05$, $\delta'_0 = 0.02$, $\nu(\bar{c}) = 6$, $\rho = \mu = 1\%$, $g_0 = 2\%$, $g_{ai} = 10\%$.

background hazard, at a cost: paying a share s of consumption for ever buys a reduction from δ_0 to δ'_0 . Consider first a society with no inequality. A society with common consumption c_0 adopts AI for defence if and only if the cost share is below the tolerated cost s^* , which solves the indifference condition

$$(1 - \delta'_0) U((1 - s^*)c_0, g_0) = (1 - \delta_0) U(c_0, g_0), \quad (9)$$

and under Assumption 1 the tolerated cost rises with consumption, as Appendix A shows: richer societies defend more. To first order,

$$s^*(c_0) \approx \frac{\delta_0 - \delta'_0}{1 - \delta'_0} \nu(c_0, g_0).$$

This is the mirror of the tolerated risk $\delta^*(c_0)$, with the value of life now in the numerator rather than the denominator, so that defence becomes more attractive with consumption just as the growth gamble becomes less. Under CRRA utility the tolerated share is exactly $s^*(c_0) = 1 - [1 + (\gamma - 1) \frac{\delta_0 - \delta'_0}{1 - \delta'_0} \nu(c_0, g_0)]^{-1/(\gamma - 1)}$.

With inequality, how the defence is financed must be specified. The simplest case has every agent contribute the same share of her consumption, and implies that AI is under-developed.

Proposition 8. *With heterogeneous agents and defence a common good financed by a uniform share s , each agent favours it if and only if $s \leq s^*(c_i)$, so the median is decisive: a majority adopts if and only if $s \leq \text{med}_i s_i^* = s^*(c_{\text{med}})$, while the redistributive first best adopts if and only if $s \leq s^*(\bar{c})$. If the distribution is skewed to the right, $s^*(c_{\text{med}}) < s^*(\bar{c})$, inequality makes society under-defend.*

The logic mirrors Section 3, with the coalition reversed: the tolerated share rises with consumption, so at any cost the agents in favour are the richer ones, and the pro-AI coalition at the median's cut-off is the median and everyone richer than her, not the median and everyone poorer as in the growth gamble. The median consumer is again decisive, and with right skew she is poorer than the average and accepts a lower cost than the first best would, so society under-defends.

The optimal financing of defensive AI, a public good to which richer agents can and may want to contribute more, is a natural avenue for future research.

Heterogeneous exposure. Introducing other sources of extinction risk opens up a new margin for heterogeneity. The poor are typically more exposed to background risks, including pandemics and catastrophic climate events such as floods and droughts. I capture this with a distribution of the background hazard $\delta_{0,i}$, while AI's own gamble still entails common extinction. To separate the role of exposure from its level, I start from equal exposure and spread it preserving its mean, $\mathbb{E}[\delta_{0,i}] = \bar{\delta}_0$, with exposure falling with consumption along a decreasing schedule $\delta_0(c)$. Developing AI and using it against background risk removes a fraction π of an agent's own hazard for a common cost share s , so to first order agent i favours it whenever the benefit covers the cost,

$$\pi \delta_0(c_i) v(c_i, g_0) \geq s,$$

the benefit being the fraction of risk removed, times exposure, times the value of life.

The left-hand side is agent i 's tolerated cost, $s^*(c_i) = \pi \delta_0(c_i) v(c_i, g_0)$, the largest share she will pay for reducing life-threatening risk. Two forces now pull against each other: the value of life $v(c_i, g_0)$ that rises with consumption, and exposure $\delta_0(c_i)$ that falls with it. Unlike the homogeneous tolerated cost of (9), which moved only with v and was therefore monotone, $s^*(c_i)$ need not be increasing and, exactly as in Section 3, whether the median is decisive and the shape of the pro-AI coalition depend on whether the cut-off is monotone in consumption.

Proposition 9. *Consider the model with heterogeneous exposure to background existential risk in which AI reduces individual risk by a common fraction. The median consumer is decisive if and only if the tolerated cost $s^*(c) = \pi \delta_0(c) v(c, g_0)$ is monotone in consumption, otherwise s^* is hump-shaped, and the decisive threshold becomes the median cut-off $\text{med}_i s^*(c_i) \neq s^*(c_{\text{med}})$.*

When the median is decisive, society defends if and only if $s \leq s^*(c_{\text{med}})$, against a redistributive first best $s \leq \pi \bar{\delta}_0 \nu(\bar{c}, g_0)$, and so under-defends if and only if

$$\frac{\nu(c_{\text{med}}, g_0)}{\nu(\bar{c}, g_0)} \cdot \frac{\delta_0(c_{\text{med}})}{\bar{\delta}_0} < 1.$$

The under-defence condition factors two forces. The value-of-life factor $\nu(c_{\text{med}}, g_0)/\nu(\bar{c}, g_0)$ is less than unity for any right-skewed distribution, the same channel as in the homogeneous case, pushing the poorer median to under-defend. The exposure factor $\delta_0(c_{\text{med}})/\bar{\delta}_0$ pulls the other way when the median is more exposed than the average, dampening under-defence, and reinforces it when she is less exposed.

The potential loss of the median's decisiveness mirrors Section 3: there a widening of the gains concentrated the prize and made the risk cut-off δ^* hump-shaped; here concentrated exposure makes the cost cut-off s^* hump-shaped.

5 Conclusion

AI may transform our economies and usher in an age of greater prosperity, but it may also end them: developing it carries a risk of human extinction. Under majoritarian voting, inequality makes society accept more risk than it would under perfect equality: at the baseline calibration, the decisive median voter accepts a 5.7% probability of extinction against the 4.9% of a society with equal consumption, because her lower consumption makes her life worth less relative to the gains that AI promises.

Inequality fundamentally affects risk taking beyond majoritarian voting: a utilitarian planner unable to redistribute may take even more risk than the median consumer. Supermajority rules and votes weighted by consumption make a richer voter decisive and are partial substitutes for redistribution.

How the gains from AI are shared matters as much as how consumption is. If AI concentrates its gains, the median voter, cut out of the upside, turns cautious: a society that over-risks for a gamble with proportional gains under-risks for a regressive one. The pattern in opinion polls, in which the poor expect AI to do more harm than good in their lives and the rich expect the opposite, is consistent with gains that exacerbate inequality.

If AI also lengthens life, this gain is in the same units as the extinction risk, and it raises the risk every voter tolerates by about the same amount, shrinking the wedge caused by inequality. Where AI can also avert other extinction threats, an unequal society under-develops it instead: the median voter, poorer than the average, pays less for protection than she would do with perfect equality, and the coalition in favour of AI becomes the median and the richer, the mirror

image of the growth gamble.

Redistribution lowers the risk society takes and raises its chances of survival: it is a policy for managing extinction risk, and one that the rich, who have the most life to lose, may come to support.

References

- Acemoglu, D. and T. Lensman (2024). Regulating transformative technologies. *American Economic Review: Insights* 6(3), 359–376.
- Adler, M. D. (2012). *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. New York: Oxford University Press.
- Adler, M. D., M. Ferranna, J. K. Hammitt, and N. Treich (2021). Fair innings? the utilitarian and prioritarian value of risk reduction over a whole lifetime. *Journal of Health Economics* 75, 102412.
- Adler, M. D., J. K. Hammitt, and N. Treich (2014). The social value of mortality risk reduction: VSL versus the social welfare function approach. *Journal of Health Economics* 35, 82–93.
- Aschenbrenner, L. (2020). Existential risk and growth. Technical Report 6-2020, Global Priorities Institute.
- Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory* 2(3), 244–263.
- Bommier, A., A. Kochov, and F. Le Grand (2017). On monotone recursive preferences. *Econometrica* 85(5), 1433–1466.
- Bommier, A., B. Lanz, and S. Zuber (2015). Models-as-usual for unusual risks? on the value of catastrophic climate change. *Journal of Environmental Economics and Management* 74, 1–22.
- Cecco, L. (2026). AI godfather Geoffrey Hinton: ‘We’re creating alien beings. I’m willing to be selfish and choose humans’. *The Guardian*, 13 October. <https://www.theguardian.com/us-news/2026/oct/13/geoffrey-hinton-trump-ai-tech>.
- Congressional Budget Office (2023). The distribution of household income in 2020. CBO Publication 59757, Congressional Budget Office.

- Fisher, J., D. S. Johnson, and T. M. Smeeding (2015). Inequality of income and consumption in the U.S.: Measuring the trends in inequality from 1984 to 2011 for the same individuals. *Review of Income and Wealth* 61(4), 630–650.
- Fleurbaey, M. (2010). Assessing risky social situations. *Journal of Political Economy* 118(4), 649–680.
- Hajdari, U. (2026). ‘A gamble with our lives’: Ex-Anthropic researcher warns of AI ‘catastrophe’. *Euronews*, 9 September. <https://www.euronews.com/2026/09/09/a-gamble-with-our-lives-ex-anthropic-researcher-warns-of-ai-catastrophe>.
- Hall, R. E. and C. I. Jones (2007). The value of life and the rise in health spending. *Quarterly Journal of Economics* 122(1), 39–72.
- Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63(4), 309–321.
- Jones, C. I. (2016a). The facts of economic growth. In J. B. Taylor and H. Uhlig (Eds.), *Handbook of Macroeconomics*, Volume 2(1), pp. 3–69. Elsevier.
- Jones, C. I. (2016b). Life and growth. *Journal of Political Economy* 124(2), 539–578.
- Jones, C. I. (2024). The AI dilemma: Growth versus existential risk. *AER: Insights* 6(4), 575–590.
- Karger, E., O. Kuusela, J. Abaluck, K. Bryan, B. Halperin, T. Jones, C. Murphy, P. Trammell, M. Reynolds, D. Mayland, R. Viswanathan, A. Mittal, R. Ceppas de Castro, J. Rosenberg, and P. E. Tetlock (2026). Forecasting the economic effects of AI. Working Paper 2026-07, Federal Reserve Bank of Chicago.
- Kuziemko, I., R. W. Buell, T. Reich, and M. I. Norton (2014). “Last-place aversion”: Evidence and redistributive implications. *Quarterly Journal of Economics* 129(1), 105–149.
- Martin, I. W. R. and R. S. Pindyck (2015). Averting catastrophes: The strange economics of scylla and charybdis. *American Economic Review* 105(10), 2947–2985.
- Meltzer, A. H. and S. F. Richard (1981). A rational theory of the size of government. *Journal of Political Economy* 89(5), 914–927.
- Moene, K. O. and M. Wallerstein (2001). Inequality, social insurance, and redistribution. *American Political Science Review* 95(4), 859–874.
- Murphy, K. M. and R. H. Topel (2006). The value of health and longevity. *Journal of Political Economy* 114(5), 871–904.

Payne, B. K., J. L. Brown-Iannuzzi, and J. W. Hannay (2017). Economic inequality increases risk taking. *Proceedings of the National Academy of Sciences* 114(18), 4643–4648.

Persson, T. and G. Tabellini (2000). *Political Economics: Explaining Economic Policy*. MIT Press.

Quinnipiac University Poll (2025). The age of artificial intelligence: Americans wary of impact on daily life, see harm to education, benefits to medical advances. National poll release, 16 April 2025, fielded 3–7 April 2025, 1,562 adults.

Rosen, S. (1988). The value of changes in life expectancy. *Journal of Risk and Uncertainty* 1(3), 285–304.

Trammell, P. and L. Aschenbrenner (2026). Existential risk and growth. Working paper.

A Proofs

Proof of Proposition 1. (i) From (2), writing U_0, U_{ai} for the two lifetime utilities and primes for c -derivatives, $\frac{d\delta^*}{dc} = -\frac{U'_0 U_{ai} - U_0 U'_{ai}}{U_{ai}^2} = -\frac{U_0}{U_{ai}} \left[\frac{U'_0}{U_0} - \frac{U'_{ai}}{U_{ai}} \right]$. Since $U_{ai} \geq U_0 > 0$, the prefactor is negative; multiplying the bracket by $c > 0$ turns the log-derivatives into elasticities, and by (3) $\partial \ln U(c, g) / \partial \ln c = 1 / \nu(c, g)$, so $\text{sign } d\delta^* / dc = \text{sign}[1 / \nu(c, g_{ai}) - 1 / \nu(c, g_0)] = \text{sign}[\nu(c, g_0) - \nu(c, g_{ai})]$, both values being positive.

(ii) Three steps. *Step 1: the value of life along a path is a weighted average of ν .* In (1) change the variable of integration to the cumulative log gain $\ell = ga$, $da = d\ell / g$: $U(c, g) = \frac{1}{g} \int_0^\infty e^{-\frac{\rho+\mu}{g}\ell} u'(ce^\ell) d\ell$. Differentiating under the integral (finiteness), $\frac{\partial U}{\partial \ln c} = \frac{1}{g} \int_0^\infty e^{-\frac{\rho+\mu}{g}\ell} u'(ce^\ell) ce^\ell d\ell$. Writing $u(y) = \nu(y) u'(y) y$ inside the expression for U and dividing the two integrals, the factor $1/g$ cancels and

$$\nu(c, g) = \frac{\int_0^\infty e^{-\frac{\rho+\mu}{g}\ell} u'(ce^\ell) ce^\ell \nu(ce^\ell) d\ell}{\int_0^\infty e^{-\frac{\rho+\mu}{g}\ell} u'(ce^\ell) ce^\ell d\ell} :$$

on each path, $\nu(c, g)$ is the mean of the same function $\ell \mapsto \nu(ce^\ell)$ under the path's own weight $e^{-\frac{\rho+\mu}{g}\ell} u'(ce^\ell) ce^\ell$.

Step 2: the AI weight dominates. Let F_0 and F_{ai} (this proof only) be the cumulative distributions of the two weights, normalised to mass one on $[0, \infty)$. The ratio of the AI weight to the status-quo weight is $\exp\{[\frac{\rho+\mu}{g_0} - \frac{\rho+\mu}{g_{ai}}]\ell\}$, strictly increasing in ℓ because $g_{ai} > g_0$. A strictly increasing likelihood ratio L implies $F_0 \geq F_{ai}$ pointwise, strictly on $(0, \infty)$: for any $\bar{\ell}$, $1 - F_{ai}(\bar{\ell}) = \int_{\bar{\ell}}^\infty L dF_0 / \int_0^\infty L dF_0 \geq \int_{\bar{\ell}}^\infty dF_0 = 1 - F_0(\bar{\ell})$, because L is larger on $[\bar{\ell}, \infty)$ than its F_0 -average.

Step 3: the gradient as a weighted integral of ν' . Integrating by parts, with $\frac{d}{d\ell} \nu(ce^\ell) = \nu'(ce^\ell) ce^\ell$ and the boundary terms vanishing ($F_0 = F_{ai} = 0$ at 0; $\nu(ce^\ell)[1 - F_j(\ell)] \rightarrow 0$ at ∞

because v is eventually monotone or bounded and the integrals of Step 1 are finite),

$$v(c, g_{\text{ai}}) - v(c, g_0) = \int_0^\infty v(c e^\ell) d[F_{\text{ai}} - F_0](\ell) = \int_0^\infty v'(c e^\ell) c e^\ell [F_0(\ell) - F_{\text{ai}}(\ell)] d\ell.$$

Writing the bracket in (i) as $\frac{1}{v(c, g_0)} - \frac{1}{v(c, g_{\text{ai}})} = \frac{v(c, g_{\text{ai}}) - v(c, g_0)}{v(c, g_0) v(c, g_{\text{ai}})}$ and substituting,

$$\frac{d\delta^*}{dc} = - \frac{U(c, g_0)}{c U(c, g_{\text{ai}}) v(c, g_0) v(c, g_{\text{ai}})} \int_0^\infty v'(c e^\ell) c e^\ell [F_0(\ell) - F_{\text{ai}}(\ell)] d\ell.$$

The prefactor is negative and the integrand is v' times a weight that is positive on $(0, \infty)$ by Step 2, so the sign of the gradient is minus the sign of v' on $[c, \infty)$: a v' that is positive on a set of positive measure gives a positive integral and a strictly decreasing δ^* ; the other two cases follow in the same way. $\square$

Derivation of (5), and the cut-off in terms of the yearly value of life. By direct integration of (1), $U(c, g) = \bar{u}/(\rho + \mu) + c^{1-\gamma}/[(1-\gamma)(\rho + \mu + (\gamma-1)g)]$ for $\gamma \neq 1$ (eq. (4) of Jones, 2024) and $U(c, g) = (\bar{u} + \log c)/(\rho + \mu) + g/(\rho + \mu)^2$ for $\gamma = 1$, finite under Assumption 1. Differentiating, $\partial U/\partial \ln c = c^{1-\gamma}/(\rho + \mu + (\gamma-1)g)$ for $\gamma \neq 1$ and $1/(\rho + \mu)$ for $\gamma = 1$; dividing gives (4), and $c \partial_c v(c, g) = (\gamma - 1)[v(c, g) - \frac{1}{1-\gamma}] = 1 + (\gamma - 1)v(c, g)$. For $\gamma \neq 1$, $U(c, g) = \bar{u}/(\rho + \mu) - \frac{c^{1-\gamma}}{\gamma-1} \cdot \frac{1}{\rho + \mu + (\gamma-1)g}$, so

$$\delta^*(c) = \frac{U(c, g_{\text{ai}}) - U(c, g_0)}{U(c, g_{\text{ai}})} = \frac{\frac{c^{1-\gamma}}{\gamma-1} \left[\frac{1}{\rho + \mu + (\gamma-1)g_0} - \frac{1}{\rho + \mu + (\gamma-1)g_{\text{ai}}} \right]}{\frac{\bar{u}}{\rho + \mu} - \frac{c^{1-\gamma}}{\gamma-1} \cdot \frac{1}{\rho + \mu + (\gamma-1)g_{\text{ai}}}}.$$

Multiply numerator and denominator by $(\rho + \mu)(\gamma - 1)c^{\gamma-1}$ and use $(\gamma - 1)\bar{u}c^{\gamma-1} = (\gamma - 1)v(c) + 1$, which follows from $v(c) = \bar{u}c^{\gamma-1} + 1/(1 - \gamma)$; this gives

$$\delta^*(c) = \frac{\frac{\rho + \mu}{\rho + \mu + (\gamma-1)g_0} - \frac{\rho + \mu}{\rho + \mu + (\gamma-1)g_{\text{ai}}}}{(\gamma - 1)v(c) + 1 - \frac{\rho + \mu}{\rho + \mu + (\gamma-1)g_{\text{ai}}}}. \quad (10)$$

Its numerator is a constant and its denominator is $(\gamma - 1)v(c) + \text{const}$, so

$$\frac{d\delta^*}{dc} = - \frac{\text{numerator of (10)} \cdot (\gamma - 1)v'(c)}{[\text{denominator of (10)}]^2}, \quad \text{numerator of (10)} = \frac{(\rho + \mu)(\gamma - 1)(g_{\text{ai}} - g_0)}{(\rho + \mu + (\gamma - 1)g_0)(\rho + \mu + (\gamma - 1)g_{\text{ai}})},$$

which gives

$$\frac{d\delta^*}{dc} = - \frac{(\gamma - 1)^2(\rho + \mu)(g_{\text{ai}} - g_0)}{(\rho + \mu + (\gamma - 1)g_0)(\rho + \mu + (\gamma - 1)g_{\text{ai}})} \cdot \frac{v'(c)}{\left[(\gamma - 1)v(c) + 1 - \frac{\rho + \mu}{\rho + \mu + (\gamma - 1)g_{\text{ai}}} \right]^2}. \quad (11)$$

The denominator of (10) equals $(\rho + \mu)(\gamma - 1)c^{\gamma-1}U(c, g_{ai}) > 0$ under Assumption 1 whenever $U(c, g_0) > 0$, since $U(c, g_{ai}) > U(c, g_0)$, and so is the numerator, $(\rho + \mu)(\gamma - 1)c^{\gamma-1}[U(c, g_{ai}) - U(c, g_0)]$, so $\delta^* \in (0, 1)$. For (5), by (4) and $(\gamma - 1)\bar{u}c^{\gamma-1} = (\gamma - 1)v(c) + 1$,

$$(\gamma - 1)v(c, g_{ai}) = [(\gamma - 1)v(c) + 1] \frac{\rho + \mu + (\gamma - 1)g_{ai}}{\rho + \mu} - 1 = \frac{\rho + \mu + (\gamma - 1)g_{ai}}{\rho + \mu} \left[(\gamma - 1)v(c) + 1 - \frac{\rho + \mu}{\rho + \mu + (\gamma - 1)g_{ai}} \right],$$

so the denominator of (10) equals $(\gamma - 1) \frac{\rho + \mu}{\rho + \mu + (\gamma - 1)g_{ai}} v(c, g_{ai})$, while its numerator equals $\frac{(\rho + \mu)(\gamma - 1)(g_{ai} - g_0)}{(\rho + \mu + (\gamma - 1)g_0)(\rho + \mu + (\gamma - 1)g_{ai})}$; dividing gives (5). For $\gamma = 1$, (4) gives $(\rho + \mu)v(c, g_{ai}) = (\rho + \mu)v(c) + g_{ai}$, so the log cut-off is $(g_{ai} - g_0)/[(\rho + \mu)v(c, g_{ai})]$, which is (5) at $\gamma = 1$. For $\gamma = 1$, $U(c, g_{ai}) - U(c, g_0) = (g_{ai} - g_0)/(\rho + \mu)^2$ and $U(c, g_{ai}) = [(\rho + \mu)(\bar{u} + \log c) + g_{ai}]/(\rho + \mu)^2$; dividing gives $\delta^*(c) = (g_{ai} - g_0)/[(\rho + \mu)v(c) + g_{ai}]$ with $v(c) = \bar{u} + \log c$, and differentiating, since the numerator is constant, $d\delta^*/dc = -(\rho + \mu)(g_{ai} - g_0)v'(c)/[(\rho + \mu)v(c) + g_{ai}]^2$. In both cases $\text{sign } d\delta^*/dc = -\text{sign } v'(c) < 0$ by Assumption 1. Asymptotics: for $\gamma > 1$ the denominator of (10) is $(\gamma - 1)\bar{u}c^{\gamma-1} + \text{const}$, so $\delta^* \sim \text{const} \cdot c^{-(\gamma-1)}$ and its elasticity tends to $-(\gamma - 1)$; for $\gamma = 1$ the elasticity is $-(\rho + \mu)/[(\rho + \mu)v(c) + g_{ai}] \rightarrow 0$.

Proof of Proposition 2. δ^* is strictly decreasing (Proposition 1 under Assumption 1); right-skew is $c_{\text{med}} < \bar{c}$, which the decreasing δ^* reverses to $\delta^*(c_{\text{med}}) > \delta^*(\bar{c})$. The redistributing planner maximises $\int U(c_i, g) di$ s.t. $\int c_i di = \bar{c}$; $U(\cdot, g)$ is strictly concave in c because u is, so $c_i = \bar{c}$. With $\bar{c}$ fixed, $\delta^*(\bar{c})$ is fixed and $\delta^*(c_{\text{med}})$ depends on the distribution only through c_{med} , i.e. through $c_{\text{med}}/\bar{c}$; since δ^* is strictly decreasing, a lower ratio raises $\delta^*(c_{\text{med}})$ and hence the over-risking. $\square$

Proof of Proposition 4. At risk δ the agents in favour are those with $\delta^*(c_i) \geq \delta$; since δ^* is decreasing this is the set $\{c_i \leq c\}$ for $\delta^*(c) = \delta$, a population fraction $F(c)$. Development under the q -rule requires $F(c) \geq q$, i.e. $c \geq c_q = F^{-1}(q)$, equivalently $\delta \leq \delta^*(c_q)$. As δ^* is decreasing and c_q increases in q with $c_{1/2} = c_{\text{med}}$, $\delta^*(c_q) \leq \delta^*(c_{\text{med}})$. $\square$

Proof of Proposition 3. With $U(c_i, g) = \bar{u}/(\rho + \mu) + c_i^{1-\gamma}/[(1-\gamma)(\rho + \mu + (\gamma - 1)g)]$, $\int_0^1 U(c_i, g) di = \bar{u}/(\rho + \mu) + \frac{\int_0^1 c_i^{1-\gamma} di}{(1-\gamma)(\rho + \mu + (\gamma - 1)g)} = U(c_{\text{EDE}}, g)$ with $c_{\text{EDE}} = (\int_0^1 c_i^{1-\gamma} di)^{1/(1-\gamma)}$, on both paths (for $\gamma = 1$, U is affine in $\log c_i$ and $c_{\text{EDE}} = \exp \int \log c_i di$), so the planner's condition $(1 - \delta) \int U(c_i, g_{ai}) di \geq \int U(c_i, g_0) di$ is $\delta \leq \delta^*(c_{\text{EDE}})$. Since $c^{1-\gamma}/(1-\gamma)$ is strictly concave for $\gamma \neq 1$, as is $\log c$, Jensen's inequality gives $c_{\text{EDE}} < \bar{c}$ for any non-degenerate distribution, so $\delta^*(c_{\text{EDE}}) > \delta^*(\bar{c})$ by the decreasing δ^* . For the log-normal $\mathbb{E}[c^t] = \bar{c}^t e^{t(t-1)\sigma^2/2}$ gives, at $t = 1 - \gamma$, $c_{\text{EDE}} = \bar{c} e^{-\gamma\sigma^2/2} < c_{\text{med}} = \bar{c} e^{-\sigma^2/2}$ for $\gamma > 1$; apply the decreasing δ^* . $\square$

Proof of Proposition 5. A log-normal with mean $\bar{c}$ and log-dispersion σ has median $\bar{c}e^{-\sigma^2/2}$, so the median's status quo is $c_{\text{med}} = \bar{c}e^{-\sigma^2/2}$ and the level from which she grows after adoption

is $\bar{c}e^{-\sigma'^2/2}$; her cut-off is $1 - U(c_{\text{med}}, g_0)/U(\bar{c}e^{-\sigma'^2/2}, g_{\text{ai}})$, decreasing in σ' because $U(\cdot, g_{\text{ai}})$ is increasing in its first argument. Since $\mathbb{E}[c'_i] = \bar{c}$ for every σ' , the first best equalises to $\bar{c}$ at adoption and $\delta^*(\bar{c})$ is independent of σ' ; the median's cut-off falls in σ' and crosses the fixed benchmark once. She is harmed if and only if $U(\bar{c}e^{-\sigma'^2/2}, g_{\text{ai}}) < U(c_{\text{med}}, g_0)$. For $\gamma > 1$, by the closed form of U this is $\frac{(\bar{c}e^{-\sigma'^2/2})^{1-\gamma}}{(1-\gamma)(\rho+\mu+(\gamma-1)g_{\text{ai}})} < \frac{(\bar{c}e^{-\sigma'^2/2})^{1-\gamma}}{(1-\gamma)(\rho+\mu+(\gamma-1)g_0)}$; dividing by $\bar{c}^{1-\gamma}/(1-\gamma)$, which is negative and flips the inequality, and taking logarithms, $(\gamma-1)\frac{\sigma'^2-\sigma^2}{2} > \ln \frac{\rho+\mu+(\gamma-1)g_{\text{ai}}}{\rho+\mu+(\gamma-1)g_0}$. For $\gamma = 1$ the closed form gives $-\sigma'^2/2 + g_{\text{ai}}/(\rho+\mu) < -\sigma^2/2 + g_0/(\rho+\mu)$. This proves (i).

For (ii) and (iii), index agents by rank x as in Lemma 1 and let m be the median of the cut-offs, so that the pro-risk coalition $\{x : \delta^*(x) \geq m\}$ has mass one half. For $\gamma > 1$ the lemma makes the profile $x \mapsto \delta^*(x)$ single-peaked on the ranks where the AI path's lifetime utility is positive, and the cut-off is zero where it is not, since the status-quo utility is positive; the coalition is therefore an interval of ranks $[x_{lo}, x_{hi}]$ (this proof only) with $\Phi(x_{hi}) - \Phi(x_{lo}) = 1/2$. Since $\Phi(x_{hi}) \geq 1/2 \geq \Phi(x_{lo})$, the median consumer, $x = 0$, belongs to the coalition, so $m \leq \delta^*(0)$. As $x \rightarrow \infty$ both lifetime utilities converge to $\bar{u}/(\rho+\mu)$ when $\gamma > 1$, so $\delta^*(x) \rightarrow 0$ and $x_{hi} < \infty$ whenever $m > 0$: the richest are always excluded. If every $x < 0$ has $\delta^*(x) \geq \delta^*(0)$, the peak lies at a rank $x \leq 0$, because a peak at a positive rank would make the profile increasing on $(-\infty, 0)$; the profile is then strictly decreasing on $(0, \infty)$, the set $\{\delta^* \geq \delta^*(0)\}$ is $(-\infty, 0]$, of mass one half, and $m = \delta^*(0)$: the median consumer is decisive, the coalition is the bottom half, and m inherits from (i) its strict decrease in σ' . Otherwise $x_{lo} > -\infty$ and the coalition excludes a tail at each end. Finally, by (i), $m \leq \delta^*(0) < \delta^*(\bar{c})$ for every σ' above the dispersion at which the median consumer's cut-off crosses the first best. $\square$

Lemma 1. *Index agents by their rank $x \in \mathbb{R}$, the standard-normal quantile (the percentile is $\Phi(x)$), so that $c(x) = \bar{c}e^{\sigma x - \sigma^2/2}$ and, under the rank-preserving widening, $c'(x) = \bar{c}e^{\sigma'x - \sigma'^2/2}$. For an agent with $U(c(x), g_0) > 0$ and $U(c'(x), g_{\text{ai}}) > 0$, the cut-off $\delta^*(x) = 1 - U(c(x), g_0)/U(c'(x), g_{\text{ai}})$ satisfies*

$$\text{sign } \partial_x \delta^*(x) = \text{sign}[\sigma' v(c(x), g_0) - \sigma v(c'(x), g_{\text{ai}})],$$

so δ^* is increasing in rank where $v(c', g_{\text{ai}})/v(c, g_0) < \sigma'/\sigma$, decreasing where $v(c', g_{\text{ai}})/v(c, g_0) > \sigma'/\sigma$, and stationary where the two are equal. Under Assumption 1 with $\gamma > 1$, the ratio crosses the level σ'/σ at most once and from below, and is unbounded above; hence δ^* has at most one interior turning point, a maximum, and is single-peaked. The peak is interior if and only if the ratio dips below σ'/σ somewhere on the range where both lifetime utilities are positive, which happens when the lower edge of that range is the AI path's utility reaching zero, and not when it is the status quo's.

Proof of Lemma 1. With $\partial_x c = \sigma c$ and $\partial_x c' = \sigma' c'$, the chain rule and (3) give $\partial_x \ln U(c(x), g_0) = \sigma/v(c(x), g_0)$ and $\partial_x \ln U(c'(x), g_{\text{ai}}) = \sigma'/v(c'(x), g_{\text{ai}})$. From $\delta^* = 1 - U_0/U_{\text{ai}}$, $\partial_x \delta^* = (U_0/U_{\text{ai}})[\partial_x \ln U_{\text{ai}} -$

$\partial_x \ln U_0] = (U_0/U_{ai})[\sigma'/\nu(c', g_{ai}) - \sigma/\nu(c, g_0)]$; with $U_0/U_{ai} > 0$ and both values of life positive, multiplying by $\nu(c, g_0)\nu(c', g_{ai}) > 0$ gives the sign rule. For the remaining claims write $R = \nu(c'(x), g_{ai})/\nu(c(x), g_0)$ (this proof only). With $c \partial_c \nu(c, g) = 1 + (\gamma - 1)\nu(c, g)$ (from (4)),

$$\partial_x \ln R = \sigma' \frac{1 + (\gamma - 1)\nu(c', g_{ai})}{\nu(c', g_{ai})} - \sigma \frac{1 + (\gamma - 1)\nu(c, g_0)}{\nu(c, g_0)} = (\gamma - 1)(\sigma' - \sigma) + \left[\frac{\sigma'}{\nu(c', g_{ai})} - \frac{\sigma}{\nu(c, g_0)} \right].$$

At a rank where $R = \sigma'/\sigma$ the bracket vanishes, so $\partial_x \ln R = (\gamma - 1)(\sigma' - \sigma) > 0$: every crossing of σ'/σ is strictly upward, so there is at most one. As $x \rightarrow \infty$, (4) gives $R \rightarrow (c'/c)^{\gamma-1} \frac{\rho+\mu+(\gamma-1)g_{ai}}{\rho+\mu+(\gamma-1)g_0} \rightarrow \infty$. Since $\text{sign } \partial_x \delta^* = \text{sign}(\sigma'/\sigma - R)$, the cut-off rises then falls across the single crossing (and is monotone if there is none), hence single-peaked. An interior peak exists if and only if that crossing occurs, i.e. if and only if $\inf_x R < \sigma'/\sigma$ on the range where both lifetime utilities are positive. Since $R \rightarrow \infty$ at the top, the infimum is governed by the lower edge of the range: as $U(c', g_{ai}) \rightarrow 0^+$, $\nu(c', g_{ai}) = U(c', g_{ai})/(\partial U/\partial \ln c) \rightarrow 0$ while $\nu(c, g_0) > 0$, so $R \rightarrow 0$; as $U(c, g_0) \rightarrow 0^+$, $R \rightarrow \infty$. In the first case $\inf_x R = 0 < \sigma'/\sigma$ and δ^* humps; in the second no downward crossing is possible and δ^* is monotone. $\square$

Proof of Proposition 6. With AI developed at the decisive voter's risk $\delta^*(c_{\text{med}}(\tau))$ and the AI path starting from $c_i(\tau)$, voter i 's expected payoff is $V_i(\tau) = (1 - \delta^*(c_{\text{med}}(\tau)))U(c_i(\tau), g_{ai})$. From $c_i(\tau) = (1 - \tau)c_i + \tau \bar{c}$ one has $c'_i(\tau) = \bar{c} - c_i$ and $c'_{\text{med}}(\tau) = \bar{c} - c_{\text{med}}$, so

$$V'_i(0) = -\delta^{*'}(c_{\text{med}})(\bar{c} - c_{\text{med}})U(c_i, g_{ai}) + (1 - \delta^*(c_{\text{med}})) \frac{\partial U(c_i, g_{ai})}{\partial c_i} (\bar{c} - c_i).$$

Since $\delta^{*'} < 0$, $V'_i(0) > 0$ iff $|\delta^{*'}(c_{\text{med}})|(\bar{c} - c_{\text{med}})U(c_i, g_{ai}) > (1 - \delta^*(c_{\text{med}})) \frac{\partial U(c_i, g_{ai})}{\partial c_i} (c_i - \bar{c})$. Dividing by $\frac{\partial U(c_i, g_{ai})}{\partial c_i} c_i > 0$ and using $\nu(c_i, g_{ai}) = U(c_i, g_{ai})/(\frac{\partial U(c_i, g_{ai})}{\partial c_i} c_i)$ and dividing by $1 - \delta^*(c_{\text{med}})$ gives the condition. $\square$

Derivation of (8) and (12). The identity (8) multiplies and divides $U(c, g_0, \mu_0)/U(c, g_{ai}, \mu_{ai})$ by $U(c, g_{ai}, \mu_0)$. For the closed form, from the closed form of U , using $c^{1-\gamma}/(\gamma - 1) = \bar{u}/[(\gamma - 1)\bar{u}c^{\gamma-1}] = \bar{u}/[(\gamma - 1)\nu(c) + 1]$,

$$U(c, g, \mu) = \frac{\bar{u}}{\rho + \mu} - \frac{c^{1-\gamma}}{(\gamma - 1)(\rho + \mu + (\gamma - 1)g)} = \frac{\bar{u}}{(\rho + \mu)[(\gamma - 1)\nu(c) + 1]} \left[(\gamma - 1)\nu(c) + 1 - \frac{\rho + \mu}{\rho + \mu + (\gamma - 1)g} \right].$$

Take the ratio of the status-quo to the AI path: $\bar{u}$ and $(\gamma - 1)v(c) + 1$ cancel, the prefactors give $(\rho + \mu_{ai})/(\rho + \mu_0)$, and the brackets give

$$\delta^*(c) = 1 - \frac{\rho + \mu_{ai}}{\rho + \mu_0} \cdot \frac{(\gamma - 1)v(c) + 1 - \frac{\rho + \mu_0}{\rho + \mu_0 + (\gamma - 1)g_0}}{(\gamma - 1)v(c) + 1 - \frac{\rho + \mu_{ai}}{\rho + \mu_{ai} + (\gamma - 1)g_{ai}}}, \quad (12)$$

Write X for $(\gamma - 1)v(c) + 1$ and A_0, A_1 for the two fractions in (12), here and in the proof of Proposition 7 only. By the same computation with (g_{ai}, μ_0) and (g_{ai}, μ_{ai}) , the longevity factor in (8) equals $\frac{\rho + \mu_{ai}}{\rho + \mu_0} \cdot \frac{X - A_{10}}{X - A_1}$ with $A_{10} = \frac{\rho + \mu_0}{\rho + \mu_0 + (\gamma - 1)g_{ai}}$. Since $\frac{d}{dX} \frac{X - A_{10}}{X - A_1} = \frac{A_{10} - A_1}{(X - A_1)^2}$ and $X = (\gamma - 1)\bar{u}c^{\gamma-1}$ is increasing in c for $\gamma > 1$ and $\bar{u} > 0$, the factor is constant in c if and only if $A_{10} = A_1$, that is $g_{ai} = 0$ given $\mu_{ai} < \mu_0$, and strictly decreasing in c if and only if $A_{10} > A_1$, that is $g_{ai} > 0$.

Proof of Proposition 7. (i) Bringing the two fractions in (12) to a common denominator,

$$A_0 - A_1 = \frac{(\gamma - 1)(\rho + \mu_0)(\rho + \mu_{ai}) \left[\frac{g_{ai}}{\rho + \mu_{ai}} - \frac{g_0}{\rho + \mu_0} \right]}{(\rho + \mu_0 + (\gamma - 1)g_0)(\rho + \mu_{ai} + (\gamma - 1)g_{ai})},$$

which is positive when $g_{ai} > g_0$ and $\mu_{ai} \leq \mu_0$. As $X \rightarrow \infty$ the ratio in (12) tends to 1. Moreover, expanding (12),

$$\delta^*(c) = \frac{(\mu_0 - \mu_{ai})X + (\rho + \mu_{ai})A_0 - (\rho + \mu_0)A_1}{(\rho + \mu_0)(X - A_1)}, \quad \delta^*(c) - \frac{\mu_0 - \mu_{ai}}{\rho + \mu_0} = \frac{(\rho + \mu_{ai})(A_0 - A_1)}{(\rho + \mu_0)(X - A_1)} > 0$$

when $A_0 > A_1$ and $X > A_1$, i.e. for $\gamma > 1$, $\bar{u} > 0$ and a positive bracket in (i), the second display following from the first by subtracting $\frac{(\mu_0 - \mu_{ai})(X - A_1)}{(\rho + \mu_0)(X - A_1)}$: the terms in X cancel and $(\mu_0 - \mu_{ai})A_1 - (\rho + \mu_0)A_1 = -(\rho + \mu_{ai})A_1$. So every agent's cut-off exceeds the floor, and any δ at or below it is accepted by all.

Let $\bar{X}$ and $X_{med} < \bar{X}$ denote $(\gamma - 1)v(\cdot) + 1$ at $\bar{c}$ and c_{med} , and $B = (\rho + \mu_{ai})A_0 - (\rho + \mu_0)A_1$ (this proof only); A_{10} is as in the derivation above. From the expansion above,

$$\frac{\delta^*(c_{med})}{\delta^*(\bar{c})} = \frac{\bar{X} - A_1}{X_{med} - A_1} \cdot \frac{(\mu_0 - \mu_{ai})X_{med} + B}{(\mu_0 - \mu_{ai})\bar{X} + B}.$$

Without longevity ($\mu_{ai} = \mu_0$) the second factor is 1 and the first is $(\bar{X} - A_{10})/(X_{med} - A_{10})$. Since U is increasing in g and decreasing in μ for $\gamma > 1$, the hypothesis $U(c_{med}, g_0, \mu_0) > 0$ gives $U > 0$ on both paths for the median and the mean, so all four cut-offs are in $(0, 1)$ and $X_{med} > A_{10} > A_1$. (a) The second factor is strictly below 1: its numerator and denominator are the values of a strictly increasing affine function at $X_{med} < \bar{X}$, both positive because $\delta^*(c_{med}), \delta^*(\bar{c}) > 0$. (b) For $F(A) = (\bar{X} - A)/(X_{med} - A)$ on $A < X_{med}$, $F'(A) = \frac{-(X_{med} - A) + (\bar{X} - A)}{(X_{med} - A)^2} = \frac{\bar{X} - X_{med}}{(X_{med} - A)^2} > 0$; and

$A_1 < A_{10}$ because $\mu_{ai} < \mu_0$ lowers $\frac{\rho+\mu}{\rho+\mu+(\gamma-1)g_{ai}}$ when $g_{ai} > 0$. Hence $F(A_1) < F(A_{10})$. Multiplying (a) and (b), the ratio with longevity is strictly below the ratio without. For the monotonicity, rearrange the first factor of the ratio: $\frac{\delta^*(c_{med})}{\delta^*(\bar{c})} = \left[1 - \frac{\bar{X} - X_{med}}{\bar{X} + B/(\mu_0 - \mu_{ai})} \right] \cdot \frac{\bar{X} - A_1}{X_{med} - A_1}$, where $B = (\rho + \mu_0)(\rho + \mu_{ai}) \left[\frac{1}{\rho + \mu_0 + (\gamma-1)g_0} - \frac{1}{\rho + \mu_{ai} + (\gamma-1)g_{ai}} \right]$ is nonnegative if and only if $\mu_0 - \mu_{ai} \leq (\gamma-1)(g_{ai} - g_0)$. As μ_{ai} falls, $\mu_0 - \mu_{ai}$ rises, $\rho + \mu_{ai}$ falls and the bracket falls, so on that range $B/(\mu_0 - \mu_{ai})$, a product of three nonnegative nonincreasing factors, falls, and with it the first factor. The second factor is $F(A_1)$ with F increasing, by (b), and $A_1 = (\rho + \mu_{ai})/(\rho + \mu_{ai} + (\gamma-1)g_{ai})$ increasing in μ_{ai} , so it falls strictly as μ_{ai} falls. The ratio is therefore strictly decreasing in the mortality gain on that range.

(ii) Repeat the proof of Proposition 1 with the mortality of each path in its weight: after the change of variable $\ell = g_j a$, $U(c, g_j, \mu_j) = \frac{1}{g_j} \int_0^\infty e^{-\frac{\rho+\mu_j}{g_j}\ell} u(c e^\ell) d\ell$, so the weights are $e^{-\frac{\rho+\mu_j}{g_j}\ell} u'(c e^\ell) c e^\ell$ and their ratio is $\exp\left\{\left[\frac{\rho+\mu_0}{g_0} - \frac{\rho+\mu_{ai}}{g_{ai}}\right]\ell\right\}$, strictly increasing if and only if $\frac{g_{ai}}{\rho+\mu_{ai}} > \frac{g_0}{\rho+\mu_0}$; Steps 2–3 of that proof then apply, with the reverse ordering reversing every inequality; part (i) of that proof needs $U_0 > 0$ only. In the equality case the integrals $\int_0^\infty e^{-\frac{\rho+\mu}{g}\ell} u(c e^\ell) d\ell$ coincide on the two paths, so $U(c, g_0, \mu_0)/U(c, g_{ai}, \mu_{ai}) = g_{ai}/g_0 = (\rho + \mu_{ai})/(\rho + \mu_0)$ and the cut-off is the floor. $\square$

Derivation of the tolerated cost (9) and its slope. Adoption requires $(1 - \delta'_0)U((1-s)c_0, g_0) \geq (1 - \delta_0)U(c_0, g_0)$; the left side decreases in s , so the rule is $s \leq s^*(c_0)$, where the cutoff solves $U((1-s^*)c_0, g_0) = \frac{1-\delta_0}{1-\delta'_0} U(c_0, g_0)$. Expanding $U((1-s)c_0, g_0) = U(c_0, g_0) - \frac{\partial U(c_0, g_0)}{\partial \ln c_0} s + o(s)$ in this equality gives $(\delta_0 - \delta'_0)U(c_0, g_0) = (1 - \delta'_0) \frac{\partial U(c_0, g_0)}{\partial \ln c_0} s^* + o(s^*)$, hence, by (3),

$$s^*(c_0) \approx \frac{\delta_0 - \delta'_0}{1 - \delta'_0} v(c_0, g_0).$$

Differentiating $U((1-s^*)c_0, g_0) = \frac{1-\delta_0}{1-\delta'_0} U(c_0, g_0)$ with respect to c_0 , and writing $c_d = (1-s^*)c_0$ (this proof only), gives $\frac{\partial U(c_d, g_0)}{\partial c} \left[(1-s^*) - c_0 \frac{ds^*}{dc_0} \right] = \frac{1-\delta_0}{1-\delta'_0} \frac{\partial U(c_0, g_0)}{\partial c}$; solving and using $1-s^* = c_d/c_0$, $\frac{1-\delta_0}{1-\delta'_0} = U(c_d, g_0)/U(c_0, g_0)$ and $\frac{\partial U(c, g_0)}{\partial c} c = U(c, g_0)/v(c, g_0)$,

$$\frac{ds^*}{dc_0} = \frac{U(c_d, g_0)}{\frac{\partial U(c_d, g_0)}{\partial c} c_0^2} \left[\frac{1}{v(c_d, g_0)} - \frac{1}{v(c_0, g_0)} \right].$$

At the cutoff $U(c_d, g_0) = \frac{1-\delta_0}{1-\delta'_0} U(c_0, g_0) > 0$ for any agent who prefers life, so the prefactor is positive; under Assumption 1, $v(\cdot, g_0)$ is increasing (by (4) it is v with $\bar{u}$ scaled by a positive constant), so $v(c_d, g_0) < v(c_0, g_0)$ since $c_d < c_0$ and the bracket is positive. Hence $ds^*/dc_0 > 0$ exactly.

Proof of Proposition 8. Each agent favours defence iff $s \leq s^*(c_i)$, and s^* is increasing, as shown

in the derivation of (9), so the favour set $\{c_i : s \leq s^*(c_i)\}$ is an upper interval and the median is decisive: a majority adopts iff $s \leq s^*(c_{\text{med}})$ and the redistributive first best iff $s \leq s^*(\bar{c})$. Since $c_{\text{med}} < \bar{c}$ and s^* is increasing, $s^*(c_{\text{med}}) < s^*(\bar{c})$, delivering under-defence. $\square$

Proof of Proposition 9. Favouring defence is $(1 - (1 - \pi)\delta_0(c_i))U((1 - s)c_i, g_0) \geq (1 - \delta_0(c_i))U(c_i, g_0)$; expanding $U((1 - s)c_i, g_0) = U(c_i, g_0) - \frac{\partial U(c_i, g_0)}{\partial \ln c_i} s + o(s)$, to first order in (δ_0, s) this is $\pi\delta_0(c_i)U(c_i, g_0) \geq s \frac{\partial U(c_i, g_0)}{\partial \ln c_i}$, i.e. $\pi\delta_0(c_i)\nu(c_i, g_0) \geq s$ by (3), that is $s \leq s^*(c_i)$ with $s^* = \pi\delta_0\nu(\cdot, g_0)$. The favour set $\{c : s^*(c) \geq s\}$ is an upper interval, and the median decisive, iff s^* is monotone, and $\frac{d}{d \ln c} \ln s^* = \varepsilon_\nu + \varepsilon_{\delta_0}$ with ε_ν the elasticity of $\nu(\cdot, g_0)$, so monotone iff $\varepsilon_\nu \geq |\varepsilon_{\delta_0}|$; otherwise s^* is single-peaked and the decisive standard is $\text{med}_i s^*(c_i)$. Equalizing consumption to $\bar{c}$ and pooling exposure to $\bar{\delta}_0$ makes the planner's program the homogeneous one, so it defends iff $s \leq \pi\bar{\delta}_0\nu(\bar{c}, g_0)$ and the median iff $s \leq \pi\delta_0(c_{\text{med}})\nu(c_{\text{med}}, g_0)$; planner-yes/median-no is non-empty iff $\delta_0(c_{\text{med}})\nu(c_{\text{med}}, g_0) < \bar{\delta}_0\nu(\bar{c}, g_0)$, the displayed product. $\square$