Optimal Timing for Superintelligence

Mundane Considerations for Existing People[1]

(2026) Nick Bostrom
Working paper
version: 1.0
[www.nickbostrom.com]

Abstract

Developing superintelligence is not like playing Russian roulette; it is more like undergoing risky surgery for a condition that will otherwise prove fatal. We examine optimal timing from a person-affecting stance (and set aside simulation hypotheses and other arcane considerations). Models incorporating safety progress, temporal discounting, quality-of-life differentials, and concave QALY utilities suggest that even high catastrophe probabilities are often worth accepting. Prioritarian weighting further shortens timelines. For many parameter settings, the optimal strategy would involve moving quickly to AGI capability, then pausing briefly before full deployment: swift to harbor, slow to berth. But poorly implemented pauses could do more harm than good.

Introduction

Some have called for a pause or permanent halt to AI development, on grounds that it would otherwise lead to AGI and superintelligence, which would pose intolerable dangers, including existential risks. For instance, Eliezer Yudkowsky and Nate Soares argue in their recent book *If Anyone Builds It, Everyone Dies* that nations should enforce a global ban on advanced AI and the computational infrastructure to support it, and on research into improved AI algorithms.[2] These authors are extremely pessimistic about the prospects of aligned superintelligent AI, regarding its advent as an almost certain doom. In their view, creating superintelligence would be far worse than subjecting all of humanity to a universal death sentence.[3] Others have argued that even a much lower level of risk would warrant an indefinite moratorium on AI. Would it not be wildly irresponsible, they ask, to expose our entire species to even a 1-in-10 chance of annihilation?

[1] For comments, I'm grateful to Owen Cotton-Barratt, Max Dalton, Tom Davidson, Lukas Finnveden, Rose Hadshar, Fin Moorehouse, Toby Ord, Anders Sandberg, Mia Taylor, and Lizka Vaintrob.

[2] Yudkowsky & Soares (2025a). The authors propose a treaty of unlimited duration. Yet they seem to be in favor of *eventually* building superintelligence, after some presumably very long delay. They suggest the creation of a crack team of genetically engineered supergeniuses to help the planet safely navigate the transition (2025b).

[3] In the U.S., average survival time after an initial death sentence is about 22 years, and only 16% of death sentences are eventually carried out (Snell, T., 2021; Baumgartner et al., 2017).

However, sound policy analysis must weigh potential benefits alongside the risks of any emerging technology. Yudkowsky and Soares maintain that if anyone builds AGI, everyone dies. One could equally maintain that if *nobody* builds it, everyone dies. In fact, most people are already dead. The rest of us are on course to follow within a few short decades. For many individuals—such as the elderly and the gravely ill—the end is much closer. Part of the promise of superintelligence is that it might fundamentally change this condition.

For AGI and superintelligence (we refrain from imposing precise definitions of these terms, as the considerations in this paper don't depend on exactly how the distinction is drawn), the potential benefits are immense. In particular, sufficiently advanced AI could remove or reduce many other risks to our survival, both as individuals and as a civilization.

Superintelligence would be able to enormously accelerate advances in biology and medicine—devising cures for all diseases and developing powerful anti-aging and rejuvenation therapies to restore the weak and sick to full youthful vigor.[4] (There are more radical possibilities beyond this, such as mind uploading, though our argument doesn't require entertaining those.[5]) Imagine curing Alzheimer's disease by regrowing the lost neurons in the patient's brain. Imagine treating cancer with targeted therapies that eliminate every tumor cell but cause none of the horrible side effects of today's chemotherapy. Imagine restoring ailing joints and clogged arteries to a pristine youthful condition. These scenarios become realistic and imminent with superintelligence guiding our science.

Aligned superintelligence could also do much to enhance humanity's *collective safety* against global threats. It could advise us on the likely consequences of world-scale decisions, help coordinate efforts to avoid war, counter new bioweapons or other emerging dangers, and generally steer or stabilize various dynamics that might otherwise derail our future.

In short, if the transition to the era of superintelligence goes well, there is tremendous upside both for saving the lives of currently existing individuals and for safeguarding the long-term survival and flourishing of Earth-originating intelligent life. The choice before us, therefore, is not between a risk-free baseline and a risky AI venture. It is between different risky trajectories, each exposing us to a different set of hazards. Along one path (forgoing superintelligence), 170,000 people die every day of disease, aging, and other tragedies; there is widespread suffering among humans and animals; and we are exposed to some level of ongoing existential risk that looks set to increase (with the emergence of powerful technologies other than AI). The other path (developing superintelligence) introduces unprecedented risks from AI itself, including the possibility of catastrophic misalignment and other failure modes; but it also offers a chance to eliminate or greatly mitigate the baseline threats and misfortunes, and unlock wonderful new levels of flourishing. To decide wisely between these paths, we must compare their complex risk profiles—along with potential upsides—for each of us alive today, and for humanity as a whole.

With this in mind, it becomes clear (*pace* Hunt, Yampolskiy, and various other writers) that analogies likening AGI development to a game of Russian roulette are misplaced.[6] Yes, launching superintelligence entails substantial risk—but a better analogy is a patient with severe

[4] Cf. Freitas (1999), Bostrom (2014), and Amodei (2024).
[5] Sandberg, A. & Bostrom, N. (2008)
[6] E.g. Hunt, T & Yampolskiy, R. (2023) and Russell, S. (2024)

heart disease deciding whether to undergo risky surgery. Imagine a patient with advanced coronary artery disease who must weigh the immediate risk of bypass surgery against the ongoing risk of leaving the condition untreated. Without an operation, they might expect to live for several more months, with a gradually increasing daily risk of a fatal cardiac event. The risk of dying on any given day remains small, but it relentlessly accumulates over time. If they opt for surgery, they face a much higher risk of dying immediately on the operating table. However, if the procedure succeeds, the reward is many additional years of life in better health.

Whether the patient should undergo the operation, and if so *when*, depends on many variables—their tolerance for risk, their discount rate on future life years, whether a more skillful surgeon is likely to become available at some point, how much better their quality of life would be if the condition is cured, and so on. All these considerations have clear parallels in deciding whether and when to deploy transformative superintelligent AI.[7]

When we take both sides of the ledger into account, it becomes clear that our individual life expectancy is *higher* if superintelligence is developed reasonably soon. Moreover, the life we stand to gain would plausibly be of immensely higher quality than the life we risk forfeiting. This conclusion holds even on highly pessimistic "doomer" assumptions about the probability of misaligned AI causing disaster.

Evaluative framework

To analyze all the facets of our predicament is possibly infeasible—certainly too complex to attempt in a single paper. However, we can examine some of the tradeoffs through a few different lenses, each providing a view on some of the relevant considerations. By breaking the issue down in this way, we can clarify some aspects of the macrostrategic choices we face, even if a comprehensive evaluation remains out of reach.

One distinction that may usefully be made is between what we could term *mundane* and *arcane* realms of consideration. By the former we refer to the ordinary kinds of secular considerations that most educated modern people would understand and not regard as outlandish or weird (given the postulated technological advances). The latter refers to all the rest—anthropics, simulation theory, aliens, trade between superintelligences, theology, noncausal decision theories, digital minds with moral status, infinite ethics, and whatnot. The arcane is, in the author's view, relevant and important; but it is harder to get to grips with, and rolling it in upfront would obscure some simpler points that are worth making. In this paper, we therefore limit our purview to mundane considerations (leaving more exotic issues to possibly be addressed in subsequent work).[8]

[7] There may of course not be a specific moment at which "superintelligence is launched", but rather a more continuous and distributed process of incremental advances, deployments, and integration into the economy. But the structural considerations we point to in this paper can be seen more clearly if we consider a simplified model with a discrete launch event, and they should carry over to cases with more complicated deployment processes.

[8] Cf. Bostrom (2024)

Within either the mundane or arcane domain, we must also decide which evaluative standard to apply. In particular, we may distinguish between a person-affecting perspective, which focuses on the interests of existing people, and an impersonal perspective, which extends consideration to all possible future generations that may or may not come into existence depending on our choices. Individual mortality risks are salient in the person-affecting perspective, whereas existential risks emerge as a central concern in the impersonal perspective. In what follows, we adopt the person-affecting perspective (leaving an analysis from the impersonal perspective for future work).

We begin by introducing a very simple model. Subsequent sections will explore various complications and elaborations.[9]

A simple go/no-go model

Suppose that without superintelligence, the average remaining life expectancy is 40 years.[10] With superintelligence, we assume that rejuvenation medicine could reduce mortality rates to a constant level similar to that currently enjoyed by healthy 20-year-olds in developed countries, which corresponds to a life expectancy of around 1,400 years.[11] This is conservative, since superintelligence could also mitigate many non-aging causes of death—such as infectious diseases, accidents, and suicidal depression. It is also conservative because it ignores more radical possibilities (like mind uploading with periodic backup copies), which could yield vastly longer lifespans.[12]

---

[9] Previous work has mostly looked at the tradeoffs from the impersonal perspective. For example, Bostrom (2003) shows that even very long delays in technological development can theoretically be impersonally optimal if they lower existential risk. Hall & Jones (2007) point out that as societies get richer, the marginal utility of consumption falls rapidly while the value of additional life-years remains high, leading them to spend a larger fraction of GDP on life-extension (e.g. health spending). Jones (2016) argues that this "safety as a luxury good" mechanism can—depending on utility curvature—make it optimal to restrain economic growth or redirect innovation toward life-saving and safety. Aschenbrenner (2020) applies the mechanism to existential risk in a directed-technical-change model, suggesting that we are in a "time of perils" (advanced enough to build doomsday technologies but not yet rich enough to spend heavily on mitigation) and arguing that faster growth can shorten this phase and increase long-run survival even if it raises near-term risk. Binder (2021) presents a minimalist timing model trading accumulated background ("state") risk against one-off superintelligence ("transition") risk, with the optimum when the proportional rate of safety improvement equals the background hazard. Jones (2024) then studies a utilitarian planner choosing how long to run growth-boosting AI that carries a constant annual extinction risk; optimal run time is highly sensitive to diminishing marginal utility, and allowing AI-driven mortality reductions greatly increases tolerable cumulative risk. Houlden (2024) summarizes Jones and explores extensions adding non-AI growth and safety progress from pausing/investment.

[10] Global life expectancy at birth is roughly 73 years and the median age of the global population is a little over 30 years (United Nations, 2024): we round the difference to 40, for simplicity. In a later section we explore scenarios in which remaining life expectancy increases even without advanced AI.

[11] In developed countries, the annual mortality rate for healthy individuals aged 20–25 is approximately 0.05–0.08% per year, with most deaths in this age group attributable to external causes. If mortality were held constant at this rate throughout life, expected remaining lifespan would be approximately $1/0.0007 \approx$ 1,400 years. See, e.g., Arias et al. (2024) for U.S. actuarial life tables; similar figures obtain in other developed countries.

[12] Sandberg & Bostrom (2008), Moravec (1988)

Now consider a choice between never launching superintelligence or launching it immediately, where the latter carries an x% risk of immediate universal death. Developing superintelligence increases our life expectancy if and only if:

$$(1 - x) \cdot 1400 > 40 \quad \Rightarrow \quad x \lesssim 97\%$$

In other words, under these conservative assumptions, developing superintelligence increases our remaining life expectancy provided that the probability of AI-induced annihilation is below 97%.

More generally, let m_0 be the annual mortality hazard before AGI, and let m_1 be the hazard after a successful AGI launch. Assign positive quality-of-life weights q_0 and q_1 to life before and after AGI, respectively. Launching immediately increases (quality-adjusted) life expectancy for those alive today iff:

$$x \; < \; 1 - \frac{q_0 \, m_1}{q_1 \, m_0}$$

Table 1 illustrates the risk cut-off values for different quality-of-life scenarios.

TABLE 1: Acceptable AI-risk if post-AGI life expectancy is 1,400 years

Pre-AGI LE (y)	Post-AGI LE (y)	q_1/q_0	Max P_{doom}
40	1,400	1	97.1%
40	1,400	2	98.6%
40	1,400	10	99.7%

Table 2 shows the corresponding thresholds if the gain in life expectancy were only 20 years (so post-AGI life expectancy is 60 years instead of 40)—perhaps a case in which the underlying aging processes for some reason remain unaddressed.

TABLE 2: Acceptable AI-risk if post-AGI life expectancy is 60 years

Pre-AGI LE (y)	Post-AGI LE (y)	q_1/q_0	Max P_{doom}
40	60	1	33.3%
40	60	2	66.7%
40	60	10	93.3%

We observe that, from a mundane person-affecting perspective—even without a difference in quality of life and with very modest assumptions about superintelligence-enabled life extension—developing superintelligence now would *increase* expected remaining lifespan even with fairly high levels of AI risk.[13]

[13] Again, we're restricting the discussion to mundane facts and considerations. (Otherwise the expected remaining lifespan may be infinite both with and without AGI.)

Incorporating time and safety progress

The previous section treated the choice as binary: either launch superintelligence now or never launch it at all. In reality, however, we may instead face a timing decision. We may be able to make AGI safer by slowing its development or delaying its deployment, allowing further alignment research (and other precautions) to reduce the risk of catastrophic failure. This introduces a new tradeoff. Launching earlier means accepting a higher level of AI risk; launching later means extending the period during which people continue to die from ordinary causes and remain vulnerable to other background dangers.

This mirrors the medical analogy introduced earlier. A patient might postpone a risky operation in the hope that a safer method becomes available, but waiting exposes them to the ongoing risk of the underlying disease (and postpones their enjoying a state of improved health).

To formalize this idea (details in Appendix A), we assume that before AGI, individuals face a constant mortality hazard m_0; after a successful launch, this drops to a much lower value m_1. We also assume that the probability of catastrophic failure if AI is launched at time t declines gradually as safety work advances. The central question becomes: How long is it worth waiting for additional safety progress?

Table 3 shows representative "optimal waiting times" under different assumptions about the initial level of AGI risk and the (relative) rate at which that risk is reduced through further safety work. We include some perhaps unrealistically extreme values for initial P_{doom} (at $t = 0$) and rate of safety progress to get a sense of the full space of possibilities.

TABLE 3: Optimal delay for various initial risks and rates of safety progress

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Launch asap	Launch asap	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Launch asap	Wait 16.9 y	Wait 58.1 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 14.3 y	Wait 31.4 y	Wait 35.5 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Wait 8.1 m	Wait 9.4 y	Wait 13.8 y	Wait 15.5 y	Wait 15.9 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 6.8 m	Wait 2.6 y	Wait 3.9 y	Wait 4.6 y	Wait 4.8 y	Wait 4.9 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 8.2 m	Wait 1.3 y	Wait 1.7 y	Wait 1.9 y	Wait 2.0 y	Wait 2.0 y
Ultra-fast safety progress (99.0%/yr)	Wait 1.7 m	Wait 5.9 m	Wait 9.5 m	Wait 11.9 m	Wait 1.1 y	Wait 1.1 y	Wait 1.1 y

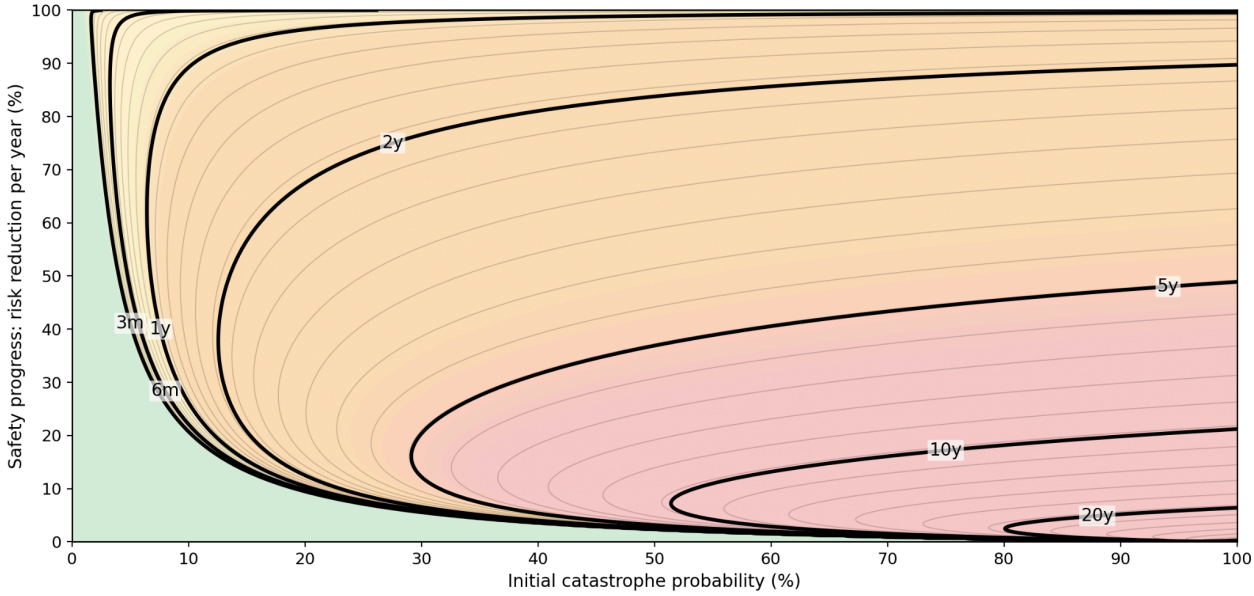
We observe a clear pattern. When the initial risk is low, the optimal strategy is to launch AGI as soon as possible—unless safety progress is exceptionally rapid, in which case a brief delay of a couple of months may be warranted. As the initial risk increases, optimal wait times become longer. But unless the starting risk is very high *and* safety progress is sluggish, the preferred delay remains modest—typically a single-digit number of years. The situation is further illustrated in Figure 1, which shows iso-delay contours across the parameter space.

Interestingly, both very fast and very slow rates of safety progress favor *earlier* launch. In the fast-progress case, the risk drops so quickly that there is no need to wait long. In the slow-progress case, waiting yields little benefit, so it is better to act sooner—while the potential

gains are still reachable for many. It is intermediate-to-slow progress rates that produce the longest optimal delays: just slow enough that safety improvements accumulate only gradually, but fast enough that waiting still buys some benefit. (There is also a corner case: if the initial risk is extremely high and safety improvements are negligible or non-existent, the model recommends never launching at all.)

If we measured outcomes in quality-adjusted life years (QALYs) rather than raw life-years, we would in most cases become even more impatient to launch. However, in the current model, this effect is modest. The prospect of reducing mortality to that of a healthy 20-year-old already dominates the tradeoff, making the value of the short pre-AGI period relatively insignificant by comparison. What drives the result is the balance between the risk of dying before AGI arrives, and the risk of dying because the launch goes wrong.

FIGURE 1: Iso-delay contours (cf. Table 3)



Temporal discounting

Thus far, we have assumed that future life-years are valued equally regardless of when they occur. In practice, decision-makers often apply a temporal discount rate, which downweights benefits that occur further in the future. Various pragmatic factors that are sometimes baked into an economic discount rate can be set aside here. For example, we should not use the discount rate to account for the fact that we may prefer to frontload good things in our lives on the ground that we might not be around to enjoy them if they are postponed far into the future (since we are modeling mortality risks separately). But decision-makers are sometimes supposed to also have a “pure time preference”, where they simply care less about what happens further into the future, and this is what we will examine here.

Discounting weakens the incentive to “rush” for the vast long-term life extension that successful AGI might bring. The enormous benefit of gaining centuries of expected life is no longer valued

at its full magnitude; whereas the risk of dying soon—either from a misaligned AGI or from current background hazards—remains at nearly full weight. As a result, introducing a discount rate shifts the optimal launch date later.

Table 4 illustrates the effect of a medium (3%) annual discount rate on optimal AGI timing. (Technical details appear in Appendix B, along with results for other discount rates.)

TABLE 4: Optimal delay with a 3% annual discount rate

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Never	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Wait 142.3 y	Wait 612.0 y	Wait 783.8 y	Wait 825.0 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 29.1 y	Wait 75.8 y	Wait 92.9 y	Wait 97.0 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Wait 2.6 y	Wait 11.3 y	Wait 15.8 y	Wait 17.4 y	Wait 17.8 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 7.5 m	Wait 2.6 y	Wait 3.9 y	Wait 4.6 y	Wait 4.9 y	Wait 4.9 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 8.2 m	Wait 1.3 y	Wait 1.7 y	Wait 1.9 y	Wait 2.0 y	Wait 2.0 y
Ultra-fast safety progress (99.0%/yr)	Wait 1.7 m	Wait 5.9 m	Wait 9.5 m	Wait 11.9 m	Wait 1.1 y	Wait 1.1 y	Wait 1.1 y

We see that some borderline cases shift from “launch immediately” to “wait a bit”; and cases that already warranted waiting now recommend longer delays. Higher discount rates would amplify this effect: if the far future counts for little, it makes sense to mostly focus on securing the near future.

Quality of life adjustment

One important hope is that developing superintelligence will not only extend life but also make it better. We can model this by assigning a quality weight q_0 to life before AGI and a higher weight q_1 to life after a successful AGI launch.

Table 5 shows optimal timing when post-AGI life is twice as good as current life ($q_1/q_0 = 2$) with a standard 3% discount rate. (See Appendix C for details and further illustrations.)

TABLE 5: Optimal delay: small quality difference ($q_1/q_0 = 2$)**, medium discount rate** ($\rho = 3\%$)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 122.2 y	Wait 294.0 y	Wait 335.2 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 27.1 y	Wait 44.2 y	Wait 48.3 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 6.7 y	Wait 11.1 y	Wait 12.8 y	Wait 13.2 y
Brisk safety progress (50.0%/yr)	Launch asap	Launch asap	Wait 1.9 y	Wait 3.2 y	Wait 3.9 y	Wait 4.2 y	Wait 4.2 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 5.7 m	Wait 1.1 y	Wait 1.5 y	Wait 1.7 y	Wait 1.8 y	Wait 1.8 y
Ultra-fast safety progress (99.0%/yr)	Wait 12.8 d	Wait 4.6 m	Wait 8.2 m	Wait 10.6 m	Wait 11.8 m	Wait 1.0 y	Wait 1.0 y

We can see that higher post-AGI quality expands the “launch asap” region, and shortens delays in the instances where waiting is optimal.

The magnitude of this shift is limited because the “launch-asap” risk bar—the level of AGI-risk below which it becomes optimal to launch immediately—is bounded above. This means that the quality-effect saturates: even arbitrarily large quality improvements cannot push all cases to immediate launch. Thus, if we postulated that post-AGI life would be 1,000 or 10,000 times better than pre-AGI life, this would not make much difference compared to more modest levels of quality improvement. Intuitively, once post-AGI life becomes sufficiently attractive (because of its length and/or quality), pre-AGI life contributes relatively little to the expected value of the future; and the chief concern then becomes maximizing the chance of actually reaching the post-AGI era—i.e. balancing the improvements in AGI safety that come from waiting against the accumulating risk of dying before AGI if the wait is too long.

Interestingly, the effect of temporal discounting can flip sign depending on the magnitude of the pre/post-AGI quality differential. When there is no quality differential, higher temporal discount rates always push towards launching *later*. However, when there is a quality differential that is sufficiently large, impatience penalizes delaying the onset of the higher-quality existence that would follow a successful superintelligence; and this pulls towards launching earlier. Consequently, while discounting always acts as a brake in the pure longevity model, it acts as an accelerator when the quality-of-life gap is sufficiently large.

Diminishing marginal utility

The preceding models have relied on a linear value assumption—essentially treating a 1,400-year lifespan as subjectively worth exactly 35 times as much as a 40-year lifespan. However, most people’s actual current preferences may exhibit diminishing marginal utility in quality-adjusted lifeyears (QALYs), meaning that e.g. a ten-year extension of a life that would otherwise be, say, 30 years is regarded as more desirable than a ten-year extension of a life that would otherwise be 1,390 years. Such a preference structure can also be viewed as a form of risk-aversion. Few people would accept a coin flip where “heads” means doubling their remaining lifespan and “tails” means dying immediately—and they may reject it even if we introduce a modest sweetener (such as a $10,000 reward or an additional bonus lifeyear if the coin lands heads).

We can model this using a standard diminishing-returns utility function—constant relative risk aversion (CRRA)—that introduces a curvature parameter, γ, representing the degree of risk-aversion. As this parameter increases, the decision-maker becomes more conservative, requiring higher probabilities of success (or greater potential upside) before betting their current life on a transformation.

Table 6 shows the results for $\gamma = 0.26$, a typical value derived from the empirical health-economics literature. Other parameters are the same as in the previous section. (See Appendix D for details and additional illustrations.)

TABLE 6: Diminishing marginal utility (CRRA, medium rate)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Wait 3.1 d	Wait 1.9 y	Wait 122.6 y	Wait 294.0 y	Wait 335.2 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Wait 4.2 d	Wait 4.4 y	Wait 31.7 y	Wait 46.3 y	Wait 50.1 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Wait 1.1 y	Wait 8.4 y	Wait 12.5 y	Wait 14.1 y	Wait 14.4 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 4.4 m	Wait 2.3 y	Wait 3.6 y	Wait 4.2 y	Wait 4.5 y	Wait 4.5 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 7.2 m	Wait 1.2 y	Wait 1.6 y	Wait 1.8 y	Wait 1.9 y	Wait 1.9 y
Ultra-fast safety progress (99.0%/yr)	Wait 1.2 m	Wait 5.4 m	Wait 9.0 m	Wait 11.3 m	Wait 1.0 y	Wait 1.1 y	Wait 1.1 y

Comparing this to Table 5, we see that diminishing marginal utility in QALYs leads to a somewhat more conservative approach: the zone of “launch asap” shrinks and optimal wait times increase. This effect is strongest for earlier dates. (See also Figure 2.)

FIGURE 2: Iso-delay contours (cf. Table 6)

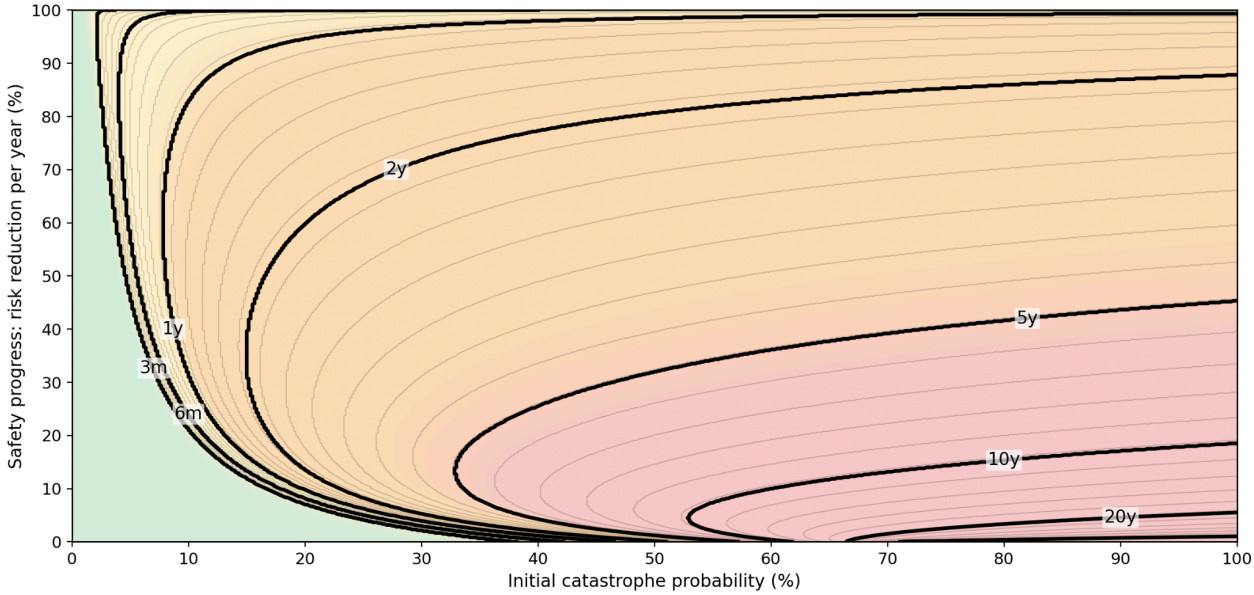


Table 7 shows what the risk is if launch occurs at the optimal time (for the same parameter settings as Table 6).

TABLE 7: Risk-at-launch (for the same model)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	1.0%	5.0%	20.0%	50.0%	Never	Never	Never
Glacial safety progress (0.1%/yr)	1.0%	5.0%	20.0%	49.9%	70.8%	70.8%	70.8%
Very slow safety progress (1.0%/yr)	1.0%	5.0%	20.0%	47.9%	58.1%	59.6%	59.9%
Moderate safety progress (10.0%/yr)	1.0%	5.0%	17.9%	20.6%	21.4%	21.6%	21.6%
Brisk safety progress (50.0%/yr)	1.0%	3.9%	4.1%	4.2%	4.2%	4.2%	4.2%
Very fast safety progress (90.0%/yr)	1.0%	1.3%	1.3%	1.3%	1.3%	1.3%	1.3%
Ultra-fast safety progress (99.0%/yr)	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%

These risk-at-launch values are somewhat—but not dramatically—reduced compared to those of a risk-neutral agent (except in cases where the risk-neutral agent would never launch or the risk-averse agent would launch asap, in which case risk-at-launch is the same for both).

Changing rates of safety progress

In the models considered so far, we assumed that AGI can be launched at any time, that background mortality remains constant until launch, that AI safety improves at a constant rate, and that no evidence about system safety is obtained beyond what that steady progress implies. In reality, however, we are not yet in a position to launch full AGI; background mortality risk could shift around the time AGI becomes available; the pace of safety progress is likely to vary across stages; and we may be able to run tests that provide direct information about whether a system is safe. We now explore how some of these factors affect the picture.

It is helpful to distinguish two timing variables:

- T_{agi}: the time from now until full AGI first becomes technically deployable. We will refer to this period as Phase 1.
- T_{pause}: any additional delay we choose after that point before deploying—a deliberate pause between AGI becoming available and being rolled out at scale. We will refer to such a period as Phase 2.

Launch thus occurs at time $T = T_{\text{agi}} + T_{\text{pause}}$.

In principle, one could try to choose both variables so as to maximize expected (discounted, quality-adjusted) life-years. In practice, T_{agi} may be harder to affect to a degree that makes a significant difference. It is largely determined by the inherent technical difficulty of attaining AGI-level capabilities and by investment choices currently driven by intense competitive dynamics; whereas T_{pause}, in at least some scenarios, may be more a matter of deliberate choice by company leaders or policymakers who at that juncture may be more focused on making macrostrategically sound deployment decisions. Furthermore, as we shall see, relatively small changes to T_{pause} plausibly make a bigger difference to expected outcomes than similarly small changes to T_{agi}.

Before considering joint optimization over both variables, therefore, let us examine a model in which only T_{pause} is subject to choice. Here we treat T_{agi} as exogenous and given by the scenario (0, 5, 10, or 20 years until AGI availability). We retain the notation and parameters from previous sections, including exponential time discounting and concave utility (both at their "medium" values unless otherwise noted).

A key feature of this multiphase setup is that the rate of safety progress need not be constant. Different stages of development offer different opportunities for progress, and the most tractable problems tend to be solved first.

During Phase 1—the period before full AGI is available—safety researchers must work without access to the systems that will ultimately matter most. They can study precursor systems, develop theoretical frameworks, and devise alignment techniques that seem likely to scale; but the exact algorithms and architectures that enable full AGI remain unknown, limiting what can be tested or verified. Safety progress during this phase is therefore likely to be moderate.

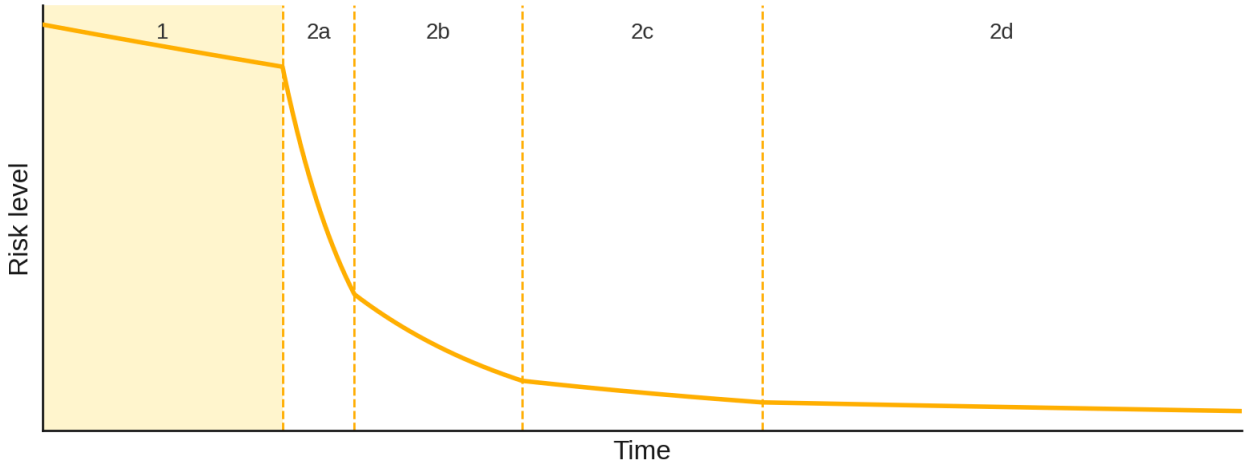
The situation changes once AGI-ready systems are attained. In Phase 2, researchers can study the actual system, run it in constrained environments, probe its behavior under controlled conditions, and potentially leverage the system's own capabilities to accelerate safety work. This suggests a burst of rapid safety progress immediately after AGI becomes available—a "safety windfall" from finally having the real artifact to work with.

Yet such rapid gains cannot continue indefinitely. The most promising interventions get explored first, and diminishing returns eventually set in. This motivates dividing Phase 2 into distinct subphases:

- *Phase 2a*: An initial period of very rapid safety progress. With the full system now available, researchers can perform interventions that were previously impossible—shaping the system, probing failure modes while slowly ramping capabilities, and implementing oversight mechanisms on the actual weights. This subphase is brief (perhaps weeks to months) but highly productive.
- *Phase 2b*: Continued fast progress, though slower than 2a. The most obvious low-hanging fruit has been picked, but researchers still benefit from working on the actual system, assisted by advanced AI tools. This might last around a year.
- *Phase 2c*: Progress slows to a rate similar to Phase 1, the benefits of having the actual system now roughly offset by the depletion of tractable problems. This subphase might last several years.
- *Phase 2d*: Ultimately progress becomes very slow, consisting of fundamental research into alignment science or the development of qualitatively new architectures. This continues indefinitely.

Figure 3 illustrates the qualitative picture. The key feature is that safety progress is front-loaded within Phase 2.

Figure 3. Qualitative picture of risk in a multiphase model



To make this concrete, Table 8 shows the optimal pause durations (from the start of Phase 2) for eight different scenarios. (For details, see Appendix E.)

TABLE 8: A multiphase model: several scenarios

#	Phase 1	2a	2b	2c	2d	1%	5%	20%	50%	80%	95%	99%
①	0y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 6.3 y	Wait 6.3 y	Wait 6.3 y
②	0y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 4.1 y	Wait 6.3 y	Wait 6.3 y	Wait 6.3 y
③	5y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 9.4 m	Wait 1.3 y	Wait 2.2 y	Wait 5.0 y	Wait 5.7 y
④	5y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Wait 1.5 m	Wait 3.6 m	Wait 1.3 y	Wait 3.0 y	Wait 4.5 y	Wait 4.9 y
⑤	10y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Wait 1.2 m	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y
⑥	10y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Launch asap	Wait 3.6 m	Wait 1.0 y	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y
⑦	20y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Launch asap	Wait 3.6 m	Wait 11.1 m	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y
⑧	20y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Launch asap	Launch asap	Wait 3.6 m	Wait 3.6 m	Wait 3.6 m	Wait 3.6 m

We see that for a wide range of initial risk levels and rates of safety progress, the optimal strategy is to implement a short pause once we enter Phase 2. If the “windfall” available in subphases 2a and 2b is significant, the optimal pause is often measured in months or a small number of years. Beyond that point, the safety benefits of further waiting tend to be outweighed by the continuing costs of mortality and temporal discounting.

If we instead consider jointly optimizing over both T_{agi} and T_{pause}—so that the decision-maker can choose how long Phase 1 lasts (up to the maximum given by each default scenario) and then also choose how long to pause after AGI-capability is attained—we get the results shown in Table 9. (For ease of comparison, the times are expressed relative to the point at which launch would have occurred “by default” in each scenario, i.e. if there were neither acceleration of Phase 1 nor any subsequent pause. For example, in scenario 4, where the default Phase 1 duration is 5 years, “Wait -3.7 y” means launch occurs 1.3 years after the beginning of Phase 1. Likewise, “launch asap” here denotes the time as it did previously, the point at which Phase 2 would have commenced by default.)

TABLE 9: Joint optimization over Phase 1 and Phase 2

#	Phase 1	2a	2b	2c	2d	1%	5%	20%	50%	80%	95%	99%
①	0y	0.3y@70%/y	1y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 6.3 y	Wait 6.3 y	Wait 6.3 y
②	0y	0.3y@70%/y	1y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 4.1 y	Wait 6.3 y	Wait 6.3 y	Wait 6.3 y
③	5y@5%/y	0.3y@70%/y	1y@25%/y	5.0y@5%/y	∞@2%/y	Wait -5.0 y	Wait -4.7 y	Wait -3.7 y	Wait -3.7 y	Wait 2.2 y	Wait 5.0 y	Wait 5.7 y
④	5y@10%/y	0.3y@70%/y	1y@25%/y	5.0y@10%/y	∞@2%/y	Wait -5.0 y	Wait -4.7 y	Wait -3.7 y	Wait -11.3 m	Wait 3.0 y	Wait 4.5 y	Wait 4.9 y
⑤	10y@5%/y	0.3y@70%/y	1y@25%/y	5.0y@5%/y	∞@2%/y	Wait -10.0 y	Wait -9.7 y	Wait -8.7 y	Wait -8.7 y	Wait -2.8 y	Launch asap	Wait 8.6 m
⑥	10y@10%/y	0.3y@70%/y	1y@25%/y	5.0y@10%/y	∞@2%/y	Wait -10.0 y	Wait -9.7 y	Wait -8.7 y	Wait -5.9 y	Wait -2.0 y	Wait -5.6 m	Wait -1.3 m
⑦	20y@5%/y	0.3y@70%/y	1y@25%/y	5.0y@5%/y	∞@2%/y	Wait -20.0 y	Wait -19.7 y	Wait -18.7 y	Wait -18.7 y	Wait -12.8 y	Wait -10.0 y	Wait -9.3 y
⑧	20y@10%/y	0.3y@70%/y	1y@25%/y	5.0y@10%/y	∞@2%/y	Wait -20.0 y	Wait -19.7 y	Wait -18.7 y	Wait -15.9 y	Wait -12.0 y	Wait -10.5 y	Wait -10.1 y

We see that in many scenarios and for many initial levels of risk, if the decision-maker is free to jointly optimize over both AGI development time and subsequent pausing, it is optimal to launch earlier than would have happened by default: these are the cells with blue background. (In scenarios 1 and 2, acceleration is impossible since Phase 1 has zero duration.)

Additionally, there are several scenarios in which, although launch occurs in Phase 2 after some period of pausing, it is still optimal to accelerate to some extent in Phase 1: these are the cells that do not have blue background but do have blue borders. This can happen because the rate of risk reduction is faster in Phase 2a and 2b than during Phase 1. There is thus a special value in being able to pause for at least a short while after AGI-capability has been attained before deploying it; and it can be worth going faster through Phase 1 in order to harvest these rapid safety gains while still keeping the overall time until AGI deployment tolerably short.

Shifting mortality rates

We have been assuming a constant background mortality rate until the launch of AGI, yet it is conceivable that it could change around the time when AGI-capability is attained (but before it is fully deployed).

Pessimistically, the world might become more dangerous with the introduction of near-AGI capabilities. For example, specialized AI systems could proliferate the capability to produce (new and more lethal) bioweapons, enable vast swarms of autonomous drones, precipitate mayhem by destabilizing our individual or collective epistemic systems and political processes, or raise geopolitical stakes and urgency in such a way as to trigger major war.

Optimistically, one might hope that near-AGI systems would enable breakthroughs in medicine that reduce mortality rates. However, substantial mortality reductions seem unlikely to materialize quickly, since many medical innovations must pass through extensive clinical trials and then require further time to achieve globally significant scale. Near-AGI systems could, of course, also have many other positive effects; yet except possibly for medical applications, it seems unlikely that they would have a big immediate impact on average death rates, since most people who are currently dying are succumbing to age-related and other medical issues.

On balance, therefore, if there is a dramatic change in global mortality just around the time when AGI becomes possible, it seems likelier to be for the worse than for the better. This adds to the reasons for keeping wait times relatively short after AGI-capability (or near-AGI capability that starts having dangerous applications) has been attained.

Yet if a medical breakthrough *were* to emerge—and especially effective anti-aging therapies—then the optimal time to launch AGI could be pushed out considerably. In principle, such a breakthrough could come from either pre-AGI forms of AI (or specialized AGI applications that don't require full deployment) or medical progress occurring independently of AI. Such developments are more plausible in long-timeline scenarios where AGI is not developed for several decades.

Note that for this effect to occur, it is not necessary for the improvement in background mortality to actually take place prior to or immediately upon entering Phase 2. In principle, the shift in optimal timelines could occur if an impending lowering of mortality becomes *foreseeable*; since this would immediately increase our *expected* lifespan under pre-launch conditions. For example, suppose we became confident that the rate of age-related decline will drop by 90% within 5 years (even without deploying AGI). It might then make sense to favor longer postponements—e.g. launching AGI in 50 years, when AI safety progress has brought the risk level down to a minimal level—since most of us could then still expect to be alive at that time. In this case, the 50 years of additional AI safety progress would be bought at the comparative bargain price of a death risk equivalent to waiting less than 10 years under current mortality conditions.

Table 10 shows the effects of postulating a precipitous drop in background mortality upon entering Phase 2—all the way to m_1, i.e. the rate that corresponds to a life expectancy of 1,400 years, same as what we have been assuming successful AGI would achieve. (Other parameters are the same as in Table 8; and we are assuming here that Phase 1 cannot be accelerated.)

TABLE 10: Pre-deployment mortality plummeting to 1/1400 (medium temporal discounting)

#	Phase 1	2a	2b	2c	2d	1%	5%	20%	50%	80%	95%	99%
①	0y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Wait 1.1 m	Wait 4.9 m	Wait 1.3 y	Wait 6.3 y	Wait 18.0 y	Wait 24.7 y	Wait 26.4 y
②	0y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Wait 1.1 m	Wait 4.9 m	Wait 3.3 y	Wait 6.3 y	Wait 8.9 y	Wait 14.5 y	Wait 15.9 y
③	5y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 6.3 y	Wait 7.4 y	Wait 13.6 y	Wait 15.1 y
④	5y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 6.1 y	Wait 6.3 y	Wait 6.3 y	Wait 6.3 y
⑤	10y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 1.5 y	Wait 6.3 y	Wait 6.3 y	Wait 6.3 y
⑥	10y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 11.2 m	Wait 1.3 y	Wait 5.2 y	Wait 6.3 y	Wait 6.3 y
⑦	20y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 9.8 m	Wait 1.3 y	Wait 1.3 y	Wait 2.5 y	Wait 3.3 y
⑧	20y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Launch asap	Wait 3.6 m	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y

We see that the optimal pause duration becomes longer—but not dramatically so. That the impact is fairly limited is due in part to safety gains being front-loaded, with diminishing returns arriving quickly after entering Phase 2. And in part it is due to the "medium"-level temporal discounting ($\rho = 3\%$) dominating the mortality rate.

Table 11 shows the same scenarios but with the "low" discount rate ($\rho = 1.5\%$). This does lead to longer wait times, especially in scenarios where the initial AI risk is so high that even after the sizable reductions during Phase 1 and Phases 2a–c, the level of risk remains too high for comfort.

TABLE 11: Pre-deployment mortality plummeting to 1/1400 (low temporal discounting)

#	Phase 1	2a	2b	2c	2d	1%	5%	20%	50%	80%	95%	99%
①	0y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Wait 3.6 m	Wait 1.3 y	Wait 5.1 y	Wait 14.9 y	Wait 33.8 y	Wait 41.2 y	Wait 43.0 y
②	0y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Wait 3.6 m	Wait 1.3 y	Wait 6.3 y	Wait 6.3 y	Wait 22.5 y	Wait 29.6 y	Wait 31.3 y
③	5y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 6.3 y	Wait 22.2 y	Wait 29.4 y	Wait 31.2 y
④	5y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Wait 1.6 m	Wait 4.6 m	Wait 3.2 y	Wait 6.3 y	Wait 6.3 y	Wait 7.8 y	Wait 9.3 y
⑤	10y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Wait 1.4 m	Wait 3.7 m	Wait 1.3 y	Wait 6.3 y	Wait 10.7 y	Wait 17.7 y	Wait 19.4 y
⑥	10y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 6.3 y	Wait 6.3 y	Wait 6.3 y	Wait 6.3 y
⑦	20y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 6.3 y	Wait 6.3 y	Wait 6.3 y
⑧	20y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Wait 1.1 m	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 2.2 y	Wait 2.6 y

Thus, if the background mortality risk is greatly reduced, then those with a low discount rate would be willing to wait a long time in order for AI risk to decline to a very low level. Note, however, that even if people stopped dying altogether, it could still be optimal to launch AGI eventually—and in fact to do so without extremely long delays—provided only there is a significant quality-of-life differential, a nontrivial temporal discount rate, and that AI safety continues to improve appreciably.

For contrast, Table 12 illustrates the situation for the opposite scenario, where mortality rates rise upon entering Phase 2. Unsurprisingly, this shortens optimal pause durations. The effect for the parameter-setting used in this table—a doubling of the mortality rate—is fairly modest. It would be more pronounced for greater elevations in the level of peril.

TABLE 12: Pre-deployment mortality rising to 1/20 (medium temporal discounting)

#	Phase 1	2a	2b	2c	2d	1%	5%	20%	50%	80%	95%	99%
①	0y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Wait 2.9 m	Wait 6.6 m	Wait 1.3 y	Wait 2.6 y	Wait 5.0 y	Wait 5.6 y
②	0y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Wait 2.9 m	Wait 6.6 m	Wait 1.3 y	Wait 4.8 y	Wait 6.3 y	Wait 6.3 y
③	5y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y
④	5y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y	Wait 1.7 y
⑤	10y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Launch asap	Wait 3.6 m	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y
⑥	10y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Launch asap	Wait 3.6 m	Wait 3.6 m	Wait 1.2 y	Wait 1.3 y	Wait 1.3 y
⑦	20y@5%/y	0.3y@70%/y	1.0y@25%/y	5.0y@5%/y	∞@2%/y	Launch asap	Launch asap	Wait 3.6 m	Wait 3.6 m	Wait 1.1 y	Wait 1.3 y	Wait 1.3 y
⑧	20y@10%/y	0.3y@70%/y	1.0y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Launch asap	Launch asap	Wait 3.0 m	Wait 3.6 m	Wait 3.6 m	Wait 3.6 m

Safety testing

AI safety work can provide at least two types of benefit: first, it can improve the nature of an AI system so that it is less likely to cause catastrophic harm if deployed; second, it can provide information about that nature, so that we can better judge whether to deploy it or to keep working to make it safer. The previous sections modeled both effects with a single parameter (the "rate of AI safety progress"). If we are willing to tolerate a more complicated setup, we can instead treat them separately. This leads to models where what is determined in advance is not an optimal launch time but an optimal *policy* that specifies—conditional on whatever safety information is then available—whether to launch or to continue working and testing.

To keep the setup manageable, we graft a simple testing process onto the multiphase model from the previous section. Once AGI-capable systems exist (the start of Phase 2), the true catastrophe probability at that time is unknown: it could be any of seven values, corresponding to the initial risk levels used earlier (1 %, 5 %, 20 %, 50 %, 80 %, 95 %, or 99 %). We assume a uniform prior over these possibilities. Safety work reduces the underlying risk over time following the same multiphase schedule as before: Phase 1 with moderate progress, followed (once AGI-capable systems exist) by a brief period of very rapid safety improvement (Phase 2a), a somewhat slower but still fast phase (2b), a medium-progress phase (2c), and then a long tail of very slow progress (2d).

Safety tests are triggered by safety progress rather than by clock time. Starting from the moment AGI-capable systems are available, a new test is performed every time safety work has reduced the system's intrinsic catastrophe probability by another 20 % relative to the last test. This reflects the idea that developing informative tests is itself part of safety work: as we make the system safer, we also learn how to probe it more effectively. If the underlying risk at the moment of testing is r, the test returns "fail" with probability r and "pass" with probability $1 - r$. Systems with very high intrinsic riskiness therefore tend to fail tests repeatedly, whereas fairly safe systems mostly pass—even if their remaining risk is still substantial. In particular, these tests usually cannot distinguish reliably between, say, ten and twenty per cent risk at launch; they are better at separating "clearly terrible" from "not obviously terrible".

We can formalize this setup as a partially observed Markov decision process (POMDP) and compute the optimal policy numerically (see Appendix G for details). Table 13 shows the expected delays (counting from the beginning of Phase 2).

TABLE 13: Periodic safety tests

#	Phase 1	2a	2b	2c	2d	1%	5%	20%	50%	80%	95%	99%
①	0y	0.3y@70%/y	1y@25%/y	5.0y@5%/y	∞@2%/y	Wait 1.4 y	Wait 1.7 y	Wait 2.7 y	Wait 4.9 y	Wait 7.3 y	Wait 8.6 y	Wait 8.9 y
②	0y	0.3y@70%/y	1y@25%/y	5.0y@10%/y	∞@2%/y	Wait 1.6 y	Wait 2.0 y	Wait 3.2 y	Wait 4.8 y	Wait 5.8 y	Wait 6.1 y	Wait 6.1 y
③	5y@5%/y	0.3y@70%/y	1y@25%/y	5.0y@5%/y	∞@2%/y	Wait 1.1 y	Wait 1.2 y	Wait 1.7 y	Wait 3.1 y	Wait 4.7 y	Wait 5.3 y	Wait 5.5 y
④	5y@10%/y	0.3y@70%/y	1y@25%/y	5.0y@10%/y	∞@2%/y	Wait 4.7 m	Wait 6.6 m	Wait 1.3 y	Wait 3.1 y	Wait 4.8 y	Wait 5.4 y	Wait 5.6 y
⑤	10y@5%/y	0.3y@70%/y	1y@25%/y	5.0y@5%/y	∞@2%/y	Wait 5.1 m	Wait 6.1 m	Wait 10.5 m	Wait 1.8 y	Wait 3.1 y	Wait 3.7 y	Wait 3.9 y
⑥	10y@10%/y	0.3y@70%/y	1y@25%/y	5.0y@10%/y	∞@2%/y	Wait 3.9 m	Wait 5.3 m	Wait 9.2 m	Wait 1.2 y	Wait 1.5 y	Wait 1.7 y	Wait 1.7 y
⑦	20y@5%/y	0.3y@70%/y	1y@25%/y	5.0y@5%/y	∞@2%/y	Wait 3.9 m	Wait 5.3 m	Wait 9.2 m	Wait 1.1 y	Wait 1.3 y	Wait 1.3 y	Wait 1.3 y
⑧	20y@10%/y	0.3y@70%/y	1y@25%/y	5.0y@10%/y	∞@2%/y	Launch asap	Launch asap	Wait 1.9 m	Wait 3.4 m	Wait 4.5 m	Wait 5.2 m	Wait 5.4 m

We observe that in most cases, the optimal policy results in an expected short (but greater-than-zero) delay, to take advantage of the rapid safety progress and concomitant opportunities gaining more information about the system's riskiness available in Phases 2a and 2b. Conditional on the system's initial riskiness being high when entering Phase 2, waiting times are longer; whereas when this is not the case, the optimal policy typically recommends launching within a year or two.

Note that Table 13 is not directly comparable to Table 8 (which represents the multiphase model analyzed earlier, the one most similar to the present model). This is because earlier we assumed that the decision-maker knew the initial riskiness of the system, whereas in the current model the agent starts out with a uniform probability distribution over the seven possible initial risk levels. If we want to pinpoint the difference that testing makes, we need to compare it to a baseline in which the agent starts out with the same agnostic distribution yet gains no further information from safety testing. Table 14 presents the result of such a comparison.

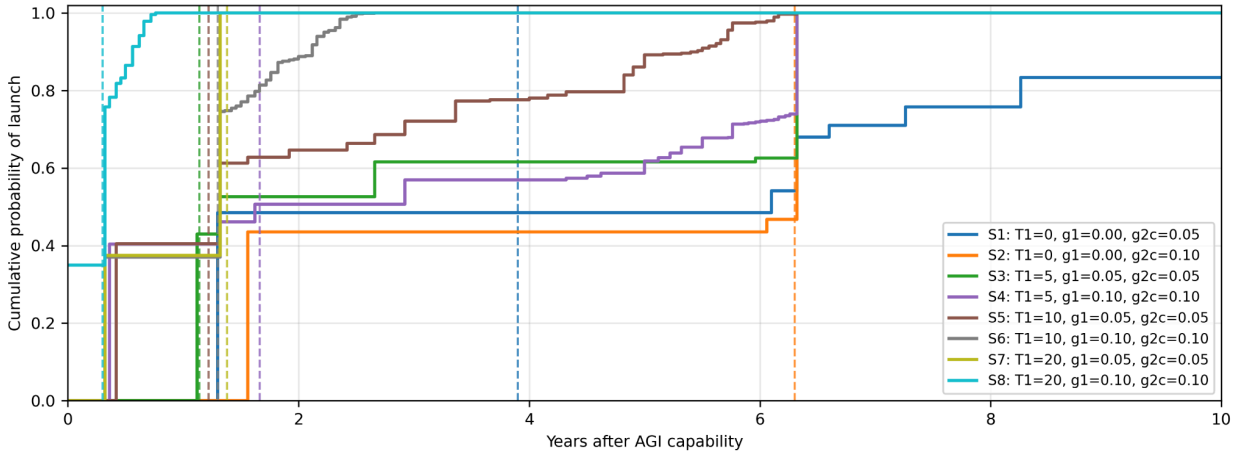
TABLE 14: Difference in outcomes from safety tests

#	Avg launch (no tests)	Avg launch (tests)	Δ wait	Risk (no tests)	Risk (tests)	Δ risk	Utility gain
①	3.90 y	5.05 y	+1.15 y	22.9%	20.6%	-2.2%	+3.58%
②	6.30 y	4.23 y	-2.07 y	15.4%	16.9%	+1.5%	+2.95%
③	1.30 y	3.23 y	+1.93 y	20.2%	17.3%	-2.9%	+1.31%
④	1.50 y	3.03 y	+1.53 y	15.1%	11.5%	-3.6%	+1.71%
⑤	1.30 y	2.05 y	+0.75 y	15.7%	14.8%	-0.9%	+0.37%
⑥	1.30 y	1.09 y	-0.21 y	9.1%	9.1%	+0.0%	+0.45%
⑦	1.30 y	0.93 y	-0.37 y	9.4%	9.6%	+0.3%	+0.28%
⑧	0.30 y	0.25 y	-0.05 y	4.2%	4.2%	+0.0%	+0.06%

We see that testing increases expected utility, sometimes by shortening the expected time-to-launch and sometimes by reducing the expected risk-at-launch. (That the expected utility gains look quite small in percentage terms is not particularly significant—this is driven by the infrequency and low sensitivity we assume of the tests and by other modeling assumptions. In reality, tests may also provide value by guiding future safety work in more productive directions.)

Figure 4 further illustrates how safety testing affects launch times. The dashed lines indicate where launches occur without safety testing (but with the agnostic prior over initial riskiness levels) for each of the eight scenarios. The solid lines show the cumulative probability distributions for the optimal policy with safety testing. We see that safety testing results in early launches in worlds where tests repeatedly pass, and later launches where tests keep failing and the posterior remains pessimistic.

FIGURE 4: Cumulative distribution functions of launch times with versus w/o safety tests



The main takeaway is that once system safety is uncertain, and future tests may provide information about how risky a system is, the relevant object is not a single optimal launch date but an optimal policy that conditions on evidence. Such a policy does something no fixed delay can do: it launches quickly when tests indicate the system is likely to be safe enough, but delays when tests reveal signs of danger. (The value of safety testing, however, depends not only on the quality of the tests themselves but—crucially—also on whether decision-makers are willing and able to let deployment decisions actually respond to what the tests reveal.)

Distributional considerations

We have analyzed the situation from the standpoint of the current world population as a whole. However, we need to acknowledge that the prudentially optimal timing for superintelligence is not the same for everyone.

One important factor of divergence is that people's mortality rates differ. Elderly people face a higher likelihood in the status quo of dying in the near future, while the young and hale could tolerate longer delays without accumulating an excessive risk of perishing before the main event.

Another factor is that those whose present quality of life is poor could rationally accept a higher risk of death for a shot at experiencing the great abundance and efflorescence that successful AGI would enable than those who are currently enjoying (what in present era is regarded as) a high standard of living.

There are therefore conflicts between different demographics over what is prudentially optimal regarding the timing of AGI. Other things equal, those who are old, sick, poor, downtrodden, miserable—or who have higher discount rates or less concave preferences over future quality-adjusted life years—should prefer earlier AGI launch dates compared to people who are comparatively satisfied and secure in the status quo.[14]

In the public policy literature, social welfare functions are often designed to include a prioritarian or egalitarian skew, such that a higher desirability is assigned (ceteris paribus) to outcomes in which the worst-off receive a given boost to their welfare than to ones in which a boost of equal magnitude is given to those who are already better-off.[15] If such priority is given to the worse off, and we combine this stipulation with the observations already made about the divergent prudential interests of different demographics, there may be implications for what is globally optimal regarding AI timelines.

In particular, the optimal timeline to superintelligence is likely shorter on a prioritarian view than it is on a neutral (person-affecting) utilitarian stance. This is partly because the worse off have less to lose and more to gain from rolling these dice. And partly it is because, in the case of the sick and the elderly, they have less ability to wait and roll the dice later when the odds may be more favourable. There is therefore a prioritarian argument for accelerating timelines beyond what the preceding analysis suggests.

Let us examine these issues a little more closely. One possible thought one might have is that the younger age structure in low-income countries would reduce the strength of the aforementioned prioritarian argument for shorter timelines, by introducing a correlation between being worse off and having longer remaining life expectancy—so that poor people in the developing world would have a prudential interest in longer AGI timelines compared to their better-off counterparts in rich countries. However, although the population does skew younger in poor countries, this is not enough to make up for the generally higher life expectancy in rich countries. The difference in life expectancy between rich and poor countries—which can exceed 25 years at birth when comparing the richest and poorest nations—narrows considerably when calculated as a population-weighted average of remaining years, due to the younger age structure in poorer countries. However, it does not close, let alone reverse.[16] While some convergence in life expectancy between poor and rich countries might be expected to occur during the remaining lifetime of people living in poor countries, it still seems plausible that, on average, people who are currently economically unfortunate can also expect to die sooner under default conditions than people who are currently economically fortunate. This positive correlation

[14] These underlying factors in narrow prudential optimality need not be consistently reflected in stated preferences. One reason is that different groups may have different empirical beliefs, such as concerning how risky AGI is or how good post-AGI life would become. They may also have different beliefs about non-mundane considerations that are outside the scope of this investigation. Furthermore, people may care about other individuals—e.g. an old person with a beloved grandchild may prefer a less aggressive timeline than if they were concerned exclusively with their own prospects. People may also have non-person-affecting preferences, such as over possible future generations. Some might have idiosyncratic reasons (e.g. glory, profit, influence) for accelerating the creation of AGI. And people may of course also misunderstand what their rational interests are, or shade the expression of their preferences in light of social desirability norms.

[15] Parfit (1997)

[16] Cf. Sanderson & Scherbov (2005); Wrigley-Field & Feehan (2022)

between poverty and lower remaining life expectancy strengthens the prioritarian case for faster timelines (compared to the distribution-agnostic analysis of the preceding sections).

One may also regard lifespan itself as a contributing factor in how fortunate a person is, and hence—on a prioritarian view—in how strong a claim they have to marginal resources or weighting of their marginal interests in the context of social planning. There are several different possible ways in which lifespan-related variation could be taken to influence somebody's baseline welfare level:

i. *Remaining life years*. One might hold that (ceteris paribus) persons with more remaining life years are better off than those with fewer years left, since it seems unfortunate to be in a condition in which one is soon about to get sick and die.

If one adopts this stance, then the prioritarian skew towards shorter timelines would be amplified. This is because older people—whose interests favor shorter timelines—would be weighted more heavily by this metric, since it would adjudge them comparatively unfortunate in the status quo.

ii. *Life years already had*. One might hold that (ceteris paribus) persons who have lived longer are better off, on grounds that they have gotten to feast more on life.

If one adopts this stance, then the prioritarian skew would be pulled in the direction favoring longer timelines, since the metric implied by (ii) would tend to deem older people as better off and hence less deserving of marginal consideration. It would not necessarily pull it far enough to make the prioritarian favor longer timelines all things considered compared to a neutral (non-prioritarian) criterion, since there are other categories of badly-off people (aside from, supposedly, the young) and who may have interests that differentially benefit from shorter timelines.

However, in any case, (ii) seems like a mistaken way to reckon. Consider two persons, a 10-year-old and a 20-year-old, both of whom have a genetic condition from which they will die at age 30, unless they receive a therapy, of which only one dose is available—in which case they live to age 50. It seems implausible to maintain that the 10-year-old has a stronger claim to the therapy just because he hasn't lived as long as the 20-year-old. It seems more plausible that their claims are equally strong—or, if not, then perhaps that the 20-year-old has a stronger claim (as would be implied by (i)).

A more plausible way to capture whatever intuition might appear to support (ii) would be:

iii. *Total life years*. One might hold that (ceteris paribus) persons whose total lifespans are longer are better off, since their endowment of life is greater.

This would accord the 10-year-old and the 20-year-old in the previous example equal weight, since they have the same baseline length of lifespan. When coupled with a prioritarian ethic, stance (iii) results in greater weight being placed on the interests of those whose lives in the default condition would be shorter.

So whose lives would, absent AGI, be shorter: the lives of the old or the lives of the young? On the one hand, the old have already survived all the hazards that kill some people prematurely.

On the other hand, the young can expect to benefit from many decades of economic and medical progress which might prolong their lives. If we extrapolate recent rates of increases in life expectancy, in wealthy countries, we may get a U-shaped curve: younger people and the very oldest people have the longest total life expectancy, with the nadir occurring for those who are around age 80. (Intuitively: somebody who's a centenarian has already lived longer than a newborn is likely to do, while a child has an advantage over people who are in their forties because the child is very likely to make it to forty and then gets benefit from four more decades of medical progress.) Since there are many more people who are substantially younger than 80 than who are substantially older than 80, this means there is a positive correlation between youth and total life expectancy. Hence (iii) induces an overall prioritarian downweighting of the interests of the young in wealthy countries. This would shorten the optimal timeline to AGI. In poor countries, however, the relationship may be more complicated due to high infant mortality: newborns have low expected total lifespans; young adults, high expected total lifespans, older adults, lower expected total lifespans; and the very old, high expected total lifespans. Absent a detailed quantitative analysis, it is not obvious how that adds up.

If one expects a radical breakthrough in life extension will happen, even in the absence of AGI, x years from now, which will enable people to live very long lives, such as two hundred years (or even to attain longevity "escape velocity"), then a discontinuity is introduced whereby those who would live less than x years without AGI are comparatively a lot more unfortunate according to (iii) than those who without AGI have more than x years left to live. Those with less than x years left to live without AGI would thus have their interests upweighted in a prioritarian social welfare function. This would increase the shift towards shorter timelines being optimal, assuming that x is within the lifetime of at least some significant fraction of currently living people.

Note that these effects from prioritarian upweighting of those with shorter total life expectancy—or those with shorter remaining life expectancy, if we adopt stance (i)—are additional to the effect that results from whatever extra benefit there is to adding life years to otherwise short lives that stem directly from diminishing marginal utility in life years (or QALYs). In other words, there are *two* possible reasons for giving an extra life year to a short-lived person rather than to a long-lived person, which are analogous to two possible reasons for giving a hundred dollar bill to a poor person rather than to a rich person: first, the poor person may derive a greater benefit from the hundred dollars; and second, the poor person might be overall worse off than the rich person, and would therefore—on a prioritarian ethic—have a stronger claim to marginal benefits (such that even if we suppose that the rich person would derive an equally large benefit from the hundred dollar bill—perhaps they are out of cash and need a taxi home—it would still be better for it to go to the poor person).

Yet another possible stance on how life chronology could be a prioritarian weight-factor is that there is some specific number of life years—for instance, the traditional three-score-and-ten—such that it is bad for a person to die earlier than that yet not significantly better to live beyond it. The metaphor might be that a human is like a cup of limited capacity, and once it's been filled up with life there's no value to keep pouring.

iv. *Full cup*. One might hold that it is unfortunate for somebody to die before the age of approximately seventy, but somebody who lives much beyond seventy is not thereby significantly better off, since they've already had a full life.[17]

This stance would have four relevant implications. First, it would reduce the value of AGI success, because some of the supposed upside consisted of the (exponentially time-discounted) value of lifespans much longer than the currently typical one for humans. (However, another part of the upside—the prospect of a greatly improved quality of life—would remain important.) Second, it would tilt the prioritarian skew in favor of the young, since they are not guaranteed in the pre-AGI default condition to reach the "full cup" number of life years that the old have already attained, thus making the young count as more unfortunate, thus giving their interests (which favor longer timelines) greater weight. Third, it would increase the downside for the young of early AGI launch, since—unless the risk has been brought down to quite a low level—an AGI launch could amplify the threat that the young will fail to reach their normal allotment of years. And fourth, since this increased downside pertains exclusively to the young, whereas the old, according to (iv), have little to lose from an AGI launch as they are already home and dry, it would tilt prioritarian concern even further towards favoring the interests of the young. The upshot would be that optimal AGI timelines, if one adopted the "full cup" stance, would become significantly longer.

However, even if the "full cup" stance might have some prima facie appeal, it is plausible that the intuitions that appear to support it are rooted—at least in substantial part—in a conflation between chronological age and contingently associated circumstances of age. In contemporary settings, old age is associated with multimorbidity, declining capacities, loneliness, pain, loss of autonomy, a sense of being a burden, and bleak future prospects. It would hardly be remarkable if additional life years *under those conditions* have limited appeal to many.[18] This might lead one to believe that seventy years (or some "normal lifespan" in that neighborhood) is all we need to max out our utility function in life years. But the most it would really show is that in present circumstances we gain little from living much beyond that age. In other circumstances, we may gain a lot. In particular, if an AGI-breakthrough enables the restoration of full health and youthful vigor, and a return or even strengthening of our previously lost capacities—and pulls open the curtains to a long continued existence, together with friends and family who can also expect to stick around for a long time, in a world that is dawning on a new age, immeasurably richer, more promising, and teeming with marvels than any earlier era—*then* why should additional life years stop being valuable for somebody just because seventy life years have passed since they were born? In such a scenario, would we not rather all be like children again—with the potential before us so greatly outstripping our comparatively paltry past?

---

[17] This can be compared to what is known in the bioethics literature as the "fair innings" view; see e.g. Harris (1985) and Williams (1997). But the latter is often focused on a comparative fairness intuition—that somebody who fails to attain the normal lifespan has been "cheated" of their fair share of years, and that individuals who would not otherwise attain this normal lifespan should therefore get priority in the allocation of healthcare resources. That view would presumably entail that if it became normal for humans to live to 500, then the fair innings would increase correspondingly. By contrast, what I call the "full cup" stance alleges that there is much less value to a person of an extra life year once they have lived for about seventy years.

[18] Tsevat, J. et al. 1998

This suggests that we should reject the "full cup" stance as a fundamental evaluative principle, and specifically reject its application in the context of transformative AI, where many of the usual conditions of life years at old age are stipulated not to obtain. It is also worth noting that even under current (often very bad) conditions, those who seem best placed to judge the value of continued life at old age—namely, those who actually are in that situation and have first-hand knowledge of what it is like—often deny the stance and place a high value on remaining alive longer. For example, in one multicenter study of hospitalized patients aged 80+, more than two-thirds were willing to give up at most one month of a remaining year for "excellent health".[19] Surrogate decision-makers systematically underestimated their reluctance to trade away time. When patients who were still alive a year later were asked the same question again, they were willing to trade even less time for better health than at baseline.

We have focused on distributional considerations that are fairly directly tied to *when* AGI is developed. There are of course many other potentially important distributional considerations that arise in the context of AGI. For example, citizens of a country that leads AGI development might benefit more than citizens of other countries; and individuals who directly participate in a successful AGI launch might gain disproportionate profits and glory. Although *who* and *how* may be correlated in various ways to *when*, these broader distributional questions fall outside the scope of this paper.

Other-focused prudential concerns

A different set of considerations arises if we expand our conception of what might lie in the prudential interest of a person to include the welfare of other persons they strongly care about. For example, while it might be in the narrow self-interest of an old person for superintelligence to be launched very soon, they might prefer a somewhat delayed launch because they also care about their grandchildren who have a much longer remaining life expectancy under pre-AGI conditions than they themselves do.

However, if we take into account these kinds of preferences, we should also take into account preferences going in the other directions: younger people who, for their own part, might benefit from longer timelines yet may prefer somewhat shorter timelines because they care about others who are closer to dying. Just as we can love our children and grandchildren, we can also love our parents and grandparents. So this type of concern for kin might total up to roughly a wash.

With regard to caring for our friends (or admired strangers), it is likewise unclear which way the correlation goes between somebody's age and the number of people who care about them. The very old may have fewer people who care about them because many of their friends have already died; but the very young may also have fewer friends who care about them because they have not met many people yet or have not known them for long.

On a prioritarian view, including other-focused concerns among our prudential interests might induce a slight shift in the direction of longer timelines. Suppose we assume a symmetric degree of average care between the young and the old. Suppose, further, that the old are on average worse off than the young in the default condition (because of their shorter remaining and total life

[19] Tvesat, Dawson, Wu, et al. (1998)

expectancy); so that a prioritarian reckoning upweights the interests of the old in determining the optimal social policy. Then the prioritarian upweighting of the interests of the old means that the interests of those whom the old care about receive extra weight (relative to what they would get if we didn't include other-focused concerns in our conception of what is prudentially desirable for somebody). Since on average the people whom old people care about are younger than they are themselves, this would shift some emphasis towards younger people, whose interests are served by longer timelines. Any such effect, however, is quite subtle and second-order.

Theory of second best

We have thus far asked the question about the optimal timing for superintelligence (from a person-affecting perspective) in an abstracted way—as if the world had a knob for different dates and your job was to turn it to the correct setting. In reality, the situation is more complex. Nobody has full control over AGI timelines, and different actors have different preferences. The ideal timing may not be achievable, or might be achievable only through methods that would carry a significant risk of making the timing much worse than it would otherwise have been. Furthermore, interventions aimed at influencing when superintelligence arrives may have other important consequences besides their effect on timing. For these reasons, while the preceding discussion highlights some relevant background considerations, it does not on its own imply particular policy recommendations.

While a full policy analysis would require bringing into consideration many facts and arguments that are out of scope for this paper, it may be useful to briefly list some of the ways that an AI pause, or efforts to bring about such a pause, could have undesirable effects (aside from simply delaying the arrival of the benefits that successful AGI could bring):

- The pause occurs too early. People conclude that it was pointless, and become less willing to pause later when it would have been useful.
- The call for a pause results in poorly designed or incomplete regulation, producing safety theater that adds costs and bureaucracy and slows useful applications, while doing nothing to reduce the real risks. Compliance and box-ticking crowd out substantive work on risk reduction.
- A pause is implemented, but the developments it aims to forestall continue anyway—just elsewhere. Work may be driven underground, or shift towards less scrupulous actors or less cooperative states.
- The pause has an exemption for national security, pushing AI activities away from the civilian into the military sector. The result may be greater emphasis on destructive uses, lower transparency and democratic oversight, amplified AI-assisted coup risk or power concentration risk, and perhaps less competent alignment efforts.
- There are calls for a pause but they go unheeded—and no catastrophe occurs. Those who warned of danger are discredited, making it harder for future calls for AI safety work to be taken seriously.
- The push for a pause highlights the strategic importance of the technology, intensifying geopolitical AI competition.

- An international agreement is reached on pausing, but this creates a prisoner's dilemma in which some parties cheat (driving developments into covert programs) or triggers geopolitical conflict when some countries accuse others of cheating.
- A pause is implemented, leading to economic recession and general pessimism and lowered hopes for the future. People see the world more as a zero-sum battle for a limited set of resources, increasing conflict and tribalism.
- A pause prolongs the period during which the world is exposed to dangers from applications of already developed levels of AI (and to risks independent of AI), which more advanced AI could have helped mitigate.
- To enforce a pause, a strong control apparatus is created. The future shifts in a more totalitarian direction.
- There is a pause on AI development, yet progress in hardware and algorithm development continues. When the pause is eventually lifted, there is a massive compute and/or algorithm overhang that leads to explosive advances in AI that are riskier than if AI had advanced at a steadier pace throughout. The world will also not have had the opportunity to learn from and adapt to living with weaker AI systems. (Or in a more extreme case, the pause holds until dangerous models or superintelligence can be implemented on consumer-grade hardware, making it ungovernable.)
- Agitation for a pause leads to extremism. Some people become radicalized or violent. Attitudes towards AI become polarized to such an extent as to make constructive dialogue difficult and destroy the ability of institutions to pass nuanced adaptive safety policy.
- The push for a pause galvanizes supporters of AI to push back. Leading AI firms and AI authorities close ranks to downplay risk, marginalizing AI safety researchers and policy experts concerned with AI risk, reducing their resourcing and influence.
- A pause, initially sold as a brief moratorium to allow social adjustments and safety work to catch up, calcifies into a de facto permaban that prevents the immense promise of superintelligence from ever being realized—or is indefinitely extended without ever being formally made permanent.[20]

Of course, there are also some potentially positive side effects that might come from calls to bring about a pause even if they fail in their main aim. For example, they might lead to an increase in funding for AI safety work as a more acceptable alternative to pausing, or generally stimulate the world to more seriously prepare for AGI. Still, the potential ways that pausing or pushing for pausing could backfire are many and quite plausible.

[20] One mechanism that could theoretically produce indefinite extension is hyperbolic discounting. People often discount imminent consequences far more steeply than distant ones. Consider someone who resolves to swim but balks at entering the cold water; fitting exponential discounting to this behavior would require a rate on the order of 50% *per minute*—absurdly high for other contexts. Applied to AGI: when the launch date arrives, we think "not today", and repeat this reasoning each time the rescheduled date comes around. A structurally similar dynamic can arise even without individual time-inconsistency. If those with influence over deployment are always drawn from a demographic—e.g. neither very old nor very sick—that prudentially favors waiting decades, then when that future arrives, a new cohort may have taken their place with its own reasons for delay. While competitive pressures among multiple actors would probably prevent such indefinite procrastination, the dynamic becomes more concerning in scenarios involving a single dominant actor or a coordinated international body with broad discretion over timing.

The profile of potential upsides and downsides of a pause or delay looks different depending on the mechanics of implementation and the context in which it takes place. We have already touched on the idea that the safety benefit of a pause of a given duration seems likely to be much greater if it occurs at a late stage—ideally, once the capacity for AGI exists, and perhaps even a fully implemented system, yet prior to maximum scaleup or general deployment; since extra time for safety testing, oversight, and final adjustment may be especially impactful during that stage. The scope of and causal process inducing the pause is also relevant. Consider the following cases:

1. *Frontrunner unilaterally burning lead*. At the time when AGI becomes possible, one developer might have a technological lead over its competitors. It could choose to burn some or all of its lead to implement extra precautions while remaining ahead. This type of pause is relatively attractive, as it has less risk of producing many of the downsides listed above. It does not rely on the creation of a regulatory apparatus or enforcement regime, and it is less likely to result in a permanent abandonment of superintelligence. The pause is self-limiting, as it expires once a competitor catches up. If the case for additional safety precautions is very clear and strong, this competitor may also be persuaded to agree to halt (either unilaterally or in coordination with the frontrunner, perhaps with some nudging from the government), thus extending its duration. But eventually, as more competitors reach similar capability levels, the pause naturally expires. The scope for this kind of pause, however, is reduced in a highly competitive environment. At present, it is unclear who is ahead; and whatever lead they have is measured in a small number of months.
2. *Government-imposed moratorium*. This brings in more of the potential failure modes and side-effects that we listed. Risks of bureaucratization, militarization, self-coups, etc. are increased. The maximum duration of the pause is extended, and there is a greater risk that it would remain in place for longer than it ought to. It matters how the government action was brought about: if it is the result of technocratic pragmatics, the risk of it becoming too long or permanent is lower than if it comes about as a result of a general political anti-AI mobilization that stigmatizes the very idea of superintelligence. Instead of an outright moratorium, there could be regulation that permits the development and deployment of AGI only when safety standards have been met—this might be theoretically superior to an outright ban, but in practice it could be difficult to specify sensible criteria with enough precision.
3. *Internationally agreed prohibition*. Since this would involve state interventions, it would bring in many of the failure modes of a government-imposed moratorium. If the international agreement prohibits all development of new frontier systems, and includes effective verification provisions, it might avoid some of the risks (such as militarization and self-coups) that may be amplified in the case of individual government-imposed moratoria that have carveouts for national security applications. Other risks would be amplified, especially the risk that the moratorium ossifies into a permanent relinquishment of advanced AI, since in a tightly enforced global regime there would be no place where AI development could continue. The enforcement regime itself might also present some risk of eventually leading towards some sort of global totalitarian system. Yet without tight global enforcement, we would instead face the risks of selection effects, where AI development continues but only in the least cooperative states who refuse to join or in covert programs operated by defecting signatories. More limited international agreements on safety standards or short pauses might reduce some of these risks: for example, if AI projects in the U.S. and China are running neck-to-neck when dangerous AI

systems are about to be developed, there may be little opportunity for a unilateral pause (of the "frontrunner burning lead" type); but some pragmatic cooperation might be possible, in which both parties agree to suspend large training runs for a finite period of time (perhaps with provisions for inspectors to verify that their biggest AI centers are idle) to allow some additional time to work out critical safety issues before resuming.

These are the merest schematics. In reality, policymakers will confront a more complicated and textured set of options, subject to many practical constraints, and in which the effect on AI timelines is only one of many consequences that need to be factored into decisions. While some of the variables may be analyzed abstractly and ahead of time, much of the essential context will only become evident as developments unfold, and will require continuing judgment calls to adjust policies to an evolving situation.

The analysis of optimal AI timelines is relevant not only to questions of whether or not to bring about an AI pause but also to other policy choices that could impact the pace of AI development and deployment. For example, chip export restrictions, taxes on data centers, or employment laws that make it harder to lay off workers are possible measures that may be proposed or rejected mainly for reasons other than their impacts on AGI timelines. Nevertheless, they would likely retard AI progress on the margin; and so, in evaluating such policies, it would be useful to know whether that effect would be desirable or undesirable.

Conclusions

We have examined optimal timing for superintelligence from a person-affecting perspective, focusing on mundane considerations, leaving aside arcane considerations and impersonal perspectives for future work. A basic point here is that the baseline is not safe—not only because there are other catastrophic risks besides AI but also because of the high rate of individual sickness and death under the status quo. The appropriate analogy for the development of superintelligence is not Russian roulette but surgery for a serious condition that would be fatal if left untreated.

A simple go/no-go model illustrated how, if aligned superintelligence would enable major life extension and quality-of-life improvements, then even very high levels of P_{doom} can be worth incurring in terms of quality-adjusted life expectancy.

Note that P_{doom} here refers to the probability of AI causing human extinction.[21] The highest tolerable probability of misaligned superintelligence could be even higher—plausibly as high as 100% with the given assumptions—since it is far from certain that all humans would die if misaligned superintelligence is deployed.[22]

We then proceeded to explore a series of models in which the decision-maker has a richer option set involving when to deploy superintelligence, rather than just the binary choice between deploying it immediately or never. Waiting can reduce catastrophic risk through safety progress,

[21] In a binary scenario—more generally, we could take P_{doom} to be the expected fraction of the human population that dies when superintelligence is launched.
[22] See, e.g., Grace (2022), Christiano (2023a, 2023b), and Greenblatt (2025).

but incurs costs of ongoing mortality and foregone (or temporally discounted) benefits. A robust qualitative pattern emerges. Long waits are favored only when initial risk is very high *and* safety progress falls within a specific intermediate range—fast enough that waiting yields meaningful risk reduction, yet slow enough that the job isn't done quickly anyway. Outside this conjunction, optimal delays tend to be modest.

Various robustness checks shift recommendations in predictable directions without overturning the basic result. Simply adding temporal discounting pushes toward later launch by downweighting far-future benefits, though it rarely produces very long delays unless the rate is quite high. Adding quality-of-life uplift pushes toward earlier launch, though this effect saturates: once post-AGI life is sufficiently attractive, pre-AGI life contributes little to expected value, and the main concern becomes simply reaching the post-AGI era. When quality-of-life uplift is present, the effect of temporal discounting can be reversed: for sufficiently large quality-of-life differentials, temporal discounting pushes towards earlier launch, as impatience penalizes the delay of the onset of that higher-quality existence. Finally, diminishing marginal utility in quality-adjusted life years makes the decision-maker more conservative, shrinking the region where immediate or early launch is optimal—but even substantial risk aversion does not radically alter the overall picture.

A more elaborate model was then introduced, which featured two timing variables: time until AGI capability exists (Phase 1, perhaps largely driven by technical difficulty), and any deliberate pause before full deployment once capability is attained (Phase 2). This matters because the rate of safety progress is unlikely to be uniform across stages. Once a deployable system exists, there is plausibly a "safety windfall"—the ability to study, probe, and stress-test the actual artifact, and to leverage its own capabilities to accelerate alignment work. Yet such gains face diminishing returns as the most tractable problems are solved. The upshot is that time early in Phase 2 purchases more safety per unit than equivalent time earlier or later. The multiphase model often recommends short but non-zero pauses—months or a small number of years—once AGI-ready systems exist.

Background conditions around the time of AGI capability also matter. If near-AGI systems destabilize the world through bioweapon proliferation, autonomous weapons, epistemic corrosion, or geopolitical escalation, the cost of waiting rises, favoring short and purposeful post-capability pauses. Conversely, a major non-AGI mortality reduction—especially effective anti-aging therapies—would lower the cost of waiting, making longer postponements potentially optimal.

We also considered a variation of the multiphase model where system risk is uncertain and tests can provide information. This changes the object of evaluation from an optimal launch *date* to an optimal *policy*: launch when evidence looks sufficiently favorable, delay when it does not. Safety testing can shorten or lengthen expected wait times, and can increase or decrease risk at launch, but in either case increases expected utility.

Prudentially optimal timing varies across individuals. The elderly and the ill face higher near-term mortality in the status quo; those with poor quality of life have less to lose and more to gain from a transition to potential post-AGI abundance. Those who are old, sick, poor, or miserable should therefore generally prefer earlier launch dates than those who are comfortable and secure. If policy incorporates prioritarian weighting, this shifts the global optimum toward shorter delays.

Some intuitions about lifespan—such as the "full cup" notion that life-years beyond approximately seventy contribute little additional value—might push in the opposite direction; but we have argued such intuitions are plausibly misguided in a transformative-AI context, where many accustomed factors (such as the deprivations of old age) need not obtain.

These models have treated timing as if there were a simple knob to turn. In reality, no one has full control; different actors have different preferences; the ideal timing may be unachievable; and interventions aimed at influencing timelines have consequences beyond their effect on timing. Even if, in an abstract sense, a perfectly implemented pause before full superintelligence deployment would be desirable, there are numerous possible ways in which a bungled moratorium or other efforts to slow down AI developments could have bad effects in practice—for instance, by shifting developments to less regulated places, by increasing militarization, by creating hardware or algorithmic overhangs that ultimately make the AI transition more explosive, or by creating stigma and bureaucratization that risk ossifying into permanent relinquishment.

For these and other reasons, the preceding analysis—although it highlights several relevant considerations and tradeoffs—does not on its own imply support for any particular policy prescriptions. If nevertheless one wishes to compress the findings into a possible practical upshot, we might express it with the words *swift to harbor, slow to berth*: move quickly towards AGI capability, and then, as we gain more information about the remaining safety challenges and specifics of the situation, be prepared to possibly slow down and make adjustments as we navigate the critical stages of scaleup and deployment. It is in that final stage that a brief pause could have the greatest benefit.

Bibliography

Abellán-Perpiñán, J., Pinto-Prades, J., Méndez-Martínez, I. & Badía-Llach, X. (2006). "Towards a Better QALY Model". *Health Economics* 15(7): pp. 665–676.

Amodei, D. (2024). "Machines of Loving Grace: How AI Could Transform the World for the Better". https://www.darioamodei.com/essay/machines-of-loving-grace

Arias, E., Xu, J., Tejada-Vera, B. & Bastian, B. (2024). "U.S. State Life Tables, 2021". *National Vital Statistics Reports* 73(6). (National Center for Health Statistics: Hyattsville, MD). https://www.cdc.gov/nchs/data/nvsr/nvsr73/nvsr73-06.pdf

Aschenbrenner, L. (2020). "Existential Risk and Growth". *Global Priorities Institute Working Paper* No. 6-2020.
https://globalprioritiesinstitute.org/leopold-aschenbrenner-existential-risk-and-growth/

Baumgartner, F. et al. *Deadly Justice: A Statistical Portrait of the Death Penalty*. (Oxford University Press: New York, 2017)

Binder, D. (2021). "A Simple Model of AGI Deployment Risk". *Effective Altruism Forum* (9 July 2021).

https://forum.effectivealtruism.org/posts/aSMexrjGXpNiWpbb5/a-simple-model-of-agi-deployment-risk

Bleichrodt, H. & Pinto, J. (2005). “The Validity of QALYs under Non-Expected Utility”. *The Economic Journal* 115(503): pp. 533–550.

Bostrom, N. (2003). “Astronomical Waste: The Opportunity Cost of Delayed Technological Development”. *Utilitas* 15(3): pp. 308–314.

Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press: Oxford, 2014)

Bostrom, N. (2024). “AI Creation and the Cosmic Host”. *Working paper*. https://nickbostrom.com/papers/ai-creation-and-the-cosmic-host.pdf

Christiano, P. (2023a). “Comment on ‘But Why Would the AI Kill Us?’”. *LessWrong* (17 April 2023). https://www.lesswrong.com/posts/87EzRDAHkQJptLthE/but-why-would-the-ai-kill-us?commentId=sEzzJ8bjCQ7aKLSJo

Christiano, P. (2023b). “Comment on ‘Cosmopolitan Values Don’t Come Free’”. *LessWrong* (31 May 2023). https://www.lesswrong.com/posts/2NncxDQ3KBDCxiJiP/cosmopolitan-values-don-t-come-free?commentId=ofPTrG6wsq7CxuTXk

Freitas, R. *Nanomedicine: Volume 1: Basic Capabilities*. (Landes Bioscience: Austin, Texas, 1999)

Grace, K. (2022). “Counterarguments to the Basic AI Risk Case”. *World Spirit Sock Puppet* (14 October 2022). https://worldspiritsockpuppet.substack.com/p/counterarguments-to-the-basic-ai

Greenblatt, R. (2025). “Notes on Fatalities from AI Takeover”. *Unpublished manuscript*.

Hall, R. & Jones, C. (2007). “The Value of Life and the Rise in Health Spending”. *Quarterly Journal of Economics* 122(1): pp. 39–72.

Harris, J. *The Value of Life: An Introduction to Medical Ethics* (Routledge: London, 1985). Chapter 5.

Houlden, T. (2024). “‘The AI Dilemma: Growth vs Existential Risk’: An Extension for EAs and a Summary for Non-economists”. *Effective Altruism Forum* (11 November 2024). https://forum.effectivealtruism.org/posts/9zzGKfSdMeL7bGoPC/the-ai-dilemma-growth-vs-existential-risk-an-extension-for

Hunt, T. & Yampolskiy, R. (2023). “Building Superintelligence Is Riskier Than Russian Roulette”. *Nautilus* (2 August 2023). https://nautil.us/building-superintelligence-is-riskier-than-russian-roulette-358022/

Jones, C. (2016). “Life and Growth”. *Journal of Political Economy* 124(2): pp. 539–578.

Jones, C. (2024). “The A.I. Dilemma: Growth versus Existential Risk”. *American Economic Review* 6(4): pp. 575–590.

Moravec, H. *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press: Cambridge, MA, 1988)

Parfit, D. (1997). “Equality and Priority”. *Ratio* 10(3): pp. 202–221.

Russell, S. (2024). *Remarks at ITU AI for Good Summit Media Roundtable*, Geneva, 18 April. https://www.itu.int/hub/2024/04/moving-ai-governance-from-principles-to-practice/

Sandberg, A. & Bostrom, N. (2008). *Whole Brain Emulation: A Roadmap*. *Technical Report* 2008-3. Future of Humanity Institute, University of Oxford. https://ora.ox.ac.uk/objects/uuid:a6880196-34c7-47a0-80f1-74d32ab98788/files/s5m60qt58t

Sanderson, W. & Scherbov, S. (2005). “Average remaining lifetimes can increase as human populations age”. *Nature* 435(7043): pp. 811–813.

Snell, T. (2021). *Capital Punishment, 2020—Statistical Tables*. NCJ 302729. Washington, DC: U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Tsevat, J., Dawson, N., Wu, A., et al. (1998). “Health Values of Hospitalized Patients 80 Years or Older”. *JAMA* 279(5): pp. 371–375.

United Nations, Department of Economic and Social Affairs, Population Division (2024). *World Population Prospects 2024*. https://population.un.org/wpp/

Williams, A. (1997). “Intergenerational Equity: An Exploration of the ‘Fair Innings’ Argument”. *Health Economics* 6(2): pp. 117–132.

Wrigley-Field, E. & Feehan, D. (2022). “In a stationary population, the average lifespan of the living is a length-biased life expectancy”. *Demography* 59(1): pp. 207–220.

Yudkowsky, E. & Soares, N. (2025a). *If anyone builds it, everyone dies: Why superhuman AI would kill us all*. (Little, Brown and Company: New York)

Yudkowsky, E. & Soares, N. (2025b). “Why would making humans smarter help?” *If Anyone Builds It, Everyone Dies*. [Supp. online material] https://ifanyonebuildsit.com/13/why-would-making-humans-smarter-help

Appendix A: Details for the “timing and safety progress” model

Let t denote the AGI launch time.

The pre-AGI annual mortality hazard is set to correspond to an average remaining life expectancy of 40 years. This yields a continuous hazard rate of:

$$m_0 = \frac{1}{40} = 0.025$$

If AGI is launched successfully, mortality is assumed to fall to a much lower value, corresponding to a life expectancy of 1,400 years:

$$m_1 = \frac{1}{1400} \approx 0.000714$$

The probability of catastrophic failure at launch declines with safety progress. If initial catastrophic risk at $t = 0$ is p_0 and safety improves at annual fractional rate g, then the continuous decay rate is:

$$r = -\ln(1 - g)$$

and the launch-time catastrophe probability is:

$$P_{\text{doom}}(t) = p_0 e^{-rt}$$

Expected remaining life-years if AGI is launched at time t are:

$$E(t) = \frac{1 - e^{-m_0 t}}{m_0} + e^{-m_0 t}\left(1 - p_0 e^{-rt}\right)\frac{1}{m_1}$$

The optimal interior launch time is found by solving $E'(t) = 0$, yielding:

$$t^* = \frac{1}{r} \ln\left(\frac{p_0(m_0 + r)}{m_0 - m_1}\right)$$

If the expression inside the logarithm is less than or equal to 1, then $t^* = 0$, meaning immediate launch maximizes expected remaining life-years. A positive t^* exists only when initial catastrophic risk is high enough and safety improves fast enough that waiting reduces expected loss more than the background mortality accumulated during the delay.

Appendix B: Details for the “temporal discounting” model

To incorporate a constant pure time preference, we discount future life-years at rate ρ. The expected discounted remaining life-years as a function of the AGI launch time t is:

$$E_\rho(t) = \frac{1 - e^{-(m_0+\rho)t}}{m_0 + \rho} + e^{-(m_0+\rho)t}\big(1 - P_{\text{doom}}(t)\big)\frac{1}{m_1 + \rho}$$

where $P_{\text{doom}}(t) = p_0 e^{-rt}$ as in Appendix A.

Differentiating with respect to t and setting $E'_\rho(t) = 0$ gives the interior first-order condition:

$$(m_1 - m_0)e^{-(m_0+\rho)t} + p_0(m_0 + \rho + r)e^{-(m_0+\rho+r)t} = 0$$

which rearranges to the threshold equation:

$$p_0(m_0 + \rho + r) = (m_0 - m_1)e^{-rt}$$

Solving for t yields the optimal discounted launch time:

$$t^* = \frac{1}{r}\ln\left(\frac{p_0(m_0 + \rho + r)}{m_0 - m_1}\right)$$

If the expression inside the logarithm is less than or equal to 1, then $t^* = 0$, so immediate launch maximizes expected discounted life-years. A positive interior solution exists only when initial catastrophic risk is sufficiently high and safety improves sufficiently quickly that waiting reduces expected discounted loss more than the additional background mortality incurred during the delay costs.

Tables B1–B3 show the results for different values of the pure temporal discount rate (ρ).

TABLE B1: Low discount rate ($\rho = 1.5\%$)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 300.4 y	Wait 472.2 y	Wait 513.4 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 3.0 y	Wait 49.7 y	Wait 66.8 y	Wait 71.0 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Wait 1.7 y	Wait 10.4 y	Wait 14.9 y	Wait 16.5 y	Wait 16.9 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 7.1 m	Wait 2.6 y	Wait 3.9 y	Wait 4.6 y	Wait 4.8 y	Wait 4.9 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 8.2 m	Wait 1.3 y	Wait 1.7 y	Wait 1.9 y	Wait 2.0 y	Wait 2.0 y
Ultra-fast safety progress (99.0%/yr)	Wait 1.7 m	Wait 5.9 m	Wait 9.5 m	Wait 11.9 m	Wait 1.1 y	Wait 1.1 y	Wait 1.1 y

TABLE B2: Medium discount rate ($\rho = 3\%$)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Never	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Wait 142.3 y	Wait 612.0 y	Wait 783.8 y	Wait 825.0 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 29.1 y	Wait 75.8 y	Wait 92.9 y	Wait 97.0 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Wait 2.6 y	Wait 11.3 y	Wait 15.8 y	Wait 17.4 y	Wait 17.8 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 7.5 m	Wait 2.6 y	Wait 3.9 y	Wait 4.6 y	Wait 4.9 y	Wait 4.9 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 8.2 m	Wait 1.3 y	Wait 1.7 y	Wait 1.9 y	Wait 2.0 y	Wait 2.0 y
Ultra-fast safety progress (99.0%/yr)	Wait 1.7 m	Wait 5.9 m	Wait 9.5 m	Wait 11.9 m	Wait 1.1 y	Wait 1.1 y	Wait 1.1 y

TABLE B3: High discount rate ($\rho = 5\%$)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Never	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Wait 447.5 y	Wait 917.2 y	Wait 1089.0 y	Wait 1130.2 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 55.7 y	Wait 102.5 y	Wait 119.6 y	Wait 123.7 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Wait 3.8 y	Wait 12.5 y	Wait 16.9 y	Wait 18.5 y	Wait 18.9 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 7.9 m	Wait 2.7 y	Wait 4.0 y	Wait 4.7 y	Wait 4.9 y	Wait 5.0 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 8.3 m	Wait 1.3 y	Wait 1.7 y	Wait 1.9 y	Wait 2.0 y	Wait 2.0 y
Ultra-fast safety progress (99.0%/yr)	Wait 1.7 m	Wait 5.9 m	Wait 9.5 m	Wait 11.9 m	Wait 1.1 y	Wait 1.1 y	Wait 1.1 y

Appendix C: Details for the "quality-of-life-adjustment" model

We generalize the objective function to maximize expected discounted quality-adjusted life-years (QALYs). Let q_0 and q_1 be the quality of life before and after AGI, respectively. The expected value as a function of launch time t is:

$$E_{\rho,q}(t) = \int_0^t q_0 e^{-(m_0+\rho)s} ds + e^{-(m_0+\rho)t}(1 - P_{doom}(t)) \int_0^\infty q_1 e^{-(m_1+\rho)u} du$$

Defining constants $A = \frac{q_0}{m_0+\rho}$, $B = m_0 + \rho$, and $C = \frac{q_1}{m_1+\rho}$, the integrated form simplifies to:

$$E(t) = A(1 - e^{-Bt}) + Ce^{-Bt}(1 - p_0 e^{-rt})$$

Differentiating with respect to t and solving the first-order condition $E'(t) = 0$ yields the optimal risk threshold p^*:

$$p^* = \frac{B(C - A)}{C(B + r)}$$

The optimal launch time t^* is derived by solving $p_0 e^{-rt^*} = p^*$:

$$t^* = \frac{1}{r} \ln\left(\frac{p_0}{p^*}\right)$$

(If $p_0 \leq p^*$, then $t^* = 0$.)

The "launch asap" region expands as post-AGI quality increases, but it is bounded. As $q_1 \to \infty$ (implying $C \to \infty$), the threshold p^* approaches $\frac{B}{B + r} = \frac{m_0 + \rho}{m_0 + \rho + r}$. Thus, even for an infinite prize, immediate launch is optimal only if the current risk is lower than this ratio. If risk exceeds this bound, it remains optimal to wait, as the probability of success improves through safety progress (r) faster than the value of the prize diminishes through mortality and discounting ($m_0 + \rho$).

The tables below illustrate this model. We first look at the case where a post-AGI lifeyear has a quality that is twice as high as a pre-AGI lifeyear ($q_0 = 1, q_1 = 2$) for low, medium, and high discount rates.

TABLE C1: Small quality difference ($q_1/q_0 = 2$)**, low discount rate** ($\rho = 1.5\%$)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 20.2 y	Wait 192.0 y	Wait 233.2 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 21.9 y	Wait 39.0 y	Wait 43.1 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 7.7 y	Wait 12.2 y	Wait 13.8 y	Wait 14.2 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 2.3 m	Wait 2.2 y	Wait 3.5 y	Wait 4.2 y	Wait 4.4 y	Wait 4.5 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 6.7 m	Wait 1.2 y	Wait 1.6 y	Wait 1.8 y	Wait 1.8 y	Wait 1.9 y
Ultra-fast safety progress (99.0%/yr)	Wait 29.2 d	Wait 5.2 m	Wait 8.8 m	Wait 11.2 m	Wait 1.0 y	Wait 1.1 y	Wait 1.1 y

TABLE C2: Small quality difference ($q_1/q_0 = 2$)**, medium discount rate** ($\rho = 3\%$)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 122.2 y	Wait 294.0 y	Wait 335.2 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 27.1 y	Wait 44.2 y	Wait 48.3 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 6.7 y	Wait 11.1 y	Wait 12.8 y	Wait 13.2 y
Brisk safety progress (50.0%/yr)	Launch asap	Launch asap	Wait 1.9 y	Wait 3.2 y	Wait 3.9 y	Wait 4.2 y	Wait 4.2 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 5.7 m	Wait 1.1 y	Wait 1.5 y	Wait 1.7 y	Wait 1.8 y	Wait 1.8 y
Ultra-fast safety progress (99.0%/yr)	Wait 12.8 d	Wait 4.6 m	Wait 8.2 m	Wait 10.6 m	Wait 11.8 m	Wait 1.0 y	Wait 1.0 y

TABLE C3: Small quality difference ($q_1/q_0 = 2$), high discount rate ($\rho = 5\%$)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 202.6 y	Wait 374.4 y	Wait 415.6 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 31.4 y	Wait 48.5 y	Wait 52.6 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 5.7 y	Wait 10.1 y	Wait 11.8 y	Wait 12.1 y
Brisk safety progress (50.0%/yr)	Launch asap	Launch asap	Wait 1.6 y	Wait 3.0 y	Wait 3.6 y	Wait 3.9 y	Wait 3.9 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 4.6 m	Wait 11.8 m	Wait 1.4 y	Wait 1.6 y	Wait 1.7 y	Wait 1.7 y
Ultra-fast safety progress (99.0%/yr)	Launch asap	Wait 4.0 m	Wait 7.7 m	Wait 10.0 m	Wait 11.3 m	Wait 11.7 m	Wait 11.8 m

For comparison, let's also look at a version where post-AGI lifeyears are ten times as good as pre-AGI lifeyears ($q_0 = 1, q_1 = 10$). Table C4 shows the case for a median discount rate.

TABLE C4: Large quality difference ($q_1/q_0 = 10$), medium discount rate ($\rho = 3\%$)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Launch asap	Wait 24.2 y	Wait 65.4 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 2.6 m	Wait 17.3 y	Wait 21.4 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 4.1 y	Wait 8.6 y	Wait 10.2 y	Wait 10.6 y
Brisk safety progress (50.0%/yr)	Launch asap	Launch asap	Wait 1.5 y	Wait 2.8 y	Wait 3.5 y	Wait 3.8 y	Wait 3.8 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 4.3 m	Wait 11.5 m	Wait 1.4 y	Wait 1.6 y	Wait 1.6 y	Wait 1.7 y
Ultra-fast safety progress (99.0%/yr)	Launch asap	Wait 3.9 m	Wait 7.5 m	Wait 9.9 m	Wait 11.1 m	Wait 11.6 m	Wait 11.7 m

Appendix D: Details for the "diminishing marginal utility" model

To model risk aversion over (time-discounted quality-adjusted) lifespan, we employ two standard (one-parameter) utility functions from decision theory: Constant Relative Risk Aversion (CRRA) and Constant Absolute Risk Aversion (CARA).

1. Power Utility (CRRA)
The CRRA utility function—the one used in the main text—is defined as:

$$u(x) = \frac{x^{1-\gamma}}{1-\gamma}$$

where x represents the total discounted quality-adjusted life years (QALYs) and γ is the coefficient of relative risk aversion.

2. Exponential Utility (CARA)
The CARA utility function family takes the form:

$$u(x) = \frac{1 - e^{-\kappa x}}{\kappa}$$

3. Computation
For either functional form, we maximize the expected utility:

$$\mathbb{E}[u(X(t))] = P_{doom}(t) \cdot u(x_{fail}) + (1 - P_{doom}(t)) \cdot u(x_{success})$$

where:

$$x_{fail} = \frac{q_0}{m_0 + \rho}(1 - e^{-(m_0+\rho)t})$$
$$x_{success} = x_{fail} + e^{-(m_0+\rho)t}\frac{q_1}{m_1 + \rho}$$

4. Empirics
Direct estimates of utility for life duration in health-economics/decision-science settings have fit both power and exponential specifications. Exponential utility functions (CARA) for life duration have been directly estimated, but power utilities (CRRA) typically fit better.[23] We therefore treat power functions as the main specification, and include exponential function as a robustness check.

For $u(t) \propto t^{\alpha}$, estimates typically find $\alpha \approx 0.65$–0.85. From this derive $\gamma = 1 - \alpha$:

[23] Bleichrodt & Pinto (2005) estimate concave power and exponential forms for utility of life duration across health states. Abellán-Perpiñán et al. (2006) find that a power model predicts best overall.

- Low: $\gamma = 0.15$ (corresponding to $\alpha = 0.85$)
- Medium: $\gamma = 0.26$ (corresponding to $\alpha = 0.74$)
- High: $\gamma = 0.35$ (corresponding to $\alpha = 0.65$)

Because CARA exhibits constant absolute risk aversion, its relative risk aversion ($R(x) = \kappa x$) scales with the value of the outcome. To match the empirical literature and make a fair comparison, we calibrate κ such that the local relative risk aversion matches the CRRA medium case ($\gamma = 0.26$) at the scale of the post-AGI "prize" (in discounted QALYs):

$$X_{ref} \approx \frac{q_1}{m_1 + \rho} \approx 65.1$$

This yields $\kappa \approx \gamma / X_{ref} \approx 0.004$.

5. Illustrations

Tables D1–D3 illustrate optimal launch times for the CRRA model, for the low, medium, and high value of γ, respectively. (Other parameters are the same as in Appendix C.)

TABLE D1: Diminishing marginal utility (CRRA, low rate)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Wait 1.7 m	Wait 122.4 y	Wait 294.0 y	Wait 335.2 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 11.8 m	Wait 30.0 y	Wait 45.5 y	Wait 49.3 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Wait 1.8 m	Wait 7.7 y	Wait 11.9 y	Wait 13.5 y	Wait 13.9 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 2.0 m	Wait 2.1 y	Wait 3.4 y	Wait 4.1 y	Wait 4.3 y	Wait 4.4 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 6.5 m	Wait 1.1 y	Wait 1.5 y	Wait 1.7 y	Wait 1.8 y	Wait 1.8 y
Ultra-fast safety progress (99.0%/yr)	Wait 25.8 d	Wait 5.0 m	Wait 8.6 m	Wait 11.0 m	Wait 1.0 y	Wait 1.1 y	Wait 1.1 y

TABLE D2: Diminishing marginal utility (CRRA, medium rate—same as in main text)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Wait 3.1 d	Wait 1.9 y	Wait 122.6 y	Wait 294.0 y	Wait 335.2 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Wait 4.2 d	Wait 4.4 y	Wait 31.7 y	Wait 46.3 y	Wait 50.1 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Wait 1.1 y	Wait 8.4 y	Wait 12.5 y	Wait 14.1 y	Wait 14.4 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 4.4 m	Wait 2.3 y	Wait 3.6 y	Wait 4.2 y	Wait 4.5 y	Wait 4.5 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 7.2 m	Wait 1.2 y	Wait 1.6 y	Wait 1.8 y	Wait 1.9 y	Wait 1.9 y
Ultra-fast safety progress (99.0%/yr)	Wait 1.2 m	Wait 5.4 m	Wait 9.0 m	Wait 11.3 m	Wait 1.0 y	Wait 1.1 y	Wait 1.1 y

TABLE D3: Diminishing marginal utility (CRRA, high rate)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Wait 1.0 m	Wait 4.4 y	Wait 122.7 y	Wait 294.0 y	Wait 335.2 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Wait 1.3 m	Wait 7.1 y	Wait 33.0 y	Wait 47.0 y	Wait 50.6 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Wait 1.9 y	Wait 9.0 y	Wait 13.0 y	Wait 14.5 y	Wait 14.9 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 6.4 m	Wait 2.4 y	Wait 3.7 y	Wait 4.4 y	Wait 4.6 y	Wait 4.7 y
Very fast safety progress (90.0%/yr)	Wait 1.7 d	Wait 7.8 m	Wait 1.2 y	Wait 1.6 y	Wait 1.8 y	Wait 1.9 y	Wait 1.9 y
Ultra-fast safety progress (99.0%/yr)	Wait 1.5 m	Wait 5.7 m	Wait 9.2 m	Wait 11.6 m	Wait 1.1 y	Wait 1.1 y	Wait 1.1 y

Finally, Table D4 shows the corresponding medium case for the CARA utility function.

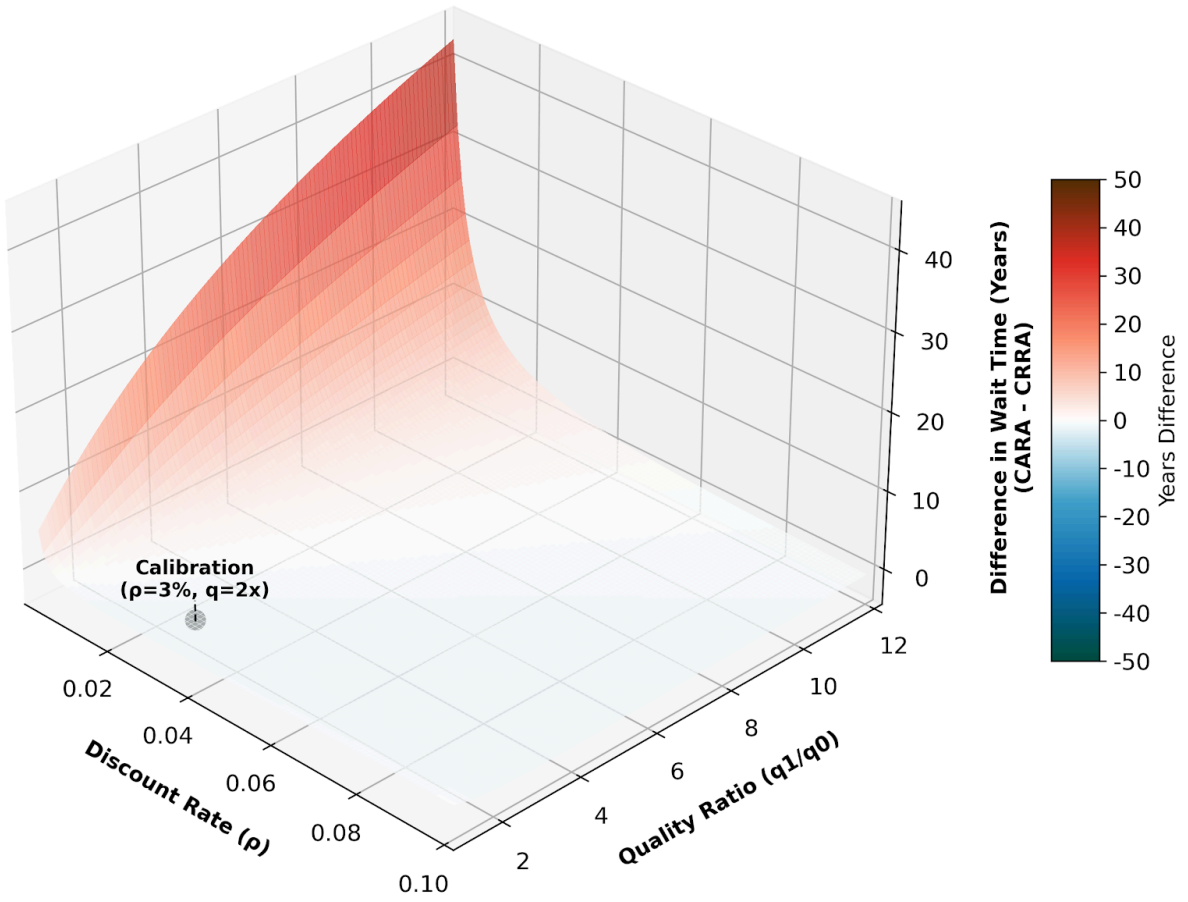
TABLE D4: Diminishing marginal utility (CARA, medium rate)

	1%	5%	20%	50%	80%	95%	99%
No safety progress (0.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Never	Never	Never
Glacial safety progress (0.1%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 122.3 y	Wait 294.0 y	Wait 335.2 y
Very slow safety progress (1.0%/yr)	Launch asap	Launch asap	Launch asap	Launch asap	Wait 28.8 y	Wait 44.9 y	Wait 48.8 y
Moderate safety progress (10.0%/yr)	Launch asap	Launch asap	Launch asap	Wait 7.4 y	Wait 11.7 y	Wait 13.3 y	Wait 13.7 y
Brisk safety progress (50.0%/yr)	Launch asap	Wait 1.2 m	Wait 2.1 y	Wait 3.4 y	Wait 4.1 y	Wait 4.3 y	Wait 4.4 y
Very fast safety progress (90.0%/yr)	Launch asap	Wait 6.3 m	Wait 1.1 y	Wait 1.5 y	Wait 1.7 y	Wait 1.8 y	Wait 1.8 y
Ultra-fast safety progress (99.0%/yr)	Wait 23.3 d	Wait 5.0 m	Wait 8.6 m	Wait 10.9 m	Wait 1.0 y	Wait 1.1 y	Wait 1.1 y

6. Comparison between CRRA and CARA

Both functional forms of diminishing marginal utility / risk-aversion in time-discounted QALYs delay launch relative to the risk-neutral case ($\gamma = 0$). Calibrated to the same reference scale and fit to the empirical literature, they give broadly similar timelines for the examined range of scenarios. However, because the relative risk aversion of CARA (κx) rises with scale, CARA can be significantly more conservative than CRRA in high-value regions (with low temporal discount factor and large quality differential). Figure 5 shows the difference surface between the two functions.

Figure 5: Difference between CARA and CRRA (for the medium rate case)



Appendix E: Details for the “changing rates of progress” model

The basic ingredients are the same as in Appendices A–D: a pre-AGI mortality hazard m_0, a post-AGI hazard m_1, a pure time-discount rate ρ, quality weights q_0 and q_1 for life before and after AGI, and CRRA utility over discounted QALYs with curvature parameter γ.

We distinguish two timing variables. Let T_{AGI} be the time from now until full AGI first becomes technically deployable (Phase 1), and let $T_{\text{pause}} \geq 0$ be any additional deliberate delay between that point and large-scale deployment (Phase 2). AGI is launched at time

$$T_L = T_{\text{AGI}} + T_{\text{pause}}$$

Let p_0 be the catastrophe probability if AGI were launched immediately. Safety work reduces this risk over time: over any sub-interval k in which the annual fractional reduction in risk is g_k, we define the corresponding continuous decay rate:

$$r_k = -\ln(1 - g_k)$$

If by time t we have spent $\Delta t_k(t)$ years in sub-interval k (capped at the maximum length of that sub-interval), the cumulative risk reduction is:

$$R(t) = \sum_k r_k, \Delta t_k(t)$$

The catastrophe probability at launch time t is then:

$$p(t) = p_0 \exp\big(-R(t)\big)$$

Phase 1 runs from time 0 to T_{AGI} with some baseline rate of safety progress. Once AGI-ready systems are available, we model a "safety windfall" by splitting Phase 2 into four subphases with front-loaded gains and diminishing returns: very rapid progress (2a), fast progress (2b), slower progress (2c), and an indefinitely long tail of very slow progress (2d). In each scenario, the first five columns ("Phase 1", "2a", "2b", "2c", "2d") of the table specify the duration and annual fractional improvement rate g_k used for these subphases.

For a given launch time T_L, let $x_{\mathrm{succ}}(T_L)$ denote the total discounted QALYs if AGI is successfully aligned at T_L, and let $x_{\mathrm{fail}}(T_L)$ denote the total discounted QALYs if launch at T_L causes catastrophe so that only pre-AGI life contributes.

With constant pre-AGI hazard m_0, post-AGI hazard m_1, and pure time discount rate ρ, the pre-AGI part is:

$$x_{\mathrm{fail}}(T_L) = \int_0^{T_L} q_0, e^{-(m_0+\rho)t}, dt = \frac{q_0}{m_0+\rho}\Big(1 - e^{-(m_0+\rho)T_L}\Big)$$

If launch succeeds at T_L, the post-AGI contribution is:

$$\int_{T_L}^{\infty} q_1, e^{-(m_0+\rho)T_L} e^{-(m_1+\rho)(t-T_L)}, dt = \frac{q_1}{m_1+\rho} e^{-(m_0+\rho)T_L}$$

so

$$x_{\mathrm{succ}}(T_L) = x_{\mathrm{fail}}(T_L) + \frac{q_1}{m_1+\rho} e^{-(m_0+\rho)T_L}$$

As in Appendix D, we use CRRA utility over discounted QALYs:

$$u(x) = \frac{x^{1-\gamma}}{1-\gamma}$$

The expected utility from launching at T_L is:

$$EU(T_L) = \big(1 - p(T_L)\big), u\big(x_{\text{succ}}(T_L)\big) + p(T_L), u\big(x_{\text{fail}}(T_L)\big)$$

In the multiphase timing table we treat T_{AGI} as fixed by the scenario (0, 5, 10, or 20 years until AGI availability). For each choice of initial catastrophe probability p_0 and each specification of baseline safety progress, we choose the pause length $T_{\text{pause}} \geq 0$ that maximizes

$$EU(T_L) = EU\big(T_{\text{AGI}} + T_{\text{pause}}\big)$$

The optimal T_{pause} is what is reported in Table 8.

Table 9 reports results when the decision-maker can also accelerate Phase 1. We allow T_{AGI} to be shortened by up to its full default duration (so that AGI could in principle become available immediately), while T_{pause} remains non-negative. The optimization problem becomes:

$$\max_{T_{\text{AGI}} \in [0, T_{\text{AGI}}^{\text{default}}],\, T_{\text{pause}} \geq 0} EU(T_{\text{AGI}} + T_{\text{pause}})$$

where safety progress during any acceleration of Phase 1 accrues at the Phase 1 rate, and the Phase 2 subphase structure (2a–2d) begins once AGI-capability is attained.

Appendix F: Details for the “shifting mortality rates” model

This extends the multiphase model of Appendix E by allowing the pre-AGI mortality hazard to change upon entering Phase 2. Let m_0 denote the mortality hazard during Phase 1, and let m_0' denote the hazard during Phase 2 (prior to launch). The discounted QALYs accumulated before launch become:

$$x_{\text{fail}}(T_L) = \frac{q_0}{m_0 + \rho}\left(1 - e^{-(m_0+\rho)T_{\text{AGI}}}\right) + \frac{q_0}{m_0' + \rho} e^{-(m_0+\rho)T_{\text{AGI}}}\left(1 - e^{-(m_0'+\rho)T_{\text{pause}}}\right)$$

The post-AGI contribution and catastrophe probability remain as in Appendix E.

Appendix G: Details for the “safety testing” model

We keep the background assumptions from Appendices E–F (mortality hazards, discounting, CRRA utility over discounted QALYs, and the four post-AGI subphases 2a–2d). At the moment AGI-capable systems first exist (start of Phase 2), the true catastrophe probability at that instant is unknown. It is known only that it equals one of seven discrete “type”, $p_{\text{init}} \in \{0.01, 0.05, 0.20, 0.50, 0.80, 0.95, 0.99\}$, with a uniform prior over these seven possibilities.

From that point onward, conditional on each type, the catastrophe probability at time t after AGI availability follows the same multiphase risk-reduction schedule as in Appendix E. For each type i this yields a deterministic risk path $p_i(t)$ with

$$p_i(t) = p_{\text{init}i} \exp(-R(t))$$

where $R(t)$ is the cumulative integrated rate implied by the phase-specific annual fractional reductions.

Starting from AGI availability, we perform a new test whenever cumulative risk reduction since the previous test reaches another 20 % factor. If the instantaneous risk at the time of a test is r, the test output is:

- "fail" with probability r
- "pass" with probability $1 - r$

Let π_i be the current posterior probability that the system is of type i and let r_i be the corresponding instantaneous risk $p_i(t)$ at the test time. After observing an outcome, we update by Bayes' rule. For a pass,

$$\pi_i' = \pi_i(1 - r_i)Z^{-1}$$

and for a fail,

$$\pi_i' = \pi_i r_i Z^{-1}$$

where Z is the normalisation constant that makes the posteriors sum to one.

Between tests, the posterior over types remains fixed, while each type's risk level declines deterministically according to the multiphase schedule.

We treat the problem from the start of Phase 2 as a finite-horizon POMDP. The state has two components:

1. Time within the multiphase schedule (which determines the phase and thus the risk-reduction rate)
2. Belief state π over the seven risk types

At each decision time (on a grid of size $\Delta t = 0.05$ years in the numerical implementation), the agent chooses between:

- *Launch now*: terminate the process and receive utility $U(t \mid \pi) = \sum_i \pi_i U(t \mid p_i(t))$, where $U(t \mid p)$ is the discounted-QALY objective from Appendices A–D for a launch at time t with catastrophe probability p.

- *Wait*: advance time by Δt (with deterministic change in the phase and risk levels) and, if a test is due, absorb the pass/fail signal and update the belief state by Bayes' rule as above.

We solve this POMDP numerically by backward induction over the discrete time grid, using the underlying survival-and-QALY value function from the earlier timing models for the "launch" payoff. The result is an approximately Bayes-optimal stationary policy mapping each time-belief pair to "launch" or "wait".

For comparison, we also compute the best fixed-pause policy with no testing. In that case, the agent chooses a single pause length τ after AGI availability, launches at τ in all worlds, and optimizes expected utility under the uniform prior over the seven types, exactly as in the multiphase model without testing.