Existential Risk and Growth in AI Races

Timothy Currie

09.02.2026

Draft. See *here* for latest version.

Abstract. I develop 2 models of an AI race. The first looks at how growth and risk interact in a race. Finds that if racing grows consumption agents will eventually stop racing once rich enough. But equilibrium is still often inefficient. In a race to develop powerful Artificial Intelligence teams may invest less in safety than a social planner would. High trust as well as a larger advantage of the leading team over its competitors increases the resulting welfare. Especially with a large capabilities advantage it is possible that the leading team can implement the socially optimal level of safety and still win the race, then the nash equilibrium will be efficient. In the second model two teams can invest in AI capabilities or safety. This gives the leader two ways to deter the follower. Both investing more in safety or in capabilities helps deter the second team.

Contents

1 Introduction

Artificial intelligence (AI) is shaping up to be a very transformative technology. But it might also come with serious risks. Existing papers have largely focused on how a benevolent social planner would choose to develop AI, trading off additional economic growth with existential risk (Jones, 2024; 2025). However, in reality AI development is likely to be competitive, with different nations and labs racing to develop powerful AI systems first.

I extend existing models of AI adoption, by adding strategic elements. I compare the decisions a benevolent planner makes with equilibrium choices in a race setting where being the first to develop AGI is more valuable than coming second. I place particular emphasis on implications for existential risk and economic growth.

This helps us get a clearer picture of the economic forces shaping the development of AI, and the form of impact powerful AI might have. The goal is not to create a maximally realistic model of an AGI race, but to highlight the most important dynamics and economic forces shaping outcomes.

I show that races to develop artificial general intelligence (AGI) can lead to overinvestment in AI capabilities and underinvestment in safety precautions. Although the first-best outcome can be achieved, specifically if the leading actors (nations or AI labs) trust each other sufficiently, or the leading team has a sufficiently large starting advantage, and therefore does not fear being overtaken. Using available data on compute and other important factors, I estimate parameters for the model and draw practical conclusions.

Related Literature. There is a significant non-economics literature discussing existential risk from AI (Ord, 2020; Grace and al., 2024). In economics there are many papers on AI and economic growth, Aghion, Jones and Jones (2017) is a good start for understanding the economic implications of AI, as is (Trammell and Korinek, 2023), which categorizes many different scenarios the future could take. There is an active debate over the impact AI will have on economic growth in the near future, with e.g. (Acemoglu, 2024) predicting only a 0.71% increase in total factor productivity over the next 10 years from AI, while many other researchers predict a much more dramatic impact (Epoch AI, 2024; Kokotajlo and al., 2024), as do many of the AI labs themselves.

There is a growth economics literature specifically looking at existential risk from AI (Jones, 2024; 2025; Growiec and Prettner, 2025). Naudé and Dimitri (2018) considers how race dynamics can shape outcomes of AI development. Trammell and Aschenbrenner (2025) looks at the interplay of growth and existential risk more generally.

I mainly build on three papers. Jones (2024) and Jones (2025) both consider the trade-off between consumption and existential risk from AI through the lens of a social planner. Though both have a different emphasis. Jones (2024) focuses on how risk aversion and the statistical value of a life, influence decisions trading off between growth

and existential risk. Through the framework: if AI is making you richer but has some risk, how rich would one have to become until you stop using it? Jones (2025) develops a model to answer the question how much a social planner should spend to reduce x-risk. The paper's main conclusion is that if x-risk is around 1% spending around 2% of GDP on reducing risks is optimal. Jones draws an analogy to COVID, where the risk of death was around 1% and we spent around 2% of GDP on various measures to fight the pandemic.

Much of the macroeconomic environment in my paper is taken from these two papers. Similarly to them I don't try to be maximally realistic, but rather to highlight specific economic forces at play. The third paper is Armstrong, Bostrom and Shulman (2013), which considers strategic elements of an AI race in a simple game-theoretic model; I largely adopt the race mechanisms from that paper.

2 Simple Model

I develop a model based on Jones (2024). This is a model where AI is not just another technology, but one as transformative as electricity or agriculture. Society faces a fundamental trade-off: adopting AI more aggressively increases economic growth but also adds existential risk. To this environment I add strategic elements from Armstrong, Bostrom and Shulman (2013): two teams (nations or AI labs) race to develop AGI first, and the winner inherits the majority of future economic output.[1]

2.1 The Planner's Problem

A benevolent social planner chooses a level of AI adoption, $a > 0$, to maximize social welfare. The planner's objective function is

$$\max_{a} \quad u(c(a)) \cdot S(a) \tag{1}$$

where $u(c)$ is the utility from consumption and $S(a)$ is the probability of survival. The choice variable a represents anything that speeds up AI progress at the cost of increased existential risk: faster adoption, deregulation, or direct investment in capabilities. In (Jones, 2024), a is interpreted as the number of years AI is used.

Consumption, c, depends on AI adoption, a, through the growth function:

$$c(a) \equiv e^{ga},$$

[1] I changed the names of some parameters. Differences to the Jones paper are: what I call a Jones calls T, my σ he calls δ. Differences to the Armstrong paper: my a they call s and the sign is flipped, their e I call b. A high value of s represents a lot of safety, while a high a represents a lack of safety. I also call the initial capabilities d while they use c.

[2] Theoretically, one has to multiply by a constant term c_0, but I will always assume $c_0 = 1$ and leave it out.

where g represents the additional economic growth from AI.[2] The probability of survival, $S(a)$, is a decreasing function of AI adoption:

$$S(a) \equiv e^{-\sigma a}$$

σ represents the dangerousness of AI; I call $S(a)$ the survival function. Finally consumers have Constant Relative Risk Aversion (CRRA) utility:

$$u(c) \equiv \begin{cases} \overline{u} + \log(c), & \text{if } \gamma = 1 \\ \overline{u} + \frac{c^{1-\gamma}}{1-\gamma}, & \text{if } \gamma \neq 1 \end{cases}$$

The additive term $\overline{u}$ represents the utility from today's consumption level. I will mostly explain the $\gamma = 1$ version of the model since the algebra is a lot more compact and results transfer to the more risk averse case, but I mention the $\gamma \neq 1$ case in the footnotes.

Social Optimum. The first order condition for the social planner's AI use is

$$v(c) \leq \frac{g}{\delta}, \qquad \text{where} \qquad v(c) \equiv \frac{u(c)}{u'(c)c}$$

$v(c)$ is the ratio of the value of a statistical life year to annual consumption. For example in the US it is currently around 6 (=240.000/60.000), we are willing to give up 6 years of consumption to save a statistical year of life. This value grows with income, the richer one is the less one is willing to risk one's life for more consumption.

The planner uses AI until this ratio is equal to $\frac{g}{\delta}$, the growth rate divided by the hazard rate. For example if $g = 0.1$ and $\delta = 0.01$, the planner would use AI until the value of a statistical year of life is 10 times annual consumption, until both ratios are identical people are not that satisfied enough with their life yet to shut down AI. The exact value for optimal AI use, which I call a^* is less intuitive but it will be useful for later sections, it is given by[3]

$$a^* = \frac{1}{\sigma} - \frac{\overline{u}}{g}. \tag{2}$$

2.2 Rules of the Race

For the strategic aspects I follow Armstrong, Bostrom and Shulman (2013). Two teams, $i \in \{1, 2\}$, race to develop AGI. Each team starts with initial AI capabilities d_i; WLOG team 1 has an advantage ($d_1 > d_2$). Since only the difference matters, I define $\Delta = d_1 - d_2$ and normalize $d_2 = 0$, so that $d_1 = \Delta$. Both teams simultaneously choose a level of AI use $a_i > 0$, and the team with the larger sum $d_i + a_i$ wins the race.

The winner gets utility according to Equation 1, while the loser receives only a fraction $1 - b$ of the winner's consumption. The parameter $b \in [0, 1)$ captures the

[3]For $\gamma \neq 1$, the social optimum is $a^* = \frac{1}{g(1-\gamma)} \ln\left(\frac{\sigma\overline{u}}{g - \frac{\sigma}{1-\gamma}}\right)$.

enmity between the teams: more enmity means the loser is treated worse. So the respective utilities are

$$U_{\text{Win}} = u(c(a_{\text{Win}})) \cdot S(a_{\text{Win}}) \qquad \text{and} \qquad U_{\text{Lose}} = u(c(a_{\text{Win}})(1-b)) \cdot S(a_{\text{Win}}).$$

Note that the loser's choice a_{Lose} doesn't affect utilities, given a_{Win} the losing team's choice is irrelevant. This is a game of complete information; both teams know all d_i. Ties are broken in favour of team 1, consequently team 2 would need to choose $a_2 = a_1 + \Delta + \varepsilon$ to win.

2.3 Existence and uniqueness of equilibrium

Next I derive the pure strategy Nash equilibrium of this game, I start with the best response function of team 1.

Proposition 1: Team 1's best response function is

$$B_1(a_2) = \max(a^*, a_2 - \Delta).$$

The proof can be found here in the appendix. This proposition says that team 1 will always win the race, and will either play the social optimum, if this deters team 2 from racing, or will use the minimal amount of AI needed to deter team 2. With this result we can assume that $a_1 \geq a^*$ holds from now on.

For each choice of a_1, team 2 will either want to outrace team 1 or accept defeat. Recall that choosing $a_2 = a_1 + \Delta$ leads to a tie, which is broken in favour of team 1, so to win team 2 must choose $a_2 = a_1 + \Delta + \varepsilon$, with $\varepsilon > 0$. Now it is clear when team 2 will race; if an ε exists, so that $U_2(a_1 + \Delta + \varepsilon) > U_2^{\text{Lose}}$ holds. To show that team 1 will always deter team 2 I show that for large enough a_1 no such ε exists.

Proposition 2: There always exists a large enough a_1 such that team 2 prefers losing to racing.

See here for the proof, from this proposition we get the following best response function for team 2

$$B_2(a_1) = \begin{cases} a_1 + \Delta + \varepsilon & \text{if } a_1 > a_{\text{indiff}} \\ a_2 \in [0, a_1 + \Delta] & \text{if } a_1 \leq a_{\text{indiff}} \end{cases}$$

where a_{indiff} is the unique solution to $U_2^{\text{Win}}(a_1) = U_2^{\text{Lose}}(a_1)$, i.e. the a_1 that makes team 2 indifferent between racing and losing. When team 1 plays a_{indiff} and team 2 plays $a_2 = a_1 + \Delta$ (a tie, broken in team 1's favour), both are playing best responses so this is a Nash equilibrium.

Team 1 won't decrease a_1 (they'd lose) or increase it (lower utility). Team 2 won't increase a_2: this would win them the race but the extra risk isn't worth it, they are indifferent between lowering and not lowering a_2. The intuition for this result as that Team 1 would prefer to choose the social optimum a^*. But this might not deter team

2 from racing, if it does the social optimum is played. Otherwise, team 1 increases a_1 just enough to make team 2 indifferent between racing and accepting defeat.

Again we can use $v(c)$ to get and intuitive understanding of the choosen level of AI use. Solving the indifference equation we get

$$U_2^{\text{race}} = U_2^{\text{lose}}$$
$$S(a_1 + \Delta)u(c) = S(a_1)u(c \cdot (1-b)).$$

Since, for log utility $v(a)u(a)$ we get

$$S(a_1 + \Delta)v(a) = S(a_1)(v(a) + \ln(1-b)).$$

now solving for $v(c)$ we get

$$S(a_1 + \Delta)v(a) = S(a_1)(v(a) + \ln(1-b))$$
$$\frac{S(a_1+\Delta)}{S(a_1)}v(a) = v(a) + \ln(1-b)$$
$$v(a) - \frac{S(a_1+\Delta)}{S(a_1)}v(a) = -\ln(1-b)$$
$$v(a)\left(1 - \frac{S(a_1+\Delta)}{S(a_1)}\right) = -\ln(1-b)$$
$$v(a) = -\frac{\ln(1-b)}{1 - \frac{S(a_1+\Delta)}{S(a_1)}}$$

substituting the survival function we get

$$v(c) = -\frac{\ln(1-b)}{1-e^{-\sigma\Delta}}$$

Recall, the social planner uses AI until $v(c) = \frac{g}{\sigma}$ holds. The social planner makes sure the (TODO) is equal to the benefits divided by the risks. In the equilibrium the (TODO) is equal to the extra utility gained from racing divided by the risk incured through racing. This is similar in spirit to the social planner's condition. One could call

- $\frac{g}{\sigma}$ the *social return to risk*, and
- $-\frac{\ln(1-b)}{1-e^{-\sigma\Delta}}$ the *private return to risk*

We can approximate this condition to get a better intuition:

- $-\ln(1-b) \approx b$ for small values of b, and
- $1 - e^{-\sigma\Delta} \approx \sigma\Delta$ for small $\sigma\Delta$.

So we get $v(c) \approx \frac{b}{\sigma\Delta}$. For example with $b = 0.3$, $\Delta = 2$ and $\sigma = 0.01$, the right hand side is equal to 0.3 / 0.02 = 15. If you are willing to give up 15 years of consumption to live a year longer, this implies you are willing to risk a 2% chance to your life to get a 30% increase in consumption.

We can also solve this approximation for when the social optimum is reached, this is when $\frac{b}{g} = \Delta$. There's probably some intuitive reason for this.

It is also useful to have the exact expression for the choosen value of AI use in the equilibrium, it is given by[4]

$$a_{\text{NE}} = -\frac{\ln(1-b)}{g(1-e^{-\sigma\Delta})} - \frac{\overline{u}}{g}. \tag{3}$$

The equilibrium is clearly unique. Existence is guaranteed since the players simply play a^* if $a^* > a_{\text{NE}}$ holds and zero if a^* is also negative.

We could also solve for survival probability and expected utility as functions of the parameters, but this yields little additional insight.

2.4 Comparative Statics and Efficiency

I'll start with a discussion of efficiency.

Proposition 3: The equilibrium outcome is efficient if the socially optimal amount of AI use is large enough to deter team 2 from racing, that is, when $a^* \geq a_{\text{NE}}$, or equivalently when

$$\frac{g}{\sigma} \geq -\frac{\ln(1-b)}{1-e^{-\sigma\Delta}}. \tag{4}$$

holds.

Most comparative statics for when efficiency is reached are immediately clear. Larger Δ or g lead to the equilibrium being efficient; larger b makes it less likely; $\overline{u}$ has no effect. The effect of σ is more complicated I will discuss it further below.

Comparative statics for equilibrium AI use.

Proposition 4: With $b \in (0,1)$ and $\overline{u}, g, \sigma, \Delta > 0$ we have the following comparative statics for the equilibrium AI use

$$\frac{\partial a_{\text{NE}}}{\partial b} = \frac{1}{gA(1-b)} > 0 \qquad \frac{\partial a_{\text{NE}}}{\partial \overline{u}} = -\frac{1}{g} < 0 \qquad \frac{\partial a_{\text{NE}}}{\partial g} = -\frac{1}{g} a_{\text{NE}} < 0$$

$$\frac{\partial a_{\text{NE}}}{\partial \Delta} = \sigma \frac{\ln(1-b)}{g} \frac{e^{-\sigma\Delta}}{A^2} < 0 \qquad \frac{\partial a_{\text{NE}}}{\partial \sigma} = \Delta \frac{\ln(1-b)}{g} \frac{e^{-\sigma\Delta}}{A^2} < 0$$

Since survival probability is monotonic in a we get its comparative statics by simply reversing these. Most of these signs are what one would expect; more enmity means there is more reason to race, a larger capabilities advantage makes racing harder and so decreases the amount of AI the leader has to use to win the race. The picture is similar for expected utility.

Proposition 5: With $b \in (0,1)$ and $\overline{u}, g, \sigma, \Delta > 0$ the equilibrium utility satisfies

[4] In the case $\gamma \neq 1$ the equilibrium is

$$a_{\text{NE}} = \frac{1}{g(1-\gamma)} \ln\left(\frac{\overline{u}(1-e^{-\sigma\Delta})(1-\gamma)}{e^{-\sigma\Delta} - (1-b)^{1-\gamma}} \right)$$

it is also unique and exists.

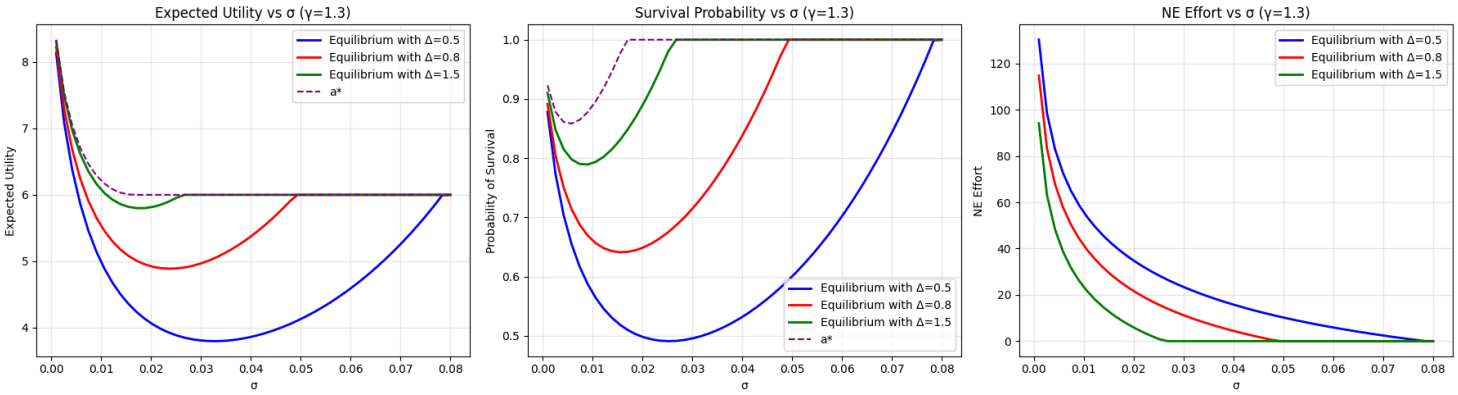

Figure 1: Comparative statics for σ for different values of Δ.

$$\frac{\partial U_{\text{NE}}}{\partial b} < 0 \qquad \frac{\partial U_{\text{NE}}}{\partial \overline{u}} > 0 \qquad \frac{\partial U_{\text{NE}}}{\partial g} > 0$$

$$\frac{\partial U_{\text{NE}}}{\partial \Delta} > 0 \qquad \frac{\partial U_{\text{NE}}}{\partial \sigma} \gtrless 0$$

The comparative statics for b, Δ, g and $\overline{u}$ have the expected signs; only σ is ambiguous. Figure 1 plots outcomes against σ (I discuss the other parameter values in the next section). The center panel shows that σ has a U-shaped effect on survival probability, even for the social planner. For high σ, it doesn't make sense to take any risks. For low σ, lots of AI is used but risks remain small. At intermediate values, survival probability is lowest. This U-shape doesn't carry over to utility for the social planner (higher σ can never improve outcomes), but in equilibrium the U-shape persists. Counterintuitively, increasing the fundamental riskiness of AI can sometimes improve equilibrium outcomes.

2.5 Numerical Results

For further results I will assume concrete values for the parameters. I choose most parameters as in Jones (2024)'s central estimate. I set $\sigma = 0.01$ and $g = 0.1$: each unit of AI use increases growth by 10% and existential risk by 1%. $\overline{u}$ is chosen so that the value of a statistical year of life is equal to 6 years of consumption, which is approximately the current value for the US[5]. The parameters b and Δ are harder to estimate. What fraction of consumption will the winner give the loser? How large is the leading team's advantage? I largely explore results by varying these two parameters.

But history can help suggest a rough estimate for b: the losers of World War 2 were not subjugated for centuries but received relatively light punishment. By 1980, Japan's

[5]Jones (2024) has a good explanation of what this means.
[6]9669$ and 12574$ respectively. Source: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

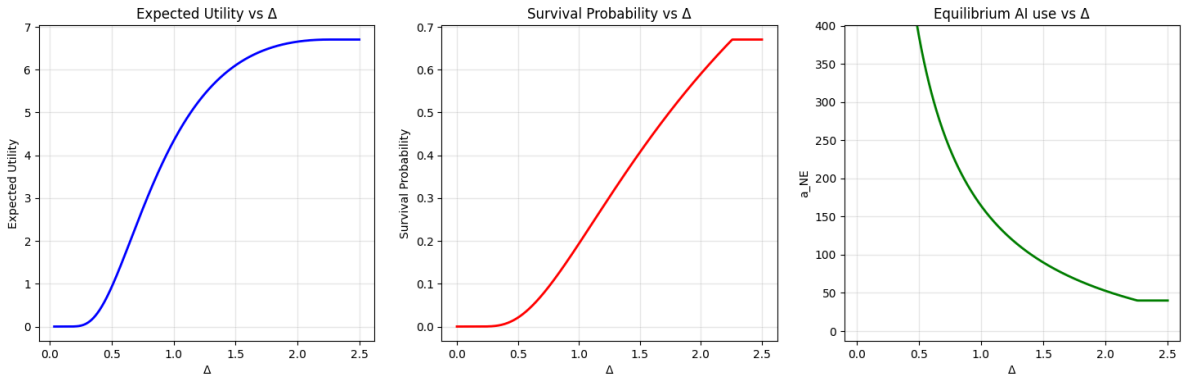

Figure 2: Comparative statics for Outcomes for different values of Δ. Other parameters are $\sigma = 0.01, b = 0.2, g = 0.1$.

GDP per capita was around 77% of that of the US.[6] This suggests b is not extremely large, perhaps not as high as 0.9. For the baseline I use around $b = 0.2 - 0.3$.

For Δ it makes sense to look at what is needed to achieve efficiency. Solving Equation 4 for Δ gives us $\Delta \geq -\frac{1}{\sigma}\ln\left(1 + \ln(1-b)\frac{\sigma}{g}\right)$. With my parameters, this requires $\Delta \geq 3.6$.[7] Comparing this to our previous approximation, $\Delta = \frac{b}{g}$ there we would have gotten $\Delta = 3$, so an ok estimate. Since Δ lacks a clear real-world interpretation in this version of the model, I vary it to explore results rather than calibrating it directly.

When the capability gap is small, equilibrium outcomes become extreme. Recall again that the social planner uses AI until $v(c) = \frac{g}{\sigma}$, with our parameters this means $v(c) = 10$, and that, in the equilibrium, AI is used until $v(c) = -\frac{\ln(1-b)}{1-e^{-\sigma\Delta}}$ holds. With $\Delta = 1.8$ (half what's needed for efficiency), this becomes: $-\frac{\ln(1-0.3)}{1-e^{-0.01\cdot 1.8}} \approx 20$. The Nash equilibrium AI use is $a_{\text{NE}} = 139$, versus $a^* = 40$ for the social optimum. Expected utility is 4.9 (compared to 6 with no AI, 6.7 with optimal AI use). Survival probability is around 0.25%, compared to 0.67% in the social optimum. The social planner would let consumption grow by a factor of 55, in the equilibrium, consumption grows by a factor of 1 million.

Figure 2 plots expected utility, survival probability, and AI use against Δ. The middle panel one can observe that the social optimum is reached for large values of Δ, hence the curve flattens. Expected utility follows an S-curve: for very small Δ, increasing the gap barely helps; then outcomes improve rapidly; finally they flatten as the social optimum is reached.

2.6 Dynamics of Deterrence

The way deterrence works in this model deserves closer examination. The intuitive way one would imagine a race is as follows: The marginal risk of additional AI use

[7] As $3.6 \approx -100 \cdot \ln(1 + 0.1 \cdot \ln(0.7))$.

increases in AI use. If the leading team isn't using much AI, it makes sense to out-race them: this adds some risk y but lets you win and gain extra utility x. But if the leader has high AI use, the extra risk of out-racing them is now higher (e.g., $10y$), while the extra benefit of winning is still x, so it doesn't make sense to race any more. A race like this is represented by the following table:

AI use Leader	Utility race	Utility lose	Extra Utility from Racing	Extra Risk from Racing
Low	9 −3	6 −1	3	2
High	12 −10	9 −5	3	5

Table 1: Hypothetical model where racing becomes more risky.

The additional utility from racing is always 3. The risk the follower adds by racing increases in the leader's AI use (from 2 to 5). Here, the total payoff of a state equals utility (in black) minus risk (in red).

But this isn't how my actual model works. Due to the memorylessness of the exponential function, the marginal danger of AI use doesn't vary with the level of AI use. Instead, deterrence works differently: when the leader uses lots of AI, the benefits of winning shrink, while the extra risk of racing stays constant.

AI use Leader	Utility race	Utility lose	Extra Utility from Racing	Extra Risk from Racing
Low	9 −3	6 −1	3	2
High	10 −7	9 −5	1	2

Table 2: Current model where racing becomes less valuable.

Racing always increases risk by 2, but the benefits of winning decrease in the leader's AI use (from 3 to 1).

This reflects the core mechanism of my model: the gain from winning over losing shrinks as AI use increases. Why? More AI use makes everyone wealthier, and wealthier agents care less about additional consumption relative to survival. It makes sense to race and increase x-risk by 3% if this doubles your consumption from $1,000 to $2,000 a year. But not when your consumption is already $100,000.

In this way, the dynamics in this race are surprisingly similar to those discussed in Trammell and Aschenbrenner (2025) and Jones (2025). As we become wealthier, we're more willing to trade away consumption for security.

While this is an interesting effect, it's unclear whether it's really what would be relevant in a race for AGI. In practice, I'd expect increasing marginal risk, not decreasing marginal benefit, to be the main driver of deterrence. Using $S(a) = e^{-\sigma a^2}$

as the survival function would implement this.[8]

2.7 Varying Levels of initial Consumption

The previous section highlights that consumption levels shape safety outcomes. To explore this further, I extend the model to allow different initial consumption, levels across teams. This also gives Δ a more concrete interpretation. We can add in consumption differences by varying c_i so that we now get:

$$c_i(a) \equiv c_i e^{ga}$$

where I still normalize c_1 to one, so we get $c_2 < 1$. The social optimum for someone with initial consumption equal to c_i is

$$a^* = \frac{1}{\sigma} - \frac{\overline{u} + \ln(c_i)}{g}.$$

The larger initial consumption is, the less AI is used. The change to the equilibrium is identical, we get

$$a = -\frac{\ln(1-b)}{g(1-e^{-\sigma\Delta})} - \frac{\overline{u} + \ln(c_i)}{g}$$

but the laggard will race more the poorer he is.

Previously I said Δ doesn't have a good interpretation, but now it does. It is the capabilities advantage that would raise the leading team's consumption from c_2 to 1 (I think this makes sense, but not 100% sure). That is, it solves $1 = c_2 e^{g\Delta}$, so we get

$$\Delta = -\frac{\ln(c_2)}{g}.$$

Writing our new equilibrium purely in terms of Δ we get

$$a_{\text{NE}} = -\frac{\ln(1-b)}{g(1-e^{-\sigma\Delta})} - \frac{\overline{u}}{g} + \Delta,$$

a curious result, the new equilibrium is exactly the previous equilibrium plus Δ.

To get the comparative statics for this new Δ, that includes also the corresponding shifts in c_2 we only need to add 1 to the previous result from Proposition 4, so we get

$$\frac{\partial a_{\text{NE}}}{\partial \Delta} = 1 + \sigma \frac{\ln(1-b)}{g} \frac{e^{-\sigma\Delta}}{A^2}.$$

Previously this expression was always negative, a larger capabilities difference would always make racing less intense. Now there is one new force at play, a higher Δ implies that the laggard is poorer, which makes them more desperate for consumption and more

[8] In this model the social optimum and equilibrium are

$$a^* = \frac{\sqrt{2g^2 + \overline{u}^2\sigma}}{2g\sqrt{\sigma}} - \frac{\overline{u}}{2g}, \qquad \text{and} \qquad a_{\text{NE}} = -\frac{\ln(1-b)}{2\Delta\sigma} - \frac{\Delta}{2}$$

respectively. Plotting the comparative statics, they seem fairly similar to the original model.

willing to race. The first effect is the dominant one, but we can find extreme parameter ranges where the second effect dominates and a larger Δ makes the equilibrium less safe.

First to simplify notation I define $Q \equiv -\frac{\ln(1-b)}{g(1-D)}$. We need three conditions for a larger Δ to make the equilibrium less safe. The derivative needs to be positive:

$$1 + \sigma Q \frac{D}{1-D} > 0,$$

the equilibrium needs to be positive (otherwise 0 is played):

$$Q + \Delta - \frac{\overline{u}}{g} > 0,$$

the equilibrium also needs to be played, so it needs to be larger than the socially optimal choice

$$Q + \Delta - \frac{1}{\sigma} > 0,$$

we can satisfy these three conditions simultaneously. In this case we typically have a very small c_2. The second place is a country with much less consumption per capita than the leader, and if they get even poorer, they become more desperate and race even more intensely. The increased risk doesn't deter them enough because they don't value their life sufficiently highly.

An interesting fourth condition

$$a_{\text{NE}} > a^*_{\text{lose}}$$

cannot also hold. a^*_{lose} is what the laggard wants to choose to maximise his utility, even when knowing he will lose. We can write this condition as

$$QD - \frac{1}{\sigma} > 0.$$

Previously we typically had

$$a^* < a^*_{\text{lose}} < a_{\text{NE}}.$$

The social optimum is smallest; then comes what the laggard would prefer even conditional on loosing (he wants more AI use because he only gets $1-b$ of consumption, so he is somewhat less risk averse); largest is the equilibrium value. With heterogeneous consumption levels, this ordering can change. The poorer team can be so poor that they want they will race just to force the leader to use more AI to raise their own consumption.

The key takeaway: increasing Δ still generally increases safety, but the effect is weaker than before. A larger capability gap now implies the laggard is poorer and more desperate to race, partially offsetting the deterrence effect.

2.8 Discussion

Most comparative statics are what one would expect: an increase in enmity makes racing more intense, while a larger gap in capabilities reduces intensity, as shown in

(Armstrong, Bostrom and Shulman, 2013). More interesting is that higher growth decreases racing intensity. One might expect that a more valuable post-AGI future makes winning more important, but since a higher growth rate increases both the winners and the losers consumption it creates a wealth effect that makes everyone more risk averse. An analogy: A person might risk more in a contest to increase his daily income from \$5 to \$10 than from \$50 to \$100. While the second increase is larger in absolute terms it is less valuable in terms of utility. This captures the same effect as Jones (2024) and (Trammell and Aschenbrenner, 2025): the richer we are (or expect to be), the more we value safety over additional consumption.

This effect may help explain why there is less violence today than in previous centuries: the richer we are, the less willing we are to risk everything in a war. For a potential AI race, the importance of this effect is hard to estimate. If you trust the other team not to enslave or destroy you, a more glorious shared future should make you less inclined to race. On the other hand, it's hard to imagine China being persuaded by the US arguing "don't race, there will be so much GDP growth that there's enough for everyone."

The effect of σ also seems relevant: worlds where a technology poses very high risk may actually be safer than intermediate-risk worlds. This has practical implications increasing public awareness of AI risks could be especially valuable, as a higher intrinsic risk only lowers equilibrium risk if agents are aware of it. But heterogeneous beliefs complicate matters: if risks are very high, there's an increased chance that some team radically underestimates them and does something foolish (Bostrom, Douglas and Sandberg, 2016), but this only strengthens the case for increasing public awareness.

For the second point it is relevant that the abilities of LLMs largely follow scaling laws (Kaplan *et al.*, 2020), smooth mappings from compute[9] to model capabilities, this makes it possible to use a compute centric model of AGI development. I use compute as a proxy for AI capabilities in general. This is helpful because the amount of compute owned by different actors can be estimated more easily than other measures of AI capabilities.

3 Richer Model

The simple model treats the arrival of AGI as exogenous. Here I develop a richer model where teams choose both safety spending (x) and capability investment (k), and the latter determines when AGI arrives. The macroeconomic environment is loosely inspired by Jones (2025).

[9]Compute is the ability to perform many computations, measured in FLOP/s (floating-point operations per second).

3.1 Environment

As before, I start by considering the problem a benevolent social planner faces. The planner maximizes the sum of pre-AGI utility and expected post-AGI utility:

$$\max_{x,k} \int_0^T u(c)e^{-\rho t}\,dt + S(x,T)\int_T^\infty u(c_t)e^{-\rho t}\,dt$$

in the pre-AGI era we integrate the utility from consumption, discounted the usual way. To this we add the value of the post-AGI future multiplied by the survival probability. Assuming no GDP growth until T and constant growth afterwards, this simplifies to:

$$\max_{x,k} u(c)\frac{1-e^{-\rho T}}{\rho} + S(x,T)e^{-\rho T}V \tag{5}$$

subject to the budget constraint:

$$c = 1 - x - k, \qquad \text{with} \qquad x, k \geq 0, \qquad x + k < 1.$$

Total output is normalized to 1 and can be allocated to consumption (c), safety spending (x), or capability investment (k).

The terms are:

- $u(c)$: (constant) flow utility from pre-AGI consumption (CRRA as before)
- $\frac{1-e^{-\rho T}}{\rho}$: adjustment for discounting and time until AGI
- $S(x,T)$: survival probability, depending on safety spending and time
- $e^{-\rho T}$: discount factor for the post-AGI future
- V: value of the post-AGI future (depends only on exogenous parameters), given by $V = \frac{1}{(1-\gamma)(\rho-(1-\gamma)G)} + \frac{\overline{u}}{\rho}$, where G is post-AGI growth

The choice variables have the following effects:

- Increasing x or k lowers pre-AGI consumption
- AGI arrival time (T) decreases in capability spending (k), this in turn decreases:
  - how much the post-AGI future is discounted (the only social upside of increasing k)
  - for how long pre-AGI utility accumulates
  - how much time there is to invest in safety
- Higher safety spending (x) raises survival probability

Survival Probability is given by

$$S(x,T) = 1 - (1-\phi)\delta_0 - \phi\delta_0 e^{-\alpha X^\lambda T}.$$

δ_0 is the baseline risk, $S(0,T) = 1 - \delta_0$ and ϕ is the fraction of risk that can be mitigated. Safety spending is multiplied by T: one can increase safety either by spending more on safety or by investing less in capabilities. If $\phi = \delta_0 = 1$ the function reduces to $1 - e^{-\alpha X^\lambda}$. The parameter $\lambda < 1$ captures diminishing returns to parallelizing safety

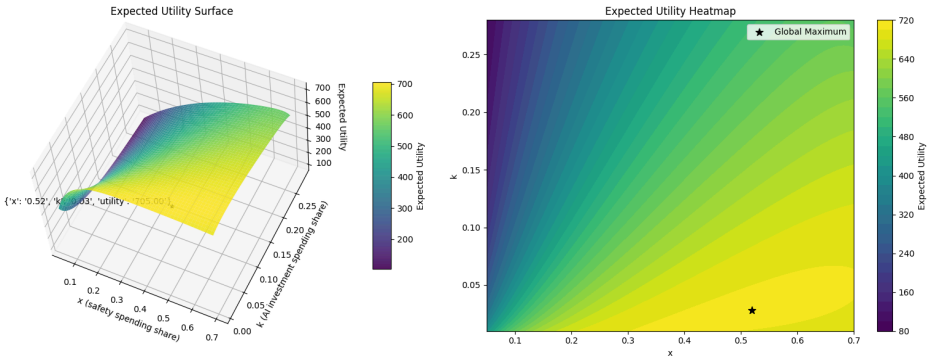

Figure 3: The social planner's optimal choice of safety and capability spending (x, k).

research. For this to work we need $X > 1$, so I define $X = \frac{x}{\overline{x}}$ with $\overline{x} \ll 1$.[10]

AGI Arrival Time. AGI arrives when a team accumulates enough compute to build AGI (K) that is when

$$K = \int_0^T \frac{k_i Y_i}{\mu_i p(t)} dt, \tag{6}$$

holds. Here

- Y is GDP (assumed constant)
- k_i is the fraction of GDP spent on compute,
- $p(t) = p_0 e^{-wt}$ is the price of compute (falling exponentially per Moore's law),
- μ_i is a country-specific cost factor (higher μ_i means compute is more expensive).

Solving for T gives us AGI arrival time as a function of k:

$$T(k) = \frac{\ln\left(1 + \frac{\mu\tau}{k}\right)}{w},$$

where $\tau = \frac{Kwp_0}{Y}$

The planner's problem This finishes the setup of the environment. Now to the social optimum. The social planner's problem involves a trade-off: increasing investment k speeds up the arrival of the valuable post-AGI future but reduces the time available for safety research to be effective and crowds out consumption and safety spending today. I solve this model numerically. Figure 3 plots the utility landscape and the optimal choice (x^*, k^*) of a social planner for one set of parameters.

3.2 Rules for the Race

I model the race as a Stackelberg-type game with a leader and a follower.

- The teams are distinguished by the different prices they pay for compute μ_i
- The leader moves first, choosing her investment in capabilities (k_1) and safety (x_1).

[10] x still represents the fraction of output spent on safety, X repsenents something like the number of total researchers working on safety.

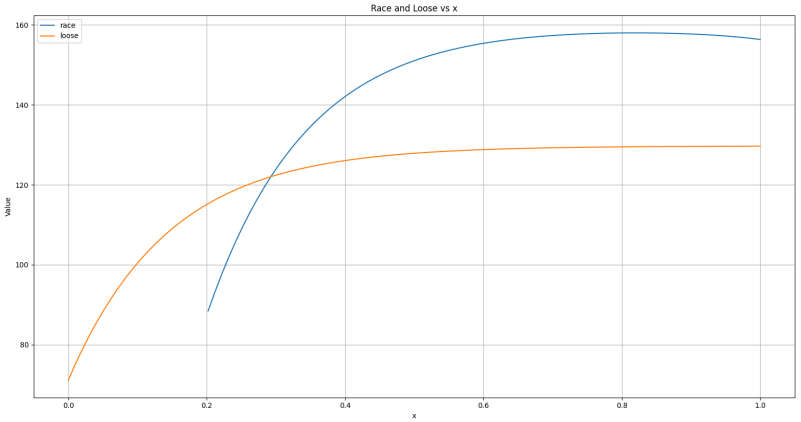

Figure 4: Racing vs. losing utility in a 1-variable illustration

- The follower observes the leader's choice and then chooses his own investments (k_2, x_2).
- Each team's pre-AGI utility depends only on their own choices.
- The winner is the team that first accumulates the required compute K to achieve AGI.
- The winner receives the same utility as the social planner, while the loser receives a fraction $(1-b)$ of the winner's post-AGI utility.

Specifically, if team i loses, their utility is:

$$U_i^{\text{lose}} = u(c_i)\frac{1-e^{-\rho T}}{\rho} + S(x_j, T)e^{-\rho T}V(1-b)$$

where $T = T_j$. I drop the subscript on T as $T_i > T_j$, so team i's arrival time never materialises. Since the winner builds AGI, their decisions on safety and capabilities spending impact both teams. Ties are broken in favour of the leader.

3.3 Solving for the Equilibrium

The formal derivation of the Nash equilibrium is in Appendix A. I give an informal outline of the solution here. First, to build intuition it helps to look at a simpler 1-variable version of this model. Figure 4 shows the follower's utility when racing and when accepting defeat as functions of the leader's choice variable x. The two curves cross at a point I call x_{cross}, where the follower is indifferent between racing and losing. If, ignoring the race, the leader's utility is concave in x, she will choose the socially optimal value x^* if it is deterring. Otherwise she chooses the closest deterring value, which turns out to be x_{cross}. So she chooses $\min\{x^*, x_{\text{cross}}\}$. Now to the equilibrium in the full model. I start with the follower's problem.

	Parameter	Value
Share of x-risk that can be eliminated	ϕ	0.9
X-risk, with no mitigation	δ_0	0.9
Effectiveness safety spending	ξ	0.5
Risk aversion parameter	γ	1.5
Discount factor	ρ	0.01
Post AGI Growth	G	0.2
Safety spending parallelization penalty	λ	0.8
Enmity	b	0.3
Capability gap	μ_2	1.6
Speed of AI progress	τ	0.17

Table 3: Baseline parameter ranges for richer model

The Follower's Problem Given the leader's decision, the follower chooses between racing and accepting defeat. When racing, he sets k_2 to just defeat the leader, i.e. $k_2 = \frac{\mu_2}{\mu_1}k_1 + \varepsilon$ for some small $\varepsilon > 0$,[11] and then chooses the safety spending x_2 that maximises his utility given k_2. He races if his racing utility exceeds his losing utility.

The Leader's Problem The leader anticipates the follower's response. Since she always wants to win, she will choose a deterring tuple, so the follower prefers losing to racing. From this set she chooses the tuple that maximises her utility. This typically results in the leader making the follower exactly indifferent between racing and losing. In the simple model there was only one such point. Now there is a range of k_1 values, each with a corresponding x_1 that induces indifference; typically one of them is chosen. This can be seen in Figure 5.

3.4 Equilibrium Analysis

For further analysis I assume concrete numerical values for parameters. For most plots and comparative statics where I vary one parameter, I use the baseline set of parameters summarised in Table 3. I have tried to find reasonable estimates for most parameters. The most noteworthy parameters are:

- μ_2, the leader's compute advantage, which I base on the US's compute advantage over China.
- τ, the difficulty of developing AGI, which I calibrate to match current trends in compute prices and estimates of when AGI will be developed, e.g. around 7 years with 2% of GDP focused on AI, 5 years with 4%.

The equilibrium is most easily understood by plotting utilities as functions of the leader's capability spending k_1; Figure 5 does this. The follower prefers to race for small

[11] As before this is not well defined, but we can ignore this because it is off equilibrium.

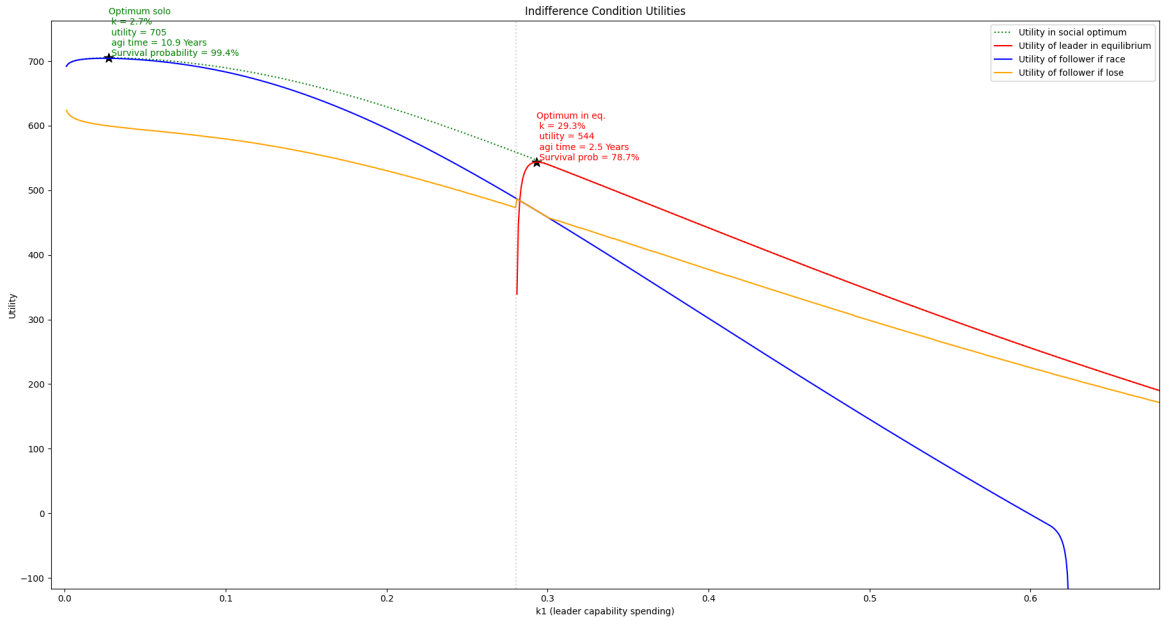

Figure 5: Equilibrium and social optimum

k_1 and to lose for large k_1. In between there is a small region where he can be made indifferent. The leader's equilibrium utility achieves its maximum at a point in that region.

An interesting dynamic of this equilibrium is that the leader must appease the follower by investing more in safety when k_1 is small. The leader has two ways to deter the follower: increasing k_1 to make racing more costly, or increasing x_1 to make losing more acceptable. For small k_1, the leader must invest more and more in safety to achieve deterrence. In equilibrium the leader chooses a k_1 that is close to the smallest k_1 that can still deter the follower, but does not go lower because the cost to consumption from the required safety spending would be too high. One implication is that both capabilities and safety spending can be higher in the equilibrium than in the social optimum. The remaining results are unsurprising: in equilibrium, risk is higher, utility is lower, and time until AGI is shorter.

Comparative Statics I plot the comparative statics for most parameters in Figure 6. I vary one parameter at a time while keeping others at baseline values from Table 3. I will not list all results as one can just look at the plot, but generally the simple model is confirmed and there are no surprising effects. The larger the capability gap, the lower enmity, and the easier safety research is, the better the outcomes, etc.

4 Discussion

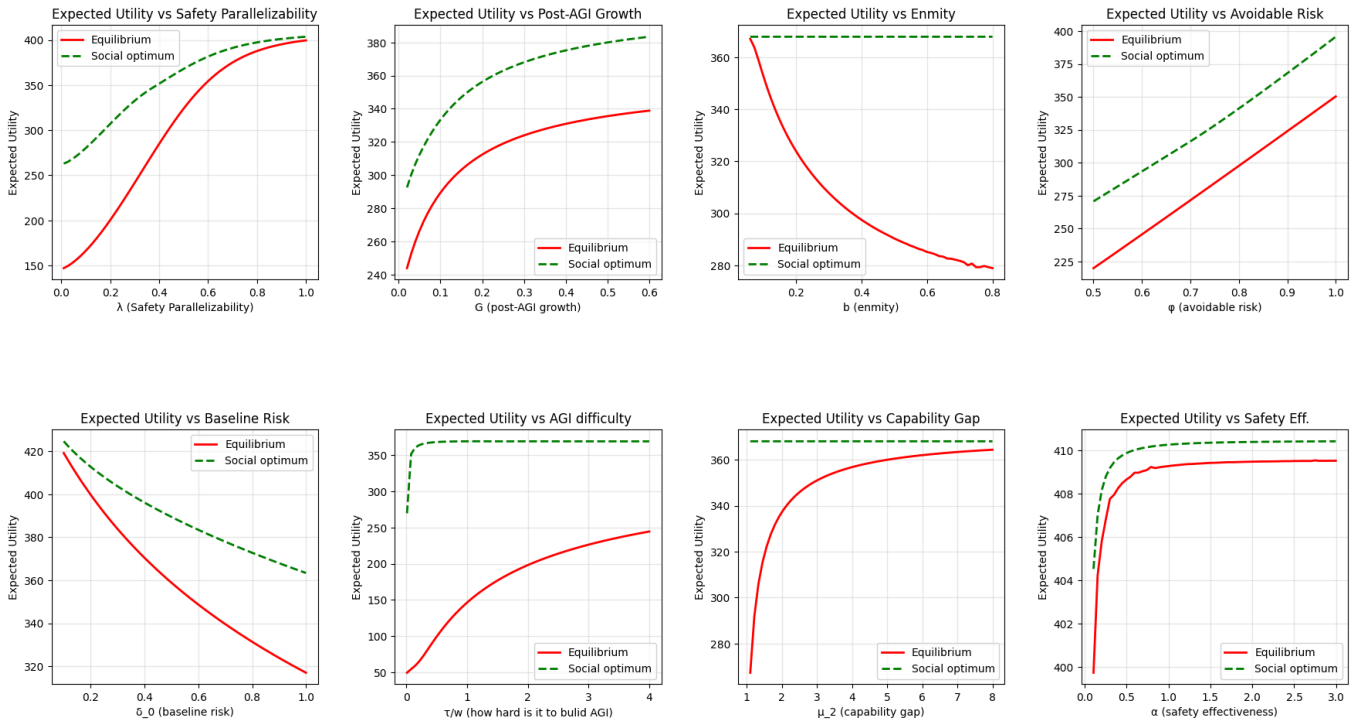

Figure 6: Comparative Statics for different parameters

I discuss the implications of the simple model in its discussion section (Section 2.8) in more detail. The most interesting result is how, if racing makes us wealthier, this makes us more risk averse and in turn less willing to race.

For the richer model, the most interesting result is the dual way the leader can deter the follower from racing: either by investing in capabilities to make racing more costly, or in safety to make losing less painful. One can ask whether such a dynamic could play out in practice. Imagine the US and China are both close to developing AGI. Both could quickly advance capabilities without investing more in safety, but they are trying to be as safe as possible while still winning. The US is considering different options:

1. Slowly advancing capabilities:
    - This leads to China outracing them unless they invest a very large amount in safety, which would satisfy China.
    - This would be very costly for the US.
2. Invest a moderate amount in both
    - China could still outrace them but it would be so costly and risky and wouldn't be worth it.
    - The US prefers this to the first option.
3. Invest a lot in capabilities and a bit in safety
    - Same as before for China.

- The US prefers the second to this option.

The real world is of course more complex, but the dynamic of China demanding extra safety measures in exchange for not racing seems plausible.

The results shared by both models are fairly standard:

1. Racing can lead to bad outcomes
   - increased risk
   - lowered expected utility
2. Outcomes are improved by
   - lowering enmity (b)
   - increasing the leading team's advantage (μ_2/Δ), as shown in (Armstrong, Bostrom and Shulman, 2013).
3. There are relevant policy/practical implications:
   - The US should not sell chips to China, as this would lower μ_2.
   - Efforts should be made to lower enmity between the competing nations or AI labs.

The models presented have several limitations:

- The models are static, there is no uncertainty, beliefs are homogeneous.
- More research needed etc.

References

Acemoglu, D. (2024) *The Simple Macroeconomics of AI*. Working Paper No. w32487. National Bureau of Economic Research.

Aghion, P., Jones, B.F. and Jones, C.I. (2017) *Artificial Intelligence and Economic Growth*. National Bureau of Economic Research.

Armstrong, S., Bostrom, N. and Shulman, C. (2013) *Racing to the precipice: a model of artificial intelligence development*. Technical Report #2013-1. Future of Humanity Institute, Oxford University, pp. 1–8.

Bostrom, N., Douglas, T. and Sandberg, A. (2016) "The Unilateralist's Curse and the Case for a Principle of Conformity." Social Epistemology.

Epoch AI (2024) *GATE: An Integrated Assessment Model for AI Automation*. Epoch AI.

Grace, K. and al. (2024) "Thousands of AI Authors on the future of AI."

Growiec, J. and Prettner, K. (2025) "The Economics of p(doom): Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI."

Jones, C.I. (2024) "The AI Dilemma: Growth versus Existential Risk," *American Economic Review: Insights*, 6, pp. 575–590.

Jones, C.I. (2025) "How Much Should We Spend to Reduce A.I.'s Existential Risk?."

Kaplan, J. *et al.* (2020) "Scaling Laws for Neural Language Models."

Kokotajlo, D. and al. (2024) *AI2027*.

Naudé, W. and Dimitri, N. (2018) "The Race for an Artificial General Intelligence: Implications for Public Policy."

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

Trammell, P. and Aschenbrenner, L. (2025) "Existential Risk and Growth."

Trammell, P. and Korinek, A. (2023) "Economic Growth under Transformative AI."

A Derivation Equilibrium

Now I will derive the Nash equilibrium of the richer model. I will start by characterising the follower's best response given the leader's choice, and then find the leader's optimal choice taking the follower's best response into account.

Notation:

- I define:
  - $c(x_2) = 1 - x_2 - k_1\mu_2$, this is the follower's consumption when racing.
  - $h = h(T) = \frac{1-e^{-\rho T}}{\rho}$ the adjustment term for pre-AGI utility. And
  - $D = D(T) = e^{-\rho T}$, the amount the post-AGI future is discounted
- Whenever k_1 is fixed I leave it implicit, writing $S(x_1)$, $c(x_2)$, h, T and D, even though they all depend on k_1.
- Because only the ratio of μ_1 and μ_2 will end up mattering I will let $\mu_1 = 1$, so that $\frac{\mu_2}{\mu_1} = \mu_2$.

The Follower's decision

Proposition 6:

1. When not racing the follower will choose $k_2 = x_2 = 0$.
2. When racing the follower chooses $k_2 = k_1\mu_2 + \varepsilon$, with $\varepsilon \to 0$,
3. And his safety spending is the unique x_2 that solves
$$\alpha\lambda x_2^{\lambda-1}\overline{x}^{-\lambda}T\phi\delta_0 e^{-\alpha\left(\frac{x_2}{\overline{x}}\right)^{\lambda}T}DV = c(x_2)^{-\gamma}h \tag{7}$$
4. If the resulting utility is higher than when losing they will race, otherwise they will not.

Proof of Proposition 6.

- *Claim 1.* This is obvious as the loser's choices do not impact the post AGI future. Thus, all resources should go to consumption.
- *Claim 4.* This is also obvious.
- *Claim 2.* This problem has the same structure as Proposition 2. The argument requires concavity of the follower's utility in k_2 (with x_2 chosen optimally). While showing this analytically is intractable, concavity must be verified numerically for each parameter configuration used. Given concavity, the follower will want to just barely outpace the leader, so that $k_2 = k_1\mu_2 + \varepsilon$ with $\varepsilon \to 0$ (We run into the same non-existence problem here as before, but it isn't a problem, see[14]).
- *Claim 3.* Given k_1 the follower maximises:
$$U_2^{\text{race}}(x_2) = u(c(x_2))h + S(x_2, T)D(T)V$$

All three functions S, u and D are concave and smooth in x_2, consequently the follower's best response exists, is unique, and depends continuously on k_1. It will be at one of the boundaries or satisfy Equation 7, which I derive by differentiating and setting equal to zero, giving us

$$\frac{dU_{2\text{race}}}{dx_2} = -c(x_2)^{-\gamma}h + S'(x_2, T)DV = 0$$

then substituting $\frac{\partial S}{\partial x} = \alpha\lambda x^{\lambda-1}\overline{x}^{-\lambda}T\phi\delta_0 e^{-\alpha\left(\frac{x}{\overline{x}}\right)^{\lambda}T}$ this gives us Equation 7 after rearranging.[12]

□

The Leader's decision

While it is clear the leader will always play to win the race, because of her compute advantage, it is slightly more complicated to determine the exact choice of x_1 and k_1. We have to determine the tuple that maximises her utility, from the set of tuples that deter the follower.

It turns out that there exists a unique x_1 for each k_1 that makes the follower indifferent between racing and losing, this is the smallest deterring x_1 for that k_1, I will call it $x_1^{\text{ind}}(k_1)$.

Proposition 7: If inducing indifference is possible for a given k_1, then there exists a unique x_1 that does this. Defining S_{target} to be the survival probability that makes the follower indifferent between racing and losing, the leader chooses

$$x_1^{\text{ind}}(k_1) = S^{-1}\left(S_{\text{target}}\right) \tag{8}$$

where

$$S_{\text{target}} = \frac{1}{1-b}\left((u(c) - u(1))\frac{D^{-1} - 1}{V\rho} + S(x_2, k_1)\right).$$

Proof of Proposition 7. Again, I am keeping as k_1 fixed. The follower is indifferent if $U_2^{\text{race}} = U_2^{\text{lose}}$ holds, as x_1 only appears once in this equations it is easy to solve for it.

$$u(1 - x_2 - k_1\mu_2)h + S(x_2)DV = u(1)h + S(x_1)DV(1-b)$$

$$\frac{u(c)h}{DV} + S(x_2) = \frac{u(1)h}{DV} + S(x_1)(1-b)$$

$$\frac{1}{1-b}\left((u(c) - u(1))\frac{h}{DV} + S(x_2)\right) = S(x_1) \coloneqq S_{\text{target}}$$

[12]This doesn't quite follow because the choice set isn't compact. We can't actually choose $x + k = 1$ as the utility function is undefined there. But because the utility goes to $-\infty$ there, this value won't ever be chosen. Thus the maximum lies in the interior of the choice set or at $x = k = 0$ which is defined.

S_{target} is the unique level of safety that makes the the follower indifferent. As S is monotonic and continuous we can invert it on its image and also get a unique x_1. Written out this is

$$x_1^{\text{ind}}(k_1) = \overline{x}\left(-\frac{1}{\alpha T}\ln\left(\frac{1-S_{\text{target}}}{\delta_0\phi} - \frac{1-\phi}{\phi}\right)\right)^{\frac{1}{\lambda}}.$$

□

So, given k_1, if inducing indifference is possible we have a unique x_1 that does this. Unlike in single-variable model where the x producing indifference was unique, we now have multiple values of k_1, each with a corresponding x_1 producing indifference. But the leader won't always play $x_1^{\text{ind}}(k_1)$, the next proposition shows what x_1 the leader chooses given k_1.

Proposition 8: For a fixed k_1 the leader chooses

$$\max\left\{x_1^{\text{ind}}(k_1),\, x_1^*(k_1)\right\} \tag{9}$$

where $x_1^*(k_1)$ is the social welfare maximizing choice of x_1 given k_1.

Proof of Proposition 8. Again this is all keeping k_1 fixed.

1. The Leaders utility is concave in x_1, this follows from Proposition 6 (since the followers problem when racing is practically identical to the leaders problem). It follows that the ideal x_1 must be the value that is closest to $x_1^*(k_1)$, while still being deterring.
2. If a x_1 is deterring all larger x_1 are also deterring. To see this recall we have deterrence if $U_2^{\text{race}} \leq U_2^{\text{lose}}$ holds. And we have $\frac{\partial U_2^{\text{race}}}{\partial x_1} = 0$ and $\frac{\partial U_2^{\text{lose}}}{\partial x_1} > 0$.
3. So the leader chooses $x_1^*(k_1)$ if it is deterring, if it isn't deterring it must be too small, so the leader chooses the smallest value that is deterring, this is $x_1^{\text{ind}}(k_1)$.

This produces the stated result. □

Now the leaders problem is reduced to a one-dimensional problem, she simply selects the k_1 that maximizes her utility while being deterring, with x_1 being chosen according to Equation 9.

B Proofs

Proof of Proposition 1. For any choice of a_2 by team 2, team 1 will want to outrace team 2 by choosing at least $a_1 = a_2 - \Delta$, as this both decreases risk from $S(a_2)$ to $S(a_2 - \Delta)$ and increases team 1s consumption by a factor of $1/(1-b)$.[13]

As U_1 is concave in a_1, team 1 will choose a^* if it wins them the race, or the minimal value that accomplishes this. □

[13]To verify this note that, given a_2, team 1s utility from outracing team 2 is always higher than from accepting defeat, $U_1^{\text{Win}} > U_1^{\text{Lose}}$ always holds. Written out this condition is$S(a_2 - \Delta)(\overline{u} + ga_2) > S(a_2)(\overline{u} + ga_2 + \ln(1-b))$.

Proof of Proposition 2. First I define

$$R(a_1) = U_2^{\text{Win}}(a_1) - U_2^{\text{Lose}}(a_1)$$

where team 2 is assumed to choose $a_2 = a_1 + \Delta + \varepsilon$ if racing. Team 2 wants to race if $R > 0$ and accepts defeat otherwise. We need to show that for sufficiently large a, R will be negative for all ε. We already know that a_1 will always be at least as large as a^*, and hence increasing a or equivalently ε always lowers the utility of the winner. Thus it suffices to show that R is negative for ε approaching 0.[14] Solving for R, while treating $\varepsilon = 0$ we get

$$\begin{aligned} R = u(e^{ga})e^{-\sigma(a+\Delta)} - u(e^{ga}(1-b))e^{-\sigma a} &< 0 \\ u(e^{ga})e^{-\sigma\Delta} - u(e^{ga}(1-b)) &< 0 \\ (\overline{u} + ga)e^{-\sigma\Delta} - [\overline{u} + ga + \ln(1-b)] &< 0 \\ \left(e^{-\sigma\Delta} - 1\right)(\overline{u} + ga) &< \ln(1-b) \\ \overline{u} + ga &> \frac{\ln(1-b)}{e^{-\sigma\Delta} - 1} \end{aligned}$$

since the right hand side is constant and the left hand side increases linearly, the condition will be met for large enough a, in addition the solution to $R = 0$ is unique.

Just to verify, with the ε included we get:

$$\overline{u} + ga > \frac{-\ln(1-b) + g\varepsilon e^{-\sigma(\Delta+\varepsilon)}}{1 - e^{-\sigma(\Delta+\varepsilon)}}.$$

The denominator is bounded away from 0 and can't make the right hand side explode. $\ln(1-b)$ is constant, and $g\varepsilon e^{-\sigma(\Delta+\varepsilon)})$ also has an upper bound, as the exponential decay dominates the linear growth in ε. Hence the right hand side is bounded from above and some a will satisfy the inequality. □

Proof of Proposition 4. First note that $A := 1 - e^{-\sigma\Delta} > 0$.

1. *For b:* Only the logarithmic term depends on b. Hence

$$\begin{aligned} \frac{\partial a_{\text{NE}}}{\partial b} &= -\frac{1}{gA} \cdot \frac{\partial}{\partial b} \ln(1-b) = -\frac{1}{gA} \cdot \left(-\frac{1}{1-b}\right) \\ &= \frac{1}{gA(1-b)} > 0. \end{aligned}$$

2. *For $\overline{u}$:* Obvious.

[14] To verify this note that, conditional on winning team 2 will choose ε to maximize

$$U_{\text{Win}} = (g(a+\varepsilon) + \overline{u})e^{-\sigma(a+\Delta+\varepsilon)}.$$

This is done by $\varepsilon^* = \max\left\{0, \frac{1}{\sigma} - \frac{\overline{u}}{g} - a\right\}$. This just tells us that if a_1 is smaller than a^* player 2 will use ε to fill this gap. But since we have $a_1 > a^*$ this will never happen and smaller ε are always better.

This leads to the problem that team 2's optimal winning choice of ε is not defined, since there is no smallest real number larger than 0. This isn't a significant problem though as team 2 only wins off-equilibrium, hence it never really happens.

3. *For g:* Since $\frac{\partial}{\partial g}\frac{1}{g} = -\frac{1}{g}\cdot\frac{1}{g}$ we have
$$\frac{\partial a_{\text{NE}}}{\partial g} = -\frac{1}{g}\cdot a_{\text{NE}} = \frac{\ln(1-b)}{g^2 A} + \frac{\overline{u}}{g^2} < 0$$
where the sign follows since $a_{\text{NE}} > 0$ and this has the opposite sign.

4. *For Δ:* Only A depends on Δ, first computing its partial derivative we have
$$\frac{\partial A}{\partial \Delta} = \sigma e^{-\sigma\Delta}$$
Using the chain rule,
$$\frac{\partial a_{\text{NE}}}{\partial \Delta} = -\frac{\ln(1-b)}{g}\cdot\frac{\partial}{\partial\Delta}\left(\frac{1}{A}\right) = \frac{\ln(1-b)}{g}\frac{\sigma e^{-\sigma\Delta}}{A^2} < 0$$
which follows since $\ln(1-b) < 0$ and all other terms are positive.

5. *For σ:* Since σ and Δ appear symmetrically in a swapping them in one partial derivative gives us the other one, hence
$$\frac{\partial a_{\text{NE}}}{\partial \sigma} = \frac{\ln(1-b)}{g}\cdot\frac{\Delta e^{-\sigma\Delta}}{A^2} < 0.$$

□

Proof of Proposition 5. Recall that
$$U_{\text{NE}} = e^{-\sigma a_{\text{NE}}}(\overline{u} + g a_{\text{NE}}).$$
We first compute the derivative with respect to equilibrium AI use:
$$\begin{aligned}\frac{\partial U_{\text{NE}}}{\partial a_{\text{NE}}} &= e^{-\sigma a_{\text{NE}}} g - \sigma e^{-\sigma a_{\text{NE}}}(\overline{u} + g a_{\text{NE}}) \\ &= e^{-\sigma a_{\text{NE}}}(g - \sigma(\overline{u} + g a_{\text{NE}})).\end{aligned}$$
Since $a_{\text{NE}} \geq a^*$ we have
$$\begin{aligned}g - \sigma(\overline{u} + g a_{\text{NE}}) &< g - \sigma(\overline{u} + g a^*) = g - \sigma\left(\overline{u} + g\left(\frac{1}{\sigma} - \frac{\overline{u}}{g}\right)\right) \\ &< g - \sigma\left(\overline{u} + \frac{g}{\sigma} - \overline{u}\right) = g - \sigma\overline{u} + g - \sigma\overline{u} \\ &< 0\end{aligned}$$
and hence $\frac{\partial U_{\text{NE}}}{\partial a_{\text{NE}}} < 0$.

1. *For b:* Since U_{NE} depends on b only through a_{NE} we have
$$\frac{\partial U_{\text{NE}}}{\partial b} = \frac{\partial U_{\text{NE}}}{\partial a_{\text{NE}}}\cdot\frac{\partial a_{\text{NE}}}{\partial b} < 0,$$
recall that the second term positive from Proposition 4.

2. *For $\overline{u}$:* Taking the derivative of U_{NE} we have

$$\begin{aligned}\frac{\partial U_{\text{NE}}}{\partial \overline{u}} &= \frac{\partial}{\partial \overline{u}}[e^{-\sigma a_{\text{NE}}}(\overline{u} + g a_{\text{NE}})] \\ &= e^{-\sigma a_{\text{NE}}}\left(1 + g\frac{\partial a_{\text{NE}}}{\partial \overline{u}}\right) - \sigma e^{-\sigma a_{\text{NE}}}(\overline{u} + g a_{\text{NE}})\frac{\partial a_{\text{NE}}}{\partial \overline{u}} \\ &= e^{-\sigma a_{\text{NE}}}(1-1) + \sigma e^{-\sigma a_{\text{NE}}}\frac{\overline{u} + g a_{\text{NE}}}{g} \\ &= \sigma \cdot e^{-\sigma a_{\text{NE}}}\left(\frac{\overline{u}}{g} + a_{\text{NE}}\right) > 0\end{aligned}$$

where I substited $\frac{\partial a_{\text{NE}}}{\partial \overline{u}} = -\frac{1}{g}$ in the third line, and the sign follows since all terms are positive.

3. *For g:* Taking the total derivative we have
$$\frac{dU_{\text{NE}}}{dg} = \frac{\partial U_{\text{NE}}}{\partial g} + \frac{\partial U_{\text{NE}}}{\partial a} \cdot \frac{\partial a_{\text{NE}}}{\partial g}.$$
Recall that $\frac{\partial a_{\text{NE}}}{\partial g} = -\frac{a_{\text{NE}}}{g}$ substituting this, as well as solving/substituting for the other two terms we get
$$\frac{dU_{\text{NE}}}{dg} = a_{\text{NE}} e^{-\sigma a_{\text{NE}}} + e^{-\sigma a_{\text{NE}}}(g - \sigma(\overline{u} + g a_{\text{NE}})) \cdot \left(-\frac{a_{\text{NE}}}{g}\right)$$
since I'm only interested in the sign I divide by $e^{-\sigma a_{\text{NE}}} a_{\text{NE}}$, which is positive to get
$$1 - \frac{g - \sigma(\overline{u} + g a_{\text{NE}})}{g} = \frac{\sigma(\overline{u} + g a_{\text{NE}})}{g} > 0.$$

4. *For Δ:* Since Utility depends on Δ only through a_{NE} we have
$$\frac{\partial U_{\text{NE}}}{\partial \Delta} = \frac{\partial U_{\text{NE}}}{\partial a_{\text{NE}}} \cdot \frac{\partial a_{\text{NE}}}{\partial \Delta} > 0$$
since both terms are negative.

5. *For σ:* The sign of the derivative is ambiguous.

□

Email address: s69tcurr@uni-bonn.de