Existential Risk and Growth

Philip Trammell and Leopold Aschenbrenner

ASSA Annual Meeting
Policy Implications of Transformative AI
3 Jan 2025

Motivation

An existential catastrophe $\equiv$ a catastrophe that kills everyone or, similarly, permanently sets global welfare to $\sim$0.

Motivation

An existential catastrophe $\equiv$ a catastrophe that kills everyone or, similarly, permanently sets global welfare to $\sim$0.

Prominent claims: (Parfit, 1984; Sagan, 1997; Ord, 2020; etc.)

- Technological progress has historically increased the risk of existential catastrophe, esp. via nuclear weapons, biotechnology, climate change, and AI.
- We are thus living through a once-in-history "time of perils", during which civilization will secure stability or perish.
- Accelerating tech progress today could increase the risk further.
  Concern for the long term should motivate slower technological development.

Motivation

An existential catastrophe $\equiv$ a catastrophe that kills everyone or, similarly, permanently sets global welfare to $\sim$0.

Prominent claims: (Parfit, 1984; Sagan, 1997; Ord, 2020; etc.)

- Technological progress has historically increased the risk of existential catastrophe, esp. via nuclear weapons, biotechnology, climate change, and AI.
- We are thus living through a once-in-history "time of perils", during which civilization will secure stability or perish.
- Accelerating tech progress today could increase the risk further.
  Concern for the long term should motivate slower technological development.

This paper: We have good reasons to believe we are living through a "time of perils".
A claim defended surprisingly little, for its importance! (Thorstad 2022, 2023)

Motivation

An existential catastrophe $\equiv$ a catastrophe that kills everyone or, similarly, permanently sets global welfare to $\sim$0.

Prominent claims: (Parfit, 1984; Sagan, 1997; Ord, 2020; etc.)

- Technological progress has historically increased the risk of existential catastrophe, esp. via nuclear weapons, biotechnology, climate change, and AI.
- We are thus living through a once-in-history "time of perils", during which civilization will secure stability or perish.
- Accelerating tech progress today could increase the risk further.
  Concern for the long term should motivate slower technological development.

This paper: We have good reasons to believe we are living through a "time of perils".
A claim defended surprisingly little, for its importance! (Thorstad 2022, 2023)

But if so, accelerating tech progress would likely *decrease* or *not affect* cumulative existential risk if optimally regulated, even by a planner with little concern for long-term survival.

Risk and survival

Time-varying hazard rate δ_t represents the flow probability of existential catastrophe.

Risk and survival

Time-varying hazard rate δ_t represents the flow probability of existential catastrophe.

Probability of survival to date t:

$$S_t \equiv e^{-\int_0^t \delta_s \, ds}.$$

Risk and survival

Time-varying hazard rate δ_t represents the flow probability of existential catastrophe.

Probability of survival to date t:

$$S_t \equiv e^{-\int_0^t \delta_s ds}.$$

$\{\delta_t\}_{t=0}^{\infty} \equiv$ the hazard curve

$X \equiv \int_0^{\infty} \delta_t dt \equiv$ cumulative risk

$S_\infty = e^{-X} \equiv$ the probability of survival: decreases in X, and > 0 iff X is finite

Framework and results

Suppose δ_t just depends on the tech level: $\delta_t = \delta(A_t) > 0$.
(Knowing how to make nukes, or deter their use; viruses, or vaccines; etc.)
Path of A is exogenous, has a positive derivative, and $\lim_{t\to\infty} A_t = \infty$.

Framework and results

Suppose δ_t just depends on the tech level: $\delta_t = \delta(A_t) > 0$.
(Knowing how to make nukes, or deter their use; viruses, or vaccines; etc.)
Path of A is exogenous, has a positive derivative, and $\lim_{t\to\infty} A_t = \infty$.

How does acceleration affect cumulative risk?
Change of variables:

$$X = \int_0^\infty \delta(A_t)\,dt = \int_{A_0}^\infty \delta(A)\left(\frac{dA}{dt}\right)^{-1} dA = \int_{A_0}^\infty \delta(A)\dot{A}_A^{-1}\,dA,$$

where $\dot{A}_A$ denotes the value of $\dot{A}$ when the technology level equals the subscripted A.

Framework and results

Suppose δ_t just depends on the tech level: $\delta_t = \delta(A_t) > 0$.
(Knowing how to make nukes, or deter their use; viruses, or vaccines; etc.)
Path of A is exogenous, has a positive derivative, and $\lim_{t\to\infty} A_t = \infty$.

How does acceleration affect cumulative risk?
Change of variables:

$$X = \int_0^\infty \delta(A_t)\,dt = \int_{A_0}^\infty \delta(A)\left(\frac{dA}{dt}\right)^{-1} dA = \int_{A_0}^\infty \delta(A)\dot{A}_A^{-1}\,dA,$$

where $\dot{A}_A$ denotes the value of $\dot{A}$ when the technology level equals the subscripted A.

Temporary level effects ($\uparrow \dot{A}$ now, $\downarrow \dot{A}$ later): <u>ambiguous</u>.

Framework and results

Suppose δ_t just depends on the tech level: $\delta_t = \delta(A_t) > 0$.
(Knowing how to make nukes, or deter their use; viruses, or vaccines; etc.)
Path of A is exogenous, has a positive derivative, and $\lim_{t\to\infty} A_t = \infty$.

How does acceleration affect cumulative risk?
Change of variables:

$$X = \int_0^\infty \delta(A_t)\,dt = \int_{A_0}^\infty \delta(A)\left(\frac{dA}{dt}\right)^{-1} dA = \int_{A_0}^\infty \delta(A)\dot{A}_A^{-1}\,dA,$$

where $\dot{A}_A$ denotes the value of $\dot{A}$ when the technology level equals the subscripted A.

Temporary level effects ($\uparrow \dot{A}$ now, $\downarrow \dot{A}$ later): ambiguous.
Temporary accelerations ($\uparrow \dot{A}$ for a bounded range of A-levels): lower X ...unless $X = \infty$.

Framework and results

Suppose δ_t just depends on the tech level: $\delta_t = \delta(A_t) > 0$.
(Knowing how to make nukes, or deter their use; viruses, or vaccines; etc.)
Path of A is exogenous, has a positive derivative, and $\lim_{t\to\infty} A_t = \infty$.

How does acceleration affect cumulative risk?
Change of variables:

$$X = \int_0^\infty \delta(A_t)\,dt = \int_{A_0}^\infty \delta(A)\left(\frac{dA}{dt}\right)^{-1} dA = \int_{A_0}^\infty \delta(A)\dot{A}_A^{-1}\,dA,$$

where $\dot{A}_A$ denotes the value of $\dot{A}$ when the technology level equals the subscripted A.

Temporary level effects ($\uparrow \dot{A}$ now, $\downarrow \dot{A}$ later): <u>ambiguous</u>.
Temporary accelerations ($\uparrow \dot{A}$ for a bounded range of A-levels): <u>lower</u> X ...unless $X = \infty$.
Permanent accelerations ($\uparrow \dot{A}$ for all A-levels): <u>lower</u> X ...perhaps even if $X = \infty$.

Framework and results

Suppose δ_t just depends on the tech level: $\delta_t = \delta(A_t) > 0$.
(Knowing how to make nukes, or deter their use; viruses, or vaccines; etc.)
Path of A is exogenous, has a positive derivative, and $\lim_{t\to\infty} A_t = \infty$.

How does acceleration affect cumulative risk?
Change of variables:

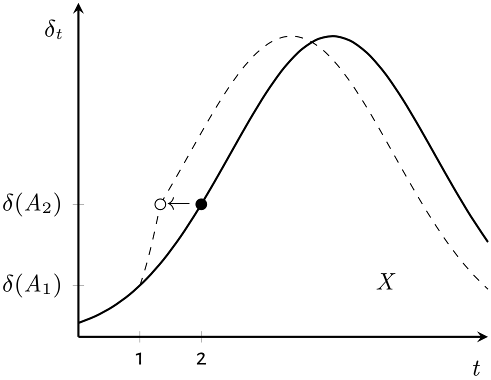
$$X = \int_0^\infty \delta(A_t)\,dt = \int_{A_0}^\infty \delta(A)\left(\frac{dA}{dt}\right)^{-1} dA = \int_{A_0}^\infty \delta(A)\dot{A}_A^{-1}\,dA,$$

where $\dot{A}_A$ denotes the value of $\dot{A}$ when the technology level equals the subscripted A.

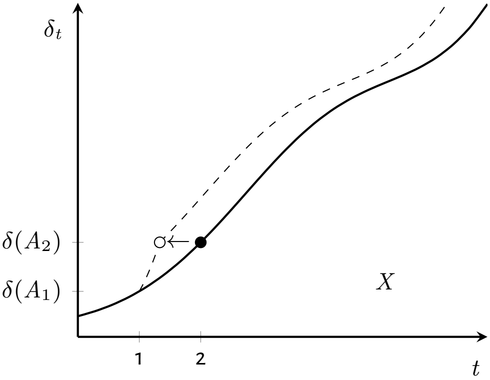
Temporary level effects ($\uparrow\ \dot{A}$ now, $\downarrow\ \dot{A}$ later): <u>ambiguous</u>.
Temporary accelerations ($\uparrow\ \dot{A}$ for a bounded range of A-levels): <u>lower</u> X ...unless $X = \infty$.
Permanent accelerations ($\uparrow\ \dot{A}$ for all A-levels): <u>lower</u> X ...perhaps even if $X = \infty$.

Permanent stagnation yields constant δ, so $S_\infty = 0$.

Illustration (temporary acceleration from A_1 to A_2)

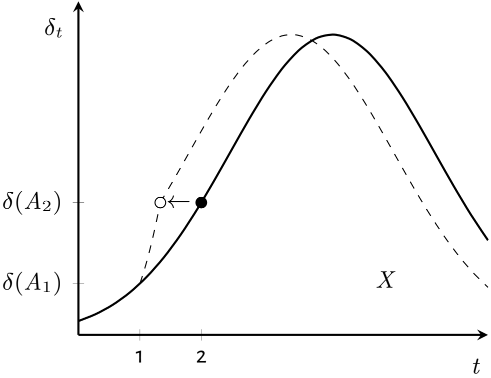


X finite $\implies$ acceleration lowers it

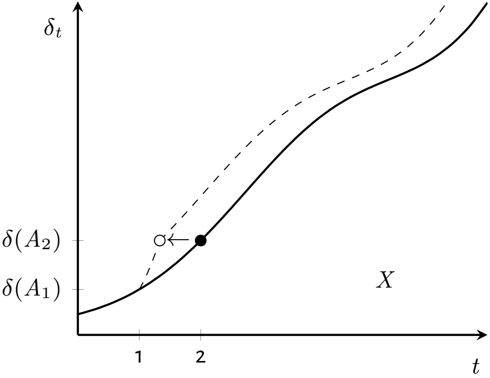


X infinite $\implies$ acceleration has no effect

Illustration (temporary acceleration from A_1 to A_2)



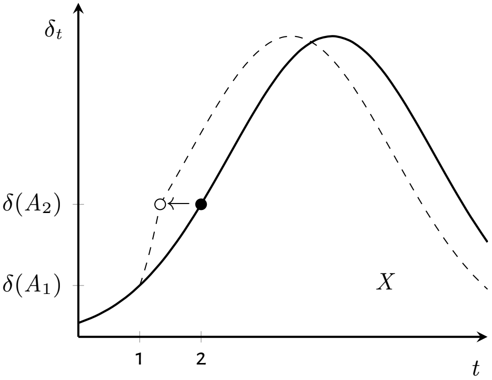
X finite $\implies$ acceleration lowers it



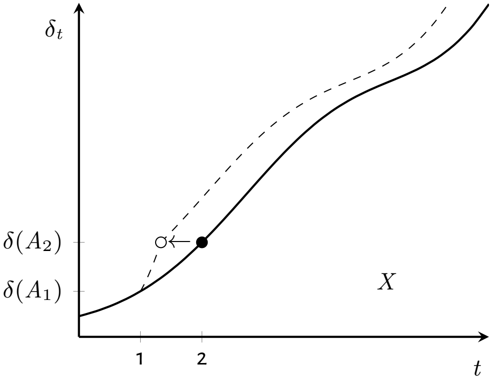
X infinite $\implies$ acceleration has no effect

Illusion of risky growth? $\uparrow \dot{A}$ may $\uparrow \delta$ in the short term. Still, cannot raise X (lower S_∞).

Illustration (temporary acceleration from A_1 to A_2)



X finite $\implies$ acceleration lowers it



X infinite $\implies$ acceleration has no effect

Illusion of risky growth? $\uparrow \dot{A}$ may $\uparrow \delta$ in the short term. Still, cannot raise X (lower S_∞).
"Longtermism" and x-risk? Concern for the *short term* can motivate slowing tech.

Exogenous policy

Let x_t denote a policy variable, and suppose $\delta_t = \delta(A_t, x_t)$.

Exogenous policy

Let x_t denote a policy variable, and suppose $\delta_t = \delta(A_t, x_t)$.

If the policy path $\{x_t\}$ is <u>exogenous</u>, the effect of acceleration on X is ambiguous.
Example where acceleration increases X:

$$\delta_t = A_t x_t, \qquad x_t = (1+t)^{-2}.$$

Consider accelerating the tech path from

$$A_t = (1+t)^k \quad \text{to} \quad (1+t)^{\tilde{k}}, \qquad k < 1 < \tilde{k}.$$

Exogenous policy

Let x_t denote a policy variable, and suppose $\delta_t = \delta(A_t, x_t)$.

If the policy path $\{x_t\}$ is exogenous, the effect of acceleration on X is ambiguous.
Example where acceleration increases X:

$$\delta_t = A_t x_t, \qquad x_t = (1+t)^{-2}.$$

Consider accelerating the tech path from

$$A_t = (1+t)^k \quad \text{to} \quad (1+t)^{\tilde{k}}, \qquad k < 1 < \tilde{k}.$$

X rises from finite ($k - 2 < -1$) to infinite ($\tilde{k} - 2 > -1$):

$$\int_0^\infty (1+t)^{k-2}\,dt \quad \text{to} \quad \int_0^\infty (1+t)^{\tilde{k}-2}\,dt.$$

Optimal policy

But “optimal” policy *strengthens* the tendency for acceleration to lower X, **in 2 ways**.

Optimal policy

But “optimal” policy *strengthens* the tendency for acceleration to lower X, **in 2 ways**.

Model: $\delta_t = \delta(A_t, x_t)$. A_t grows exogenously as before; $\delta(\cdot)$ decreases in $x_t \in [0, 1]$.
Consumption is $C_t = A_t x_t$:

- Technology A is indexed by potential consumption.
- Policy x is indexed by the fraction of consumption sacrificed to lower risk.

Optimal policy

But "optimal" policy *strengthens* the tendency for acceleration to lower X, **in 2 ways**.

Model: $\delta_t = \delta(A_t, x_t)$. A_t grows exogenously as before; $\delta(\cdot)$ decreases in $x_t \in [0,1]$. Consumption is $C_t = A_t x_t$:

- Technology A is indexed by potential consumption.
- Policy x is indexed by the fraction of consumption sacrificed to lower risk.

The optimal x path maximizes, at each t, the expected continuation value:

$$v_t \equiv \int_t^\infty e^{-\rho(s-t)} \frac{S_s}{S_t} u(C_s)\, ds.$$

Optimal policy

But "optimal" policy *strengthens* the tendency for acceleration to lower X, **in 2 ways**.

Model: $\delta_t = \delta(A_t, x_t)$. A_t grows exogenously as before; $\delta(\cdot)$ decreases in $x_t \in [0, 1]$. Consumption is $C_t = A_t x_t$:

- Technology A is indexed by potential consumption.
- Policy x is indexed by the fraction of consumption sacrificed to lower risk.

The optimal x path maximizes, at each t, the expected continuation value:

$$v_t \equiv \int_t^\infty e^{-\rho(s-t)} \frac{S_s}{S_t} u(C_s)\, ds.$$

1 Even if $\delta(\cdot)$ increases in A fixing x, $\uparrow A \implies \uparrow v, \downarrow u' \implies \downarrow x$. ("Kuznets curve.")
This may make X finite, and thus lowerable.

Optimal policy

But "optimal" policy *strengthens* the tendency for acceleration to lower X, **in 2 ways**.

Model: $\delta_t = \delta(A_t, x_t)$. A_t grows exogenously as before; $\delta(\cdot)$ decreases in $x_t \in [0,1]$.
Consumption is $C_t = A_t x_t$:

- Technology A is indexed by potential consumption.
- Policy x is indexed by the fraction of consumption sacrificed to lower risk.

The optimal x path maximizes, at each t, the expected continuation value:

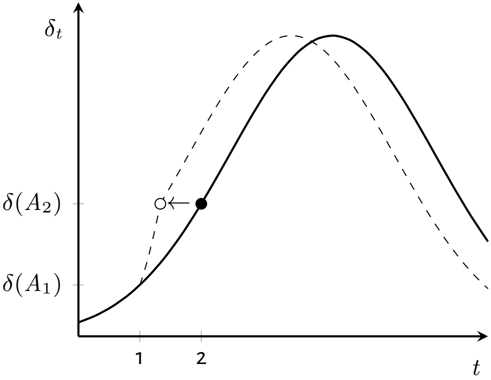
$$v_t \equiv \int_t^\infty e^{-\rho(s-t)} \frac{S_s}{S_t} u(C_s)\, ds.$$

1 Even if $\delta(\cdot)$ increases in A fixing x, $\uparrow A \implies \uparrow v, \downarrow u' \implies \downarrow x$. ("Kuznets curve.")
This may make X finite, and thus lowerable.

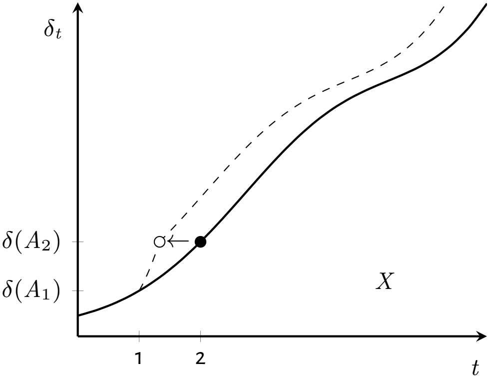
2 If x_t is just a function of A_t, back to state-risk-only: $\delta_t = \delta(A_t, x(A_t)) = f(A_t)$.
The cost (in "utils") of lowering x_t just depends on A_t. But the expected benefit $\frac{\partial \delta(A_t, x_t)}{\partial x_t} v_t$ increases in v_t, which increases in anticipated growth. So faster growth after $t \implies \downarrow x$ at A_t.

Optimal policy: illustration

State risk only



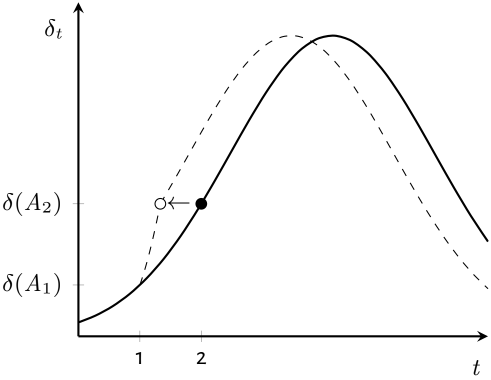
X finite $\implies$ acceleration lowers it.



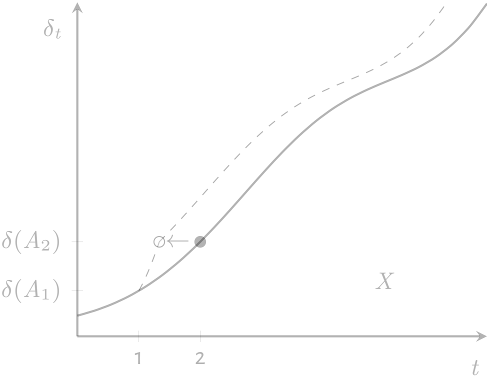
X infinite $\implies$ acceleration has no effect.

Optimal policy: illustration

State risk only + **1**



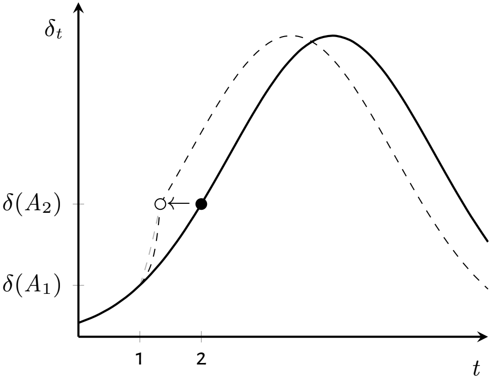
X finite $\implies$ acceleration lowers it.



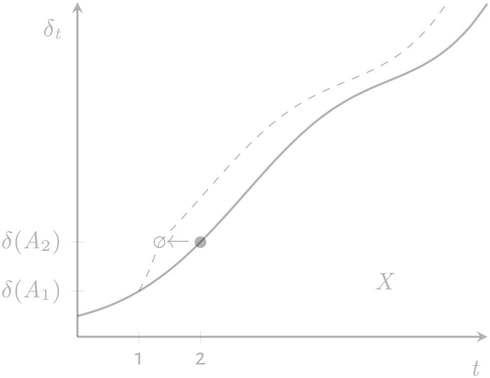
X infinite $\implies$ acceleration has no effect.

Optimal policy: illustration

State risk only + **1**, **2**



X finite $\implies$ acceleration lowers it.



X infinite $\implies$ acceleration has no effect.

Transition risk

Under hazard function $\delta(A_t, x_t)$, δ depends on the *state* of technology.
Escaping risky states more quickly lowers cumulative risk.

Transition risk

Under hazard function $\delta(A_t, x_t)$, δ depends on the *state* of technology.
Escaping risky states more quickly lowers cumulative risk.

But risk may instead be “transitional”: posed by *technological development*.
(Frontier virology lab; building an AI with unknown but immediate impact; etc.)

Transition risk

Under hazard function $\delta(A_t, x_t)$, δ depends on the *state* of technology.
Escaping risky states more quickly lowers cumulative risk.

But risk may instead be "transitional": posed by *technological development*.
(Frontier virology lab; building an AI with unknown but immediate impact; etc.)

Suppose δ_t increases in $\dot{A}_t$ rather than, or in addition to, A_t: $\quad \delta_t = \dot{A}_t^{\zeta} f(A_t, x_t), \ \ \zeta > 0.$

Transition risk

Under hazard function $\delta(A_t, x_t)$, δ depends on the *state* of technology.
Escaping risky states more quickly lowers cumulative risk.

But risk may instead be "transitional": posed by *technological development*.
(Frontier virology lab; building an AI with unknown but immediate impact; etc.)

Suppose δ_t increases in $\dot{A}_t$ rather than, or in addition to, A_t: $\delta_t = \dot{A}_t^{\zeta} f(A_t, x_t), \quad \zeta > 0.$
Developing *new* tech poses risk; $\delta = 0$ if $\dot{A} = 0$.
Different discoveries can pose different amounts of risk, depending on $f(\cdot)$.

Transition risk

Under hazard function $\delta(A_t, x_t)$, δ depends on the *state* of technology.
Escaping risky states more quickly lowers cumulative risk.

But risk may instead be “transitional”: posed by *technological development*.
(Frontier virology lab; building an AI with unknown but immediate impact; etc.)

Suppose δ_t increases in $\dot{A}_t$ rather than, or in addition to, A_t: $\quad \delta_t = \dot{A}_t^{\zeta} f(A_t, x_t), \quad \zeta > 0$.
Developing *new* tech poses risk; $\delta = 0$ if $\dot{A} = 0$.
Different discoveries can pose different amounts of risk, depending on $f(\cdot)$.

If $\zeta = 1$, as in Jones (2016, 2024), the growth rate from some $\underline{A}$ to $\overline{A}$ does not directly affect X.
Higher anticipated $\dot{A}$ may lower δ by increasing v and thus lowering x.

Transition risk

Under hazard function $\delta(A_t, x_t)$, δ depends on the *state* of technology.
Escaping risky states more quickly lowers cumulative risk.

But risk may instead be "transitional": posed by *technological development*.
(Frontier virology lab; building an AI with unknown but immediate impact; etc.)

Suppose δ_t increases in $\dot{A}_t$ rather than, or in addition to, A_t: $\quad \delta_t = \dot{A}_t^{\zeta} f(A_t, x_t), \quad \zeta > 0.$
Developing *new* tech poses risk; $\delta = 0$ if $\dot{A} = 0$.
Different discoveries can pose different amounts of risk, depending on $f(\cdot)$.

If $\zeta = 1$, as in Jones (2016, 2024), the growth rate from some $\underline{A}$ to $\overline{A}$ does not directly affect X.
Higher anticipated $\dot{A}$ may lower δ by increasing v and thus lowering x.

More generally, the effects of acceleration are ambiguous, depending on ζ and $f(\cdot)$.
Even if $\zeta > 1$, so that (all else equal) "experiments" are riskier concurrently than in sequence, acceleration may lower X due to policy interactions.

Policy implication for TAI

Suppose you believe:

1. AI will soon do/accelerate tech development across the board (so that speeding/slowing AI development amounts to speeding/slowing A)

Policy implication for TAI

Suppose you believe:

1. AI will soon do/accelerate tech development across the board (so that speeding/slowing AI development amounts to speeding/slowing A)
2. Hazard primarily
   - depends on the *state* of technology, or
   - increases not (much) faster than proportionally in the *rate* of tech development

Policy implication for TAI

Suppose you believe:

1. AI will soon do/accelerate tech development across the board (so that speeding/slowing AI development amounts to speeding/slowing A)
2. Hazard primarily
   - depends on the *state* of technology, or
   - increases not (much) faster than proportionally in the *rate* of tech development
3. Policymakers will "optimally" navigate the consumption-safety tradeoff

   Contra e.g. Shulman and Thornley (2024)

Policy implication for TAI

Suppose you believe:

1. AI will soon do/accelerate tech development across the board (so that speeding/slowing AI development amounts to speeding/slowing A)
2. Hazard primarily
   - depends on the *state* of technology, or
   - increases not (much) faster than proportionally in the *rate* of tech development
3. Policymakers will "optimally" navigate the consumption-safety tradeoff

   Contra e.g. Shulman and Thornley (2024)

Then—even if your *only* goal is to reduce cumulative risk, i.e. you're arbitrarily more risk-averse than those setting policy—you should prefer faster AI development.

Policy implication for TAI

Suppose you believe:

1. AI will soon do/accelerate tech development across the board (so that speeding/slowing AI development amounts to speeding/slowing A)
2. Hazard primarily
   - depends on the *state* of technology, or
   - increases not (much) faster than proportionally in the *rate* of tech development
3. Policymakers will "optimally" navigate the consumption-safety tradeoff

   Contra e.g. Shulman and Thornley (2024)

Then—even if your *only* goal is to reduce cumulative risk, i.e. you're arbitrarily more risk-averse than those setting policy—you should prefer faster AI development.

This observation has no bearing on:

- How stringently to regulate AI deployment, holding development fixed.
  Indeed, one benefit of high A is low x.
- Whether to attempt a targeted slowing of certain sectors of AI development.