NBER WORKING PAPER SERIES

THE AGI RACE AND EXISTENTIAL RISK

Ethan Bueno de Mesquita
Wioletta Dziuda
Mattias Polborn

Working Paper 35276
http://www.nber.org/papers/w35276

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2026

We have benefitted from comments from seminar participants at the University of Chicago, Stanford University, the University of Zurich, Yale University, and UIUC. Ethan Bueno de Mesquita has advised Meta on governance related issues. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The AGI Race and Existential Risk
Ethan Bueno de Mesquita, Wioletta Dziuda, and Mattias Polborn
NBER Working Paper No. 35276
May 2026
JEL No. D0, D20, D21, O3, O30

ABSTRACT

Concerns about the race to artificial general intelligence often assume that competition and resources increase risk by accelerating development. We study a model in which firms allocate scarce resources between speed and safety. Speed increases a firm's chance of reaching AGI first but leaves fewer resources for safety; safety lowers doom risk but slows arrival. Fragmentation increases total speed and conditional doom risk by shifting a fixed industry resource pool toward speed. The model also identifies a critical market size: below it, firms have positive expected payoff from achieving AGI, while above it, firms race even though achieving AGI has negative expected value. More per-firm resources always accelerate expected arrival, but their effect on conditional doom risk changes sign at this cutoff. Policy affects risk by changing equilibrium incentives: consolidation, resource regulation, commitment devices, and cautious public entry can improve welfare in some environments. The results show that AGI risk depends not only on technical considerations, but also on market structure, resource constraints, and institutions that shape the equilibrium allocation between speed and safety.

Ethan Bueno de Mesquita
The University of Chicago
bdm@uchicago.edu

Wioletta Dziuda
The University of Chicago
and NBER
wdziuda@uchicago.edu

Mattias Polborn
Vanderbilt University
Department of Economics
polborn@gmail.com

Artificial general intelligence (AGI) describes automated systems capable of performing a wide range of economically and strategically valuable cognitive tasks at a human level or beyond. Such systems could automate research, write and improve software, assist with strategic planning, operate digital infrastructure, and accelerate further AI innovation. The first firm to achieve AGI may therefore obtain enormous private advantages. But high capability also raises the stakes of inadequate control. If a powerful system is misaligned, deployed prematurely, or made available to malicious actors, arrival may produce a catastrophic event rather than a socially valuable innovation. Following common terminology in frontier-AI discussions, we refer to such an event as *doom*.[1]

Precisely because the first successful developer may obtain such large advantages, a central worry is that AGI development will become a race in which firms move too quickly and exercise too little caution. This concern is often linked to the possibility that, once achieved, AGI could rapidly improve itself and generate an explosion in capability (Good, 1966). OpenAI's Charter (OpenAI, 2018) states: "We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions." Anthropic's Joe Carlsmith describes the resource-allocation tradeoff underlying this worry: safety work may require "compute to do these intensive experiments on the AIs," but "that compute, we could use that for experiments for the next scaling step."[2] The concern is therefore not only that firms may care too little about safety. It is that safety and speed may compete for the same scarce inputs—compute, engineers, energy, managerial attention, and time—and that winner-take-most incentives may tilt those inputs toward speed.

These worries sit at the center of current AGI policy debates. Many policy proposals aim to slow frontier development, restrict access to compute, license large training runs, consolidate or coordinate private firms, or build public AI systems governed by objectives other than private profit.[3] This paper studies how market structure and resource constraints shape the speed-safety trade-off in the AGI race. The central point is that policies do not affect safety only by changing the technical difficulty of alignment or the amount of compute available for development. They also change the equilibrium allocation of scarce inputs between speed and caution. This makes the welfare effect of intervention depend on how it changes the strategic value of speed: by changing firm resources, the number of competitors, the allocation of resources across firms, or the credibility of restraint.

The idea that racing can induce developers to sacrifice safety is not new. Existing work on AI races and risky technology races shows that competitive pressure can lead actors to underinvest in safety, deploy unsafe systems, or cut corners on governance (Armstrong, Bostrom and Shulman, 2016; Cave and ÓhÉigeartaigh, 2018; Askell, Brundage and Hadfield, 2019; Han et al., 2021; Jensen, Emery-Xu and Trager, 2023; Emery-Xu, Park and Trager, 2024). This paper studies how those racing incentives are mediated by market structure and resource constraints. Firms allocate a fixed resource budget between speed and caution, so an intervention can affect risk by relaxing a firm's private resource constraint, changing the number of competitors, reallocating resources across firms, or making restraint credible. This framework lets us compare resource expansion, entry, consolidation, public entry, and commitment as different ways of changing the same underlying equilibrium allocation problem.

[1] Doom is not meant to represent a single mechanism; it encompasses possibilities such as autonomous systems taking over or disrupting critical infrastructure and powerful systems complementing malicious human intent. See Hendrycks, Mazeika and Woodside (2023) for an overview of catastrophic risks.

[2] Text slightly cleaned up from minutes 12–14 of "Joe Carlsmith—Preventing an AI Takeover," interview by Dwarkesh Patel, transcript, `https://www.dwarkesh.com/p/joe-carlsmith?showTranscript=true`. Accessed May 2026.

[3] See, e.g., Yudkowsky (2023); Christiano, Shlegeris and Amodei (2018); Tegmark (2017); recent scientific consensus efforts such as Bengio et al. (2024); Cohen et al. (2024); strategic policy analysis such as Black et al. (2024); and roadmap initiatives such as AI 2027 (2025).

We model a race among N firms. Each firm has a fixed resource budget, which can be interpreted as compute, engineering effort, capital, managerial attention, or another scarce input to frontier development. Each firm chooses an allocation of its budget between speed and caution. Individual speed determines the firm's arrival rate at AGI. The race ends when the first firm arrives, and the payoff from AGI is then determined probabilistically. If the winner succeeds, it receives a prize. If the winner causes doom, every firm suffers a loss. Crucially, a firm that moves faster is more likely to arrive at AGI first, but has fewer resources to devote to caution, which lowers the probability that its arrival at AGI ends in success and increases the probability that it ends in doom.

We characterize the unique symmetric equilibrium and use the equilibrium first-order condition to decompose the private incentives for speed versus safety. The decomposition highlights four forces driving the race for AGI. First, individual speed affects the expected arrival time of AGI. This encourages speed when achieving AGI is valuable in expectation, but discourages speed when achieving AGI has negative expected value. Second, moving faster makes a firm more likely to arrive at AGI before its rivals, which creates a standard business-stealing incentive to increase speed. Third, when projects differ in safety, speed can also shift arrival away from dangerous rivals. This effect encourages a slower and safer firm to speed up and a faster and riskier firm to slow down (this marginal effect is zero at a symmetric strategy profile). Finally, there is the direct safety cost of speed: allocating more resources to speed rather than caution makes the firm's own arrival more likely to end in doom. This effect encourages individual firms to slow down.

The model yields four main implications about how market structure and resources shape the equilibrium allocation between speed and safety. First, safety incentives are weaker when more firms compete for AGI. Adding firms raises total equilibrium speed and brings the expected arrival time of AGI forward. It also pushes individual firms to prioritize speed over safety, which increases the probability of doom conditional on arrival. Overall, market consolidation is safety-enhancing.

Second, there is a critical market size N_c that determines whether the race for AGI is profitable in expectation. Below this cutoff, each firm's expected payoff from achieving AGI is positive in equilibrium, while above it the expected payoff is negative. When there are few firms, each firm is relatively likely to be the one that arrives first, so the incentive to make its own arrival safe is strong. When there are many firms, each firm is less likely to arrive first, so each internalizes less of the safety cost of its own speed. The race can therefore become privately self-defeating: firms may continue to race even though, conditional on an arrival occurring, each firm's expected payoff is negative.

Third, additional resources have ambiguous effects on safety. More resources always increase equilibrium speed and bring AGI forward in expectation. But doom risk depends on the safety ratio, not on speed alone, and marginal incentives for caution depend on the number of competitors. When $N < N_c$, achieving AGI is valuable in expectation, and firms use additional resources disproportionately for caution. When $N > N_c$, achieving AGI has negative expected value, and firms use additional resources disproportionately for speed. Thus resource abundance is not mechanically good or bad for safety; its effect depends on market structure.

Fourth, the social planner has various policy levers at its disposal, and the optimal policy may involve a somewhat unexpected combination of them. An unconstrained planner who controls the allocation of resources across firms and between speed and caution has no intrinsic preference over the number of firms, but equalizes safety ratios across active projects and chooses a lower aggregate speed than the market would. A planner who controls only the number of firms prefers a monopoly.

At the same time, the planner may benefit from increasing industry resources. When the race is sufficiently concentrated and the social value of arrival is positive, increasing industry resources induces individual firms to allocate these resources disproportionately to safety. AGI arrives earlier, its arrival is safer and social welfare increases. In contrast, when the market is too crowded, firms allocate additional resources disproportionately to speed, making a race with negative value even riskier. In that case, the optimal policy is to shut the sector down. Hence, somewhat paradoxically, the planner may want to combine industry consolidation with resource subsidies.

The planner may also improve welfare by entering the race itself. We consider a publicly owned firm whose speed is chosen by the planner while private firms continue to behave strategically. A slow public entrant is safe, and hence increases welfare if it wins. Such an event is unlikely, but this possibility has a strategic spillover on other firms. By slowing down, each individual firm increases the probability of losing to a cautious public entrant, which is less dangerous than losing to a fast private rival. This can soften private incentives and make the overall race slower and safer. Public entry is therefore useful only if it is structured as a cautious alternative, not as another aggressive racer.

The model also identifies a welfare-improving role for regulation operating as a commitment device. A firm may benefit from a binding constraint on its own speed because such a constraint induces rivals to slow down as well. Locally, such a commitment is Pareto-improving for all firms. This may help explain why some calls for regulation come from within the industry itself.[4]

The paper's central implication is that AGI risk is not determined by technical progress alone. It is determined by the way market structure, resources, and policy shape the equilibrium allocation of scarce inputs between speed and safety. The same aggregate increase in frontier-AI capacity can therefore have different effects depending on whether it takes the form of larger budgets for existing firms, entry by additional firms, consolidation among incumbents, or public provision. Resource limits may improve safety in a crowded market but reduce safety in a more concentrated market. Consolidation may have costs in terms of market and political power, but it can also reduce doom risk through the speed-safety channel. Public AI options may improve welfare, but only if they are structured as cautious entrants rather than as additional aggressive racers.

The analysis should not be read as a general argument for concentration, resource restriction, or public provision. The model instead provides a framework for studying how market structure, resource constraints, and policy shape frontier-AI racing. Other considerations, including market power, distribution, misuse, geopolitical competition, antitrust enforcement, and the governance quality of public institutions, remain outside the analysis.

1 Relationship to the Literature

This paper relates to five literatures: work on AI races and catastrophic risk, formal models of risky technology races, frontier-AI governance and verification, models of innovation and market structure, and theories of strategic risk-taking and brinkmanship.

First, the paper contributes to work on AI races and catastrophic risk from advanced technologies. The closest starting point is Armstrong, Bostrom and Shulman (2016), who model teams racing to build the first AI and show how the value of being first can lead teams to skimp on safety precautions. Cave and ÓhÉigeartaigh (2018) analyze the rhetoric and reality of an AI race for strategic advantage, emphasizing the risk that competitive pressure can encourage corner-cutting on safety and governance. Askell, Brundage and Hadfield (2019) frame responsible AI development as a collective-action problem: firms may prefer safer development if others do the same, but competition can increase the incentive to move faster and underinvest in safety, security, and social-impact evaluation. Han et al. (2021) study how rewards and punishments can mediate unsafe racing

[4]See, e.g., Cecilia Kang, "OpenAI's Sam Altman Urgens A.I. Regulation in Senate Hearing," *New York Times* May 16, 2023. Available at https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

incentives in AI and related technology races. Polborn (2025) studies a related class of dangerous-technology races in which actors choose how much restraint to impose on a technology that is valuable in competition but may trigger a shared catastrophe. Policy-facing work makes similar concerns central to AGI governance: recent analyses model AGI competition as a prisoner's dilemma driven by first-mover advantages and shared catastrophic risks, and emphasize the importance of coordination, verification, and strategic stability in the transition to AGI (Abraham, Kavner and Moon, 2025; Frelinger and Mueller, 2026).

Our contribution is to model a different margin in the race. Firms do not choose only a safety level, a restraint level, or whether to accelerate development. They choose how much of a resource budget to allocate between speed and caution. This lets us ask how the speed-safety margin changes with firm resources, market size, and the distribution of resources across firms. As a result, policy-relevant changes can differ from their partial-equilibrium intuition: a policy that relaxes a firm's private resource constraint need not have the same effect as a policy that increases the total resources deployed in the race.

Second, the paper is related to recent formal models of risky technology races and safety-performance tradeoffs. Jensen, Emery-Xu and Trager (2023) study a race in which developers allocate compute between safety and performance, with performance determining competitiveness and safety determining the probability of disaster. Their analysis shows that compute prices and targeted subsidies can have non-obvious effects on equilibrium safety. Emery-Xu, Park and Trager (2024) study international races to implement dangerous new technologies under different information environments, showing how uncertainty about capabilities affects safety choices and disaster risk. These papers share our premise that strategic competition can distort safety investment. We differ by embedding the speed-safety tradeoff in a firm-level resource-allocation model and using that to study market structure. This distinction lets us compare changes in per-firm resources, entry, consolidation, public entry, and commitment within a common equilibrium framework.

Third, the paper contributes to the emerging literature on government regulation of AI (e.g., Figueroa, Guadalupi and Lemus, 2026). The governance implications also connect to work on trustworthy AI development, frontier regulation, model evaluation, and compute governance. Brundage et al. (2020) argue that responsible AI development requires mechanisms that allow claims about safety, security, fairness, privacy, and compute use to be verified rather than merely asserted. Those mechanisms include third-party auditing, red teaming, incident sharing, audit trails, and high-precision compute measurement. Recent work on frontier-AI regulation emphasizes safety standards, registration and reporting requirements, pre-deployment risk assessment, external scrutiny, and compliance mechanisms (Anderljung et al., 2023). Work on model evaluation for extreme risks studies how evaluations for dangerous capabilities and alignment can inform responsible training, deployment, transparency, and security decisions (Shevlane et al., 2023). Work on compute governance emphasizes that compute is detectable, excludable, quantifiable, and produced through a concentrated supply chain, making it a potentially important lever for regulatory visibility, allocation, and enforcement (Sastry et al., 2024; Shavit, 2023). Our model is complementary: it does not design verification or evaluation mechanisms, but it explains why such mechanisms matter. If speed and safety mechanisms compete for scarce compute, engineering effort, or managerial attention, then governance must alter the incentives that determine how resources are allocated, not merely announce safety principles.

Fourth, the paper builds on classic models of innovation races and market structure. Formally, our race technology is related to the contest-success-function approach widely used to model probabilistic competition, in which relative effort determines the probability of winning (Tullock, 1980; Hirshleifer, 1989; Skaperdas, 1996; Konrad, 2009; Dechenaux, Kovenock and Sheremeta, 2015). In patent-race and R&D-race models, firms invest to arrive before their rivals, and competition can accelerate innovation through preemption and business-stealing incentives (Kamien and Schwartz, 1972; Loury, 1979; Lee and Wilde, 1980; Dasgupta and Stiglitz, 1980; Reinganum, 1981, 1982, 1983, 1985; Harris and Vickers, 1985, 1987). Schumpeterian models likewise study how market structure shapes innovation incentives, including nonmonotone relationships between competition and innovation (Aghion and Howitt, 1992; Aghion, Harris and Vickers, 1997; Aghion et al., 2001, 2005). These papers typically treat innovation as valuable conditional on success. In our setting, arrival can be catastrophic. This changes the role of standard racing forces. Business stealing still pushes firms toward speed, but speed also makes arrival less safe. The same force that accelerates innovation in standard models can therefore raise catastrophic risk by shifting resources away from caution.

The paper is especially close in spirit to work on the direction of innovation. Bryan, Lemus and Marshall (2022) show that when innovation in a field becomes more lucrative, the rate of innovation can rise while its direction becomes distorted: higher payoffs attract entry, and more intense R&D competition shifts effort toward quicker-to-invent but less valuable projects. Our model studies an analogous distortion in frontier AI. Competition affects not only how much effort firms exert, but how they allocate effort between speed and safety. The distortion is therefore not merely that firms race too much. It is that racing changes the direction of technical effort.

The paper also differs from standard innovation-race models in the policy comparisons it emphasizes. We distinguish per-firm resources from aggregate resources. When per-firm resources increase, firms move faster, but doom risk can fall because firms may devote a smaller share of their budgets to speed. When aggregate resources are fixed, adding firms reallocates the common resource pool toward speed and increases doom risk. Thus the effect of competition depends on whether entry brings new resources into the sector or divides an existing resource pool among more competitors. This distinction is central for applying innovation-race logic to AI, where talent, compute, capital, and safety expertise may be scarce in the short run.

The paper also relates to work on antitrust, cooperation, and innovation incentives. Segal and Whinston (2007) show how market structure can affect innovative investment when innovation payoffs are nonnegative. In our setting, the payoff from arrival may be negative once doom risk is taken into account. This changes the welfare interpretation of innovation incentives. Increasing resources or preserving rents does not simply increase socially valuable innovation. It changes both the timing of arrival and the safety of arrival. Similarly, consolidation has a direct safety implication in our model, not because monopoly is intrinsically desirable, but because reducing the number of competitors weakens business stealing and shifts the allocation of a fixed resource pool toward caution. This implication must be interpreted alongside legal and institutional constraints. Hua and Belfield (2021) emphasize the tension between AI-safety cooperation and competition law: cooperation may help firms develop AI systems more safely and responsibly, but agreements among competitors can also threaten the competitive process that antitrust law is designed to protect.

Fifth, the paper relates to theories of strategic risk-taking, arms races, and brinkmanship. Schelling's analysis of threats that leave something to chance emphasizes that actors may use risk as part of strategic competition (Schelling, 1960, 1966). Powell's model of nuclear brinkmanship is especially close in spirit. Powell argues that standard brinkmanship theory largely abstracts from the balance of military power, even though actual crises often feature a tradeoff between bringing more military power to bear and increasing the risk of catastrophic escalation (Powell, 2015). Bringing more power to bear can increase the probability of prevailing if conflict remains limited, but it can also increase the probability that events get out of control.

Our model studies an analogous tradeoff in technological competition. Speed increases the chance of arriving first, but it also raises the probability that arrival ends in doom. The analogy is not that firms deliberately threaten catastrophe in the way states may manipulate escalation risk.

Rather, the common structure is a power-risk tradeoff: actors can improve their chance of winning only by accepting a higher probability of a catastrophic outcome. The difference is that our risk comes from a resource allocation between speed and safety. This allows us to study how market structure, public entry, and geopolitical rivalry change the equilibrium position on that tradeoff.

Finally, the model contributes to a growing social scientific literature that examines the broad societal implications of highly capable artificial intelligence (e.g., Webb, 2019; Acemoglu and Restrepo, 2020; Eloundou et al., 2024; Gans, 2025; Westwood, Grimmer and Hall, 2025; Miyazaki and Hall, 2026).

2 The Model

There are $N \geq 1$ ex ante identical firms, indexed by $i \in \{1, \ldots, N\}$. Firm i has resources $R > 0$ and chooses speed $s_i \geq 0$ and caution $c_i \geq 0$ subject to

$$s_i + c_i = R.$$

After firms simultaneously allocate their resources, the race unfolds in continuous time. At time t, firm i has technological capacity $t \cdot s_i$. Firm i reaches AGI when $t \cdot s_i = \theta_i$, where

$$\theta_i \sim \mathrm{Exp}(\lambda)$$

independently across firms. The parameter λ captures underlying facts about how difficult it is to develop AGI. Thus the arrival time of firm i is

$$T_i := \inf\{t \geq 0 : t \cdot s_i \geq \theta_i\}.$$

The game ends at the first arrival time,

$$T := \min_j T_j,$$

and the first firm to arrive is the winner,

$$W := \arg\min_j T_j.$$

If the winner's choice is (s_W, c_W), doom occurs with probability that depends on the winner's caution per unit of speed, $\frac{c_W}{s_W}$. We assume this probability takes the following form

$$\pi(s_W, c_W) = \exp\left(-\alpha \frac{c_W}{s_W}\right), \qquad \alpha > 0$$

with the convention that $\frac{c_i}{s_i} = +\infty$ if $s_i = 0$. With complementary probability, success occurs.[5]

If success occurs, the winner receives $B > 0$ and all other firms receive zero. If doom occurs, every firm suffers loss $D > 0$. Payoffs are discounted at rate $\rho > 0$.

The exponential-race structure gives a static representation of the game with the following expected payoff for firm i:

$$U_i(\mathbf{s}, \mathbf{c}) = \mathbb{E}\left[e^{-\rho T}\left(B\mathbf{1}_{\{W=i\}}(1 - \pi(s_i, c_i)) - D\pi(s_W, c_W)\right)\right]. \tag{1}$$

[5] One can think of R as representing total compute. Using resources on caution increases safety but an equivalent increase in safety when firm progresses quickly requires a proportional increase in compute devoted to testing.

Because any best response with positive speed exhausts the budget, a speed choice $s \in (0, R]$ implies $c = R - s$, so we can write

$$q\left(\frac{R}{s}\right) := \pi(s, R-s) = \exp\left[-\alpha\left(\frac{R}{s} - 1\right)\right]. \tag{2}$$

The parameter α is intrinsic safety: a higher α means any given amount of caution per unit of progress makes doom less likely. Note that the probability of doom at arrival depends on speed only through the ratio of resources to speed, R/s. In what follows, we refer to R/s as the *safety ratio*.

Lemma 1 (Exponential race)**.** *Conditional on speeds $(s_1, \ldots, s_N)$, the arrival times are independent with $T_i \sim \mathrm{Exp}(\lambda s_i)$. Let $S := \sum_j s_j > 0$. Then*

$$T \sim \mathrm{Exp}(\lambda S), \qquad \mathbb{P}(W = i) = \frac{s_i}{S}, \qquad W \perp T,$$

and

$$\mathbb{E}[e^{-\rho T}] = \frac{\lambda S}{\lambda S + \rho}, \qquad \mathbb{E}[e^{-\rho T}\mathbf{1}_{\{W=i\}}] = \frac{\lambda s_i}{\lambda S + \rho}.$$

All proofs are found in the appendix.

Let $S(s) := \sum_{j=1}^N s_j$. Using this notation and Lemma 1, we can write (1) as follows

$$U_i(\mathbf{s}) = \underbrace{\frac{S}{S + \frac{\rho}{\lambda}}}_{\substack{\text{effective}\\ \text{discounting}}} \left[\underbrace{\frac{s_i}{S}}_{\substack{\text{winning}\\ \text{prob.}}} B\left(1 - \underbrace{q\left(\frac{R}{s_i}\right)}_{\substack{\text{doom}\\ \text{prob.}}}\right) - D\sum_{j=1}^N \underbrace{\frac{s_j}{S}}_{\substack{\text{winning}\\ \text{prob.}}} \underbrace{q\left(\frac{R}{s_j}\right)}_{\substack{\text{doom}\\ \text{prob.}}}\right] \tag{3}$$

with the convention $U_i(0, \ldots, 0) = 0$. Since discounting (ρ) and technological difficulty (λ) enter only through the ratio ρ/λ, except where explicitly noted, set $\lambda = 1$ to avoid notational clutter.

A strategy is a speed $s_i \in [0, R]$. We study Nash equilibria in symmetric strategies.

The following facts will be useful throughout our analysis.

Lemma 2. *For any $s \in (0, R]$,*

$$\frac{dq\left(\frac{R}{s}\right)}{ds} = q\left(\frac{R}{s}\right)\frac{\alpha R}{s^2}, \qquad \frac{d}{ds}\left(sq\left(\frac{R}{s}\right)\right) = q\left(\frac{R}{s}\right)\left(1 + \frac{\alpha R}{s}\right).$$

2.1 Discussion of assumptions

A few of our assumptions merit further comment.

First, we model the race for AGI as a winner-takes-all contest. This assumption is not meant to imply that all economic value from AGI literally accrues to a single firm. Rather, it captures, in reduced form, the strong first-mover advantage often emphasized in frontier AI discussions. (Cave and ÓhÉigeartaigh (2018) document the winner-take-all, or at least winner-take-most, rhetoric surrounding AGI development.) Importantly, our results do not depend on a stark form of winner-take-all payoffs. Suppose instead that, following successful AGI development, each non-winning firm receives a payoff $b > 0$. Then firm i's expected payoff conditional on success is

$$p_i B + (1 - p_i)b = b + p_i(B - b),$$

where p_i is its probability of winning. Thus, the incentive to win is governed by the payoff spread $B - b$, not by B alone. Our qualitative results therefore remain unchanged so long as b is not too large relative to B.[6]

Second, we assume that AGI may result in a catastrophic outcome that affects all firms, and possibly society more broadly. This assumption captures a possibility emphasized by both observers and industry leaders. The Center for AI Safety statement, signed by the CEOs of OpenAI, Google DeepMind, and Anthropic, states that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." (Center for AI Safety, 2023) OpenAI states that AGI could involve "serious risk of misuse, drastic accidents, and societal disruption" and that it will "operate as if these risks are existential" (OpenAI, 2023). Anthropic similarly treats catastrophic risks as a possible outcome of advanced AI development (Anthropic, 2023*b*), while Google DeepMind's Frontier Safety Framework is designed to identify future AI capabilities that could cause "severe harm" (Dragan, King and Dafoe, 2024). Hendrycks, Mazeika and Woodside (2023) provide an analytic overview of catastrophic risks posed by AI. These statements motivate modeling the risk of a doom event that negatively affects all societal actors.

Third, we assume that each firm has its own AGI arrival threshold θ_i. The assumption that the AGI arrival thresholds are not identical across firms reflects differences in the approaches adopted by different firms; as a result, two firms devoting the same resources to speed may arrive at a breakthrough at different times. The assumption that the θ_i are i.i.d., rather than correlated, across firms is made for tractability.

We also assume exponential distributions for tractability. This assumption separates the winner's identity from calendar time, makes total speed investment govern expected arrival, and gives a closed-form payoff.[7] It also has the attractive feature that the race looks the same regardless of calendar time. Hence, if firms were allowed to revise their speeds at some point during the race, they would choose not to do so.

Finally, the assumption that resources are fixed, rather than allowing firms to choose their resource expenditure subject to a cost function, is made to focus our analysis on the speed-safety tradeoff without having to account for the scale of investment. This fixed-resource assumption is meant to capture a short-run constraint, not to suggest that the frontier-AI sector cannot expand over time. In current frontier-AI development, the relevant inputs—large-scale compute, data-center capacity, energy, specialized engineering labor, and safety expertise—are costly and slow to expand. Frontier AI executives routinely describe compute, power, and data-center capacity as binding constraints. For example, press reports quote OpenAI executives as saying that limited compute has forced the company to make tradeoffs across product launches, and Microsoft CEO Satya Nadella has described power and powered data-center shells as a binding constraint on deploying AI GPUs (Spirlet, 2026; Giffiths, Brent D., 2025; Trueman, 2025). The International Energy Agency projects that data-center electricity consumption will more than double by 2030, with AI as the most important driver of that growth (International Energy Agency, 2025). In the short run, adding another frontier project may therefore divide scarce capacity rather than create an equal amount of new capacity. That said, our analysis does provide comparative static insights into how incentives

[6]One nuance arises for the social planner. If b is interpreted as a separate private payoff accruing to each non-winning firm, then adding firms mechanically creates additional residual payoffs, which could bias the planner toward more entry. A more natural interpretation is that b represents a share of a fixed aggregate residual surplus or spillover, in which case the societal benefit from b should remain constant with the number of firms, making all our results extend as long as b is sufficiently small.

[7]Much of this extends to a broader class of distributions: distributions satisfying proportional-hazards preserve winner-time separation, and Weibull-distributions preserve calendar-time scaling. The tractability of the exponential allows us to keep the analysis focused on substantive, strategic forces rather than distributional details.

change when industry resources expand.

3 Equilibrium

In this section, we characterize the unique symmetric equilibrium of the game and the value of the game to the firms.

3.1 The marginal effects of speed

To understand firm i's incentives and best response, fix s such that $s_j = s$ for all $j \neq i$, and differentiate (3) with respect to s_i. We obtain the following Lemma.

Lemma 3 (Marginal effects of speed)**.** *Suppose firm i chooses $s_i \in [0, R]$ while each of the other $N-1$ firms chooses s. Let*

$$S = s_i + (N-1)s \qquad \text{and} \qquad Q = s_i q\left(\frac{R}{s_i}\right) + (N-1)sq\left(\frac{R}{s}\right).$$

Then

$$\frac{\partial U}{\partial s_i}(s_i; s) = \underbrace{\frac{\rho}{(S+\rho)^2}\left[\frac{s_i}{S}B\left(1-q\left(\frac{R}{s_i}\right)\right) - D\frac{Q}{S}\right]}_{\text{net discounting}} + \underbrace{\frac{N-1}{S+\rho}\frac{s}{S}B\left(1-q\left(\frac{R}{s_i}\right)\right)}_{\text{business stealing}}$$
$$\underbrace{-\frac{N-1}{S+\rho}\frac{s}{S}\left(q\left(\frac{R}{s_i}\right)-q\left(\frac{R}{s}\right)\right)D}_{\text{doom stealing}} \underbrace{-\frac{1}{S+\rho}(B+D)q\left(\frac{R}{s_i}\right)\frac{\alpha R}{s_i}}_{\text{doom penalty}}. \tag{4}$$

At a symmetric profile, the doom-stealing term vanishes and the symmetric first-order condition is

$$\underbrace{\frac{\rho}{Ns+\rho}}_{\text{timing weight}}\underbrace{\left[\frac{B}{N}\left(1-q\left(\frac{R}{s}\right)\right) - Dq\left(\frac{R}{s}\right)\right]}_{\text{expected payoff at arrival}} + \underbrace{\frac{N-1}{N}B\left(1-q\left(\frac{R}{s}\right)\right)}_{\text{business stealing}} - \underbrace{(B+D)q\left(\frac{R}{s}\right)\alpha\frac{R}{s}}_{\text{doom penalty}} = 0. \tag{5}$$

The four terms in (4) identify the incentives created by an increase in speed and reveal how the race for AGI differs from other markets. The net-discounting term comes from the fact that speed changes when AGI arrives in discounted time. Its sign is the sign of the firm's expected payoff from achieving AGI. If achieving AGI is valuable in expectation, moving arrival forward is attractive, and each individual firm has an incentive to speed up; if achieving AGI is bad in expectation, moving arrival forward is costly, and each individual firm has an incentive to slow down.[8]

The business-stealing term captures the fact that speed makes firm i more likely to arrive before its rivals. This effect pushes each individual firm to speed up. The doom-stealing term captures the effect of changing which project arrives first when projects differ in safety. At a symmetric profile this term is zero because all projects are equally dangerous. The doom-penalty term captures the direct safety cost of speed: increasing speed reduces caution per unit of progress and therefore makes firm i's own arrival more likely to end in doom.

[8]It may seem natural to conjecture that the expected profit upon arrival must be positive in equilibrium. However, because doom triggered by other firms cannot be avoided by simply exiting the market, expected profits may be negative in this model. See our discussion in Section 4.2.

Setting $s_i = s$ in (4), multiplying it by $Ns + \rho$, and rearranging terms gives (5). Equation (5) will be the main guide to the subsequent analysis. If there were no discounting or the expected profit at arrival were zero, the first term would be zero, and the equilibrium would require that firms equalize the doom penalty and business stealing. That is, each firm would choose a speed at which a marginal increase in the probability of winning the race equals the marginal increase in the probability of doom upon winning. Note that these two terms depend on speed only via the safety ratio. If the expected profit at arrival is positive, however, firms have an additional incentive to speed up in order to bring the positive profit forward, which tilts the safety ratio down. If the expected profit at arrival is negative, firms have an additional incentive to slow down, which tilts the safety ratio up.

3.2 Equilibrium

The next result establishes that there is a unique symmetric equilibrium.

Proposition 1 (A unique symmetric equilibrium exists)**.** *There exists a unique symmetric Nash equilibrium. The equilibrium speed is interior, $s_N \in (0, R)$, and is the unique symmetric speed satisfying* (5)*. At equilibrium, each firm's expected payoff is*

$$U^* = \frac{Ns_N}{Ns_N + \rho}\left[\frac{B}{N}\left(1 - q\left(\frac{R}{s_N}\right)\right) - Dq\left(\frac{R}{s_N}\right)\right]. \tag{6}$$

4 Resources, Competition, and Fragmentation

This section studies how equilibrium changes with the number of firms and the resources available to them. The first comparative static increases the number of firms while holding resources per firm fixed. This captures entry that brings new resources into the race. In that case, individual speed rises and doom risk increases (Proposition 2). As the number of firms grows, each firm is less likely to win the prize and the race becomes riskier. Expected profit upon arrival therefore falls with N. There is a cutoff N_c, independent of per-firm resources, such that arrival has positive expected value when $N < N_c$ and negative expected value when $N > N_c$ (Proposition 3). This cutoff is important because it determines how additional resources affect safety and several other later results.

The second comparative static increases the resources available to each firm while holding the number of firms fixed. This always makes arrival earlier, but its effect on doom risk depends on whether N is below or above N_c (Proposition 4). When $N < N_c$, additional resources make arrival safer; when $N > N_c$, they make arrival riskier.

The third comparative static changes the number of firms while holding total industry resources fixed. This captures short-run fragmentation: more firms divide a common pool of compute, talent, capital, or other scarce inputs. Fragmentation increases total speed and doom risk (Proposition 5).

4.1 More Firms with Fixed Per-Firm Resources

We first analyze how the race changes when more firms compete while resources per firm remain fixed. This comparative static captures entry that brings new resources into the sector: as N increases, total industry resources increase as well. The short-run fragmentation case is different. If aggregate resources are relatively fixed, then adding firms divides a common pool of talent, compute, capital, or complementary inputs more thinly. We consider that case in Section 4.4.

Proposition 2 (More firms means earlier arrival and higher doom risk)**.** *In equilibrium, individual speed s_N is strictly increasing in N. Hence as N increases:*

(i) AGI arrives earlier;

(ii) the safety ratio goes down;

(iii) equilibrium doom risk $q\left(\frac{R}{s_N}\right)$ *increases.*

The intuition is easiest to see from the symmetric first-order condition in (5). Hold the safety ratio R/s fixed and increase N. At a symmetric profile, each firm wins with probability $1/N$. The marginal doom penalty from increasing speed is therefore governed by the probability level: the firm internalizes the extra doom risk from its own arrival only in states in which it is the firm that arrives first. As N increases, this probability falls directly as $1/N$.

The business-stealing incentive is governed by a different object. It depends not on the probability of winning, but on the marginal effect of speed on the probability of winning. When firm i increases its speed, it shifts probability of arriving first from its rivals to itself. At a symmetric profile, this marginal effect is proportional to $(N-1)/N^2$: adding rivals increases the amount of rival arrival probability that can be stolen, but the incremental effect of speed is spread over a larger race. Comparing the business-stealing margin $(N-1)/N^2$ to the doom-penalty margin $1/N$ gives the relative force of these two competing effects, $(N-1)/N$, which is the term that appears in (5). Thus, holding fixed R/s, increasing N raises the relative marginal value of speed.

This is why the result differs from the standard comparative static in contest models, where individual effort typically falls as the number of competitors increases (Dechenaux, Kovenock and Sheremeta, 2015). That standard dilution force is present: a marginal increase in speed has a smaller effect on a firm's probability of winning as the race becomes more crowded. But the internalized doom penalty is also diluted. The key difference is that the doom penalty is governed by the level probability $1/N$, while business stealing is governed by the marginal effect $(N-1)/N^2$. The former falls faster than the latter. Hence, relative to the doom penalty, the marginal value of speed rises with N. Restoring (5) therefore requires a lower safety ratio R/s, which means higher speed. More firms consequently imply earlier arrival and higher doom risk.

The preceding intuition focuses on the doom-penalty and business-stealing terms in (5). The first term in that equation, which captures the private arrival benefit from a marginal increase in speed, can move in either direction with N. This does not overturn the comparative static. When that term decreases with N, it decreases more slowly than the relative business-stealing force $(N-1)/N$ increases. The net effect is therefore that, at the old safety ratio, the marginal value of speed is too high relative to the marginal value of caution. Firms restore equilibrium by tilting their resource allocation toward speed.

4.2 Expected Profit upon Arrival

As more firms compete, each individual firm is less likely to win B. Moreover, by Proposition 2, the race becomes riskier and hence doom becomes more likely. Combining these two effects implies that the expected profit at arrival decreases in N.

Proposition 3 (Achieving AGI can have negative expected value)**.** *The expected payoff upon arrival,* $\frac{B}{N}\left(1-q\left(\frac{R}{s_N}\right)\right)-Dq\left(\frac{R}{s_N}\right)$ *decreases in* N*. Moreover, there exists a cutoff* $N_c = N_c(B, D, \alpha) > 1$ *which is independent of* R *such that*

$$\frac{B}{N}\left(1-q\left(\frac{R}{s_N}\right)\right)-Dq\left(\frac{R}{s_N}\right)\begin{cases}>0, & N<N_c,\\ <0, & N>N_c.\end{cases}$$

Somewhat surprisingly, N_c does not depend on the firms' resources R. To see why, consider the speed at which the expected payoff upon arrival is zero. At that point, the timing term in the first-order condition drops out, so the safety ratio is pinned down by the balance between business stealing and the doom penalty. Both depend on speed only through the safety ratio. When resources change, speed adjusts to keep that ratio, and hence doom risk, unchanged. The cutoff N_c is therefore independent of R.

A natural question is why firms would remain active if the AGI race provides negative expected value—that is, when $N > N_c$. The answer is that firms are willing to participate in a negative expected-value race because of the doom risk created by other firms. In Appendix B, we consider an extension where firms decide whether to enter, or take an outside option while still remaining exposed to the doom risk created by any other firms that enter the AGI race. We show that $N > N_c$ can certainly arise in the equilibrium of that extended game with endogenous entry.

4.3 More Resources with a Fixed Number of Firms

Increasing a firm's resources relaxes its budget constraint. The firm can use the additional resources to move faster, to make arrival safer, or both. The first effect is unambiguous: equilibrium speed increases with R. The effect on safety is more subtle because doom risk depends on the safety ratio R/s_N, not on speed alone.

Proposition 4. *As the per-firm budget R increases:*

(a) the equilibrium speed s_N increases, so AGI arrives earlier;

(b) the probability of doom conditional on arrival

- *(i) decreases if $N < N_c$;*
- *(ii) increases if $N > N_c$.*

The first part is straightforward. A larger budget allows each firm to do more, so equilibrium speed rises and AGI arrives sooner.

The second part turns on how firms use the additional resources. Suppose first that a firm scaled up speed and caution in the same proportion. Then the safety ratio R/s_N would not change, and doom risk conditional on arrival would not change either. The question is therefore whether a larger budget leads firms to put relatively more of the new resources into caution or relatively more into speed.

As the discussion surrounding Lemma 3 showed, the doom penalty and the business-stealing effect depend on speed only through the safety ratio. If these were the only concerns, a larger budget would lead firms to scale speed and caution together, leaving the safety ratio, and therefore doom risk, unchanged. Resources change doom risk because a larger budget also increases speed, and higher speed changes the value of arriving sooner. Whether firms scale caution more than speed or speed more than caution is determined by the direction in which that timing effect pushes.

That direction depends on whether arrival is good or bad in expectation. When $N < N_c$, AGI has positive expected value. Firms therefore have a reason to bring arrival forward. But as resources increase and the race becomes faster, this timing benefit becomes less important at the margin: there is less to gain from making an already-fast race slightly faster. Firms then put relatively more of the additional resources into caution. Speed still rises, but it rises less than proportionally with resources. The safety ratio R/s_N rises, and doom risk falls.

When $N > N_c$, AGI has negative expected value. In that case, timing works in the opposite direction. Firms are reluctant to bring a bad expected outcome forward. But as resources increase and the race becomes faster, this reluctance also matters less at the margin. The force holding firms back weakens, so firms put relatively more of the additional resources into speed. Speed rises more than proportionally with resources. The safety ratio R/s_N falls, and doom risk rises.

Thus more resources always make the race faster, but they do not always make it safer. Whether doom risk falls or rises depends on the sign of the expected payoff from arrival, which is exactly what the cutoff N_c captures.

4.4 Fragmentation with Fixed Industry Resources

As argued earlier, in the short run, aggregate resources in the sector may be relatively fixed. Hence, it is reasonable to ask what happens to the race when we change the number of firms while keeping total resources fixed.

To this end, suppose total resources are fixed at $\bar{R}$, and each firm receives $R_N = \bar{R}/N$. Increasing N then means more competitors but fewer resources for each individual firm. From Proposition 2, increasing N directly pushes firms toward higher speed and higher doom risk. From Proposition 4, changing per-firm resources can have a more nuanced effect on the safety ratio. The proposition below shows that, when total industry resources are fixed, the fragmentation effect dominates: the race becomes riskier as the number of firms increases.

Proposition 5 (Fragmentation means earlier arrival and higher doom risk)**.** *Suppose total resources are fixed at $\bar{R}$, so each firm has budget $\bar{R}/N$. Then total equilibrium speed $S_N = Ns_N$ is strictly increasing in N. Hence:*

(i) AGI arrives earlier;

(ii) the fraction of each firm's budget spent on speed,

$$\frac{s_N}{\bar{R}/N} = \frac{S_N}{\bar{R}},$$

is strictly increasing in N;

(iii) equilibrium doom risk $q\left(\frac{\bar{R}}{Ns_N}\right)$ is strictly increasing in N.

Individual speed $s_N = S_N/N$ may increase or decrease.

Read in reverse, the proposition identifies one safety channel through which consolidation can be beneficial. Consolidation reduces fragmentation: fewer firms divide the same resource pool, which weakens racing pressure and shifts resources away from speed and toward safety. This does not mean that competition is intrinsically bad or that concentration is desirable. In practice, concentration may also create market-power, accountability, distributional, and misuse concerns that are outside the model.

5 Prizes, Doom Costs, and Technical Progress

The previous section studied market structure and resources. This section studies four primitives that describe the economic and technological environment of the race. The prize B captures the private value of winning the AGI race. The doom loss D captures the costs firms internalize from catastrophe. The impatience parameter ρ captures the value of time: firms care more about arriving sooner when future payoffs are discounted more heavily. Finally, λ captures the productivity of technological development. A higher λ means that a given speed choice generates progress toward AGI more quickly, as would happen after improvements in model architectures, training techniques, research tooling, or other technological improvements that make frontier AI development more efficient.

These comparative statics serve two purposes. First, they establish that the model has the expected incentive effects with respect to the value of winning and the cost of doom. A larger prize or a smaller doom cost leads firms to move faster and take more risk. Second, they show why the cutoff N_c matters beyond the resource results above. Changes that affect the value of time or the productivity of the development technology have opposite effects on firms' speed choices depending on whether arrival has positive or negative expected value.

Proposition 6 (Baseline comparative statics)**.** *Fix N and let s_N denote the unique symmetric equilibrium speed.*

(a) As B increases or D decreases, s_N increases, AGI arrives earlier, and the probability of doom conditional on arrival increases.

(b) As impatience ρ increases, s_N increases, AGI arrives earlier, and the probability of doom conditional on arrival increases if $N < N_c$. If $N > N_c$, s_N decreases, AGI arrives later, and the probability of doom conditional on arrival decreases.

(c) As λ increases, AGI arrives earlier. Moreover, s_N and the probability of doom conditional on arrival decrease if $N < N_c$, while s_N and the probability of doom conditional on arrival increase if $N > N_c$.

The comparative static with respect to D is straightforward. A smaller doom loss makes unsafe arrival less costly, so firms move faster and choose a lower safety ratio. The comparative static with respect to B is slightly less mechanical because a larger prize makes both winning quickly and arriving safe arrival valuable. The competitive effect dominates: increasing the value of winning raises the return to speed, so firms move faster and doom risk conditional on arrival rises.

The effects of impatience ρ and technical productivity λ are more subtle because they operate through the value of timing. Restoring λ, the relevant timing weight is

$$\frac{\rho/\lambda}{Ns + \rho/\lambda}.$$

This term multiplies the expected payoff from achieving AGI. When arrival has positive expected value, timing incentives push firms toward speed and away from safety: they want to speed up the arrival of a valuable outcome. When arrival has negative expected value, timing incentives push against speed. By Proposition 3, the first case obtains when $N < N_c$, and the second obtains when $N > N_c$.

Increasing ρ strengthens this timing force. When $N < N_c$, the timing force favors speed, so greater impatience makes firms move faster and take more risk. When $N > N_c$, the timing force restrains speed, so greater impatience makes firms move more slowly and take less risk.

Increasing λ represents a more productive development technology. A given speed choice now produces AGI more quickly. In the first-order condition, this lowers the effective discount rate ρ/λ, so it weakens the timing force. When $N < N_c$, the timing force favors speed: firms want to bring a valuable outcome forward. A higher λ weakens that pro-speed force, so firms choose lower speed and become safer. When $N > N_c$, the timing force restrains speed: firms are reluctant to bring a bad expected outcome forward. A higher λ weakens that restraint, so firms choose higher speed and become riskier.

There is also a direct technological effect. In symmetric equilibrium, expected time to arrival is

$$\frac{1}{\lambda N s_N}.$$

Thus a higher λ directly accelerates the race even if firms do not change their speed choices. Part (c) shows that this direct effect dominates: AGI arrives earlier when the technology improves, even in the case $N < N_c$, where firms respond by choosing a slower speed.

6 Policy and Governance

This section considers four policy instruments. First, we study an unconstrained planner who can choose the number of active firms and how they allocate resources between speed and safety. Second, we study a constrained planner who can regulate only the total resources available to private firms. Third, we ask whether a firm might want binding regulation because regulation can serve as a commitment device. Finally, we ask whether a constrained planner can improve outcomes by introducing a publicly owned firm whose speed it controls.

Equilibrium behavior has two distinct inefficiencies. First, firms do not internalize the full social cost of the doom risk they create. Second, competition generates business-stealing incentives that push firms toward speed. To isolate these forces and study possible policy interventions, we consider a planner who values successful AGI but also places social weight on catastrophic losses.

Let $K \geq N$ denote the social weight on doom losses. The special case $K = N$ corresponds to a planner who counts only the doom losses borne by the N firms. Larger values of K capture catastrophic costs borne by unmodeled societal actors besides the firms.

6.1 The unconstrained planner

Suppose the planner controls total resources R and can choose both firm-level resource allocations $(R_1, \dots, R_N)$ and firm-level speeds $(s_1, \dots, s_N)$. The planner's welfare is

$$W_K\left(\{s_i, R_i\}_{i=1}^N\right) = \frac{B \sum_i s_i - (B + KD) \sum_i s_i \exp\left[-\alpha \left(\frac{R_i}{s_i} - 1\right)\right]}{\sum_i s_i + \rho}, \tag{7}$$

where terms with $s_i = 0$ are taken to be zero.

Let s_M denote the monopolist's speed with total budget R and private doom loss D.

Proposition 7 (The social planner wants to slow down the race)**.** *Fix $S := \sum_i s_i \in (0, R]$. Among all allocations with $\sum_i R_i = R$ and $\sum_i s_i = S$, the planner equalizes the safety ratio R_i/s_i across all active projects. Thus the planner's problem reduces to*

$$\max_{S \in [0,R]} W_K(S) := \frac{S\left[B - (B + KD) q\left(\frac{R}{S}\right)\right]}{S + \rho}, \tag{8}$$

and the number of active projects is irrelevant. The planner's aggregate speed S^{SP} satisfies

$$S^{\mathrm{SP}} < s_M$$

whenever $K \geq N > 1$. Consequently, with the same total resources, the planner chooses lower aggregate speed than the monopolist.

This comparison clarifies how the planner's optimum relates to the decentralized race. Under fixed aggregate resources, the $N = 1$ decentralized equilibrium is the monopolist's problem, so $S_1 = s_M$. Proposition 5 shows that decentralized aggregate speed is strictly increasing in N. Hence, for every $N \geq 2$,

$$s_M = S_1 < S_N.$$

The monopolist is therefore slower and safer than any competitive race with the same aggregate resource pool.

Proposition 7 says that, when $K > 1$, the planner goes even slower:

$$S^{\mathrm{SP}} < s_M < S_N \qquad \text{for all } N \geq 2.$$

Thus a planner who places more than private weight on doom losses chooses lower aggregate speed than the monopolist and, a fortiori, lower aggregate speed than any competitive race. In the boundary case $K = 1$, the planner's problem coincides with the monopolist's problem, so $S^{\mathrm{SP}} = s_M$; even then, the planner is slower and safer than every race with $N \geq 2$.

The first part of the proposition shows that the planner is not intrinsically opposed to the existence of multiple firms. Conditional on aggregate speed S, what matters for doom risk is the equilibrium safety ratio. The planner therefore allocates resources so that all active firms have the same resource-to-speed ratio. Once this is done, splitting the same aggregate speed across one firm or many has no effect on social welfare.

The second part identifies a fundamental source of inefficiency. The planner chooses lower aggregate speed than a monopolist with the same total resources whenever the planner places more than private weight on doom losses. The monopolist already avoids business stealing, but it still internalizes only its own doom loss. A planner who values catastrophic losses more heavily tilts the allocation further toward caution and delays arrival.

This distinction matters for interpreting the role of market structure. The problem with a decentralized race is not multiplicity of firms by itself. The problem is that decentralized competition gives firms private incentives to spend too much on speed and too little on caution. Thus, if aggregate resources are fixed and the planner cannot directly control firms' speeds or safety ratios, but can choose market structure, the planner chooses the most consolidated market structure, $N = 1$. Reducing N weakens business stealing incentives, lowers aggregate speed, and therefore reduces doom risk.

6.2 The constrained planner: regulating total resources

The unconstrained planner above has the ability to choose both speed and resource allocation. We now consider a planner with a more limited policy instrument: the constrained planner can regulate the total resources available to the industry, but cannot directly choose firms' speeds or safety ratios.

Formally, suppose there is an available resource pool $\bar{R} > 0$ and the government allows the industry access to $R \in [0, \bar{R}]$. Each of the N firms then receives R/N, and firms play the symmetric equilibrium characterized above. Let $W_K(R)$ denote planner welfare from (7), evaluated at the symmetric equilibrium induced by total resources R, with $R_i = R/N$ for all i.

Proposition 8 (Second-best resource regulation)**.** *Suppose $K \geq N$. Suppose also that the planner can choose only total resources $R \in [0, \bar{R}]$, which are divided equally across the N firms.*

If $N > N_c$, then $W_K(R)$ is strictly decreasing in R. Hence the planner chooses

$$R^{SP} = 0.$$

If $N < N_c$, then there is a cutoff $N_\ell(K, \bar{R})$, possibly at the boundary, such that

$$R^{SP} = 0 \qquad \textit{if} \qquad N > N_\ell(K, \bar{R}),$$

and

$$R^{SP} = \bar{R} \qquad \textit{if} \qquad N < N_\ell(K, \bar{R}).$$

The cutoff $N_\ell(K, \bar{R})$ is decreasing in K and increasing in $\bar{R}$, strictly whenever the cutoff is interior. In the benchmark in which $K = N$, the cutoff coincides with N_c.

When the planner controls only resources, the second-best policy is bang-bang: the planner either shuts down the industry or makes the maximum resources available to it. The reason is that resources do not merely change the amount of activity in the race. They also determine whether the race becomes safer or more dangerous. Which extreme is optimal depends on the number of firms and on how much weight the planner places on doom losses borne outside the racing firms.

The benchmark $K = N$ gives the cleanest intuition. In this case, the planner counts only the doom losses borne by the N firms in the race. Since firms are symmetric, the planner's expected welfare from arrival has the same sign as a firm's expected payoff from arrival. Proposition 3 shows that this sign is positive when $N < N_c$ and negative when $N > N_c$.

Consider first $N > N_c$. Arrival has negative expected value. By Proposition 4, increasing resources raises equilibrium speed and, when $N > N_c$, lowers the safety ratio. Thus AGI arrives sooner and is more likely to end in doom. Because arrival is socially bad in expectation, both effects reduce welfare. The planner therefore shuts the industry down: $R^{SP} = 0$.

When $N < N_c$, the same forces point in the opposite direction. Arrival has positive expected value. Increasing resources still raises equilibrium speed, but now Proposition 4 implies that it also raises the safety ratio and lowers doom risk conditional on arrival. Resources therefore bring a socially valuable arrival forward and make arrival safer. Both effects increase welfare, so the planner chooses $R^{SP} = \bar{R}$.

The case $K > N$ adds the losses borne by actors outside the racing firms. A higher K makes catastrophe more socially costly. As a result, a race that has positive expected value for the firms themselves may have negative expected social value once the planner counts broader doom losses. In that case, speeding up arrival is undesirable, and the planner sets $R^{SP} = 0$. This is why the planner's intervention cutoff satisfies $N_\ell(K, \bar{R}) < N_c$. A larger K shrinks the range of firm counts for which the planner supplies resources.

The comparative static with respect to $\bar{R}$ points in the opposite direction. When $N < N_c$, additional resources make the maximal-resource race safer. A government that can provide more resources can therefore support the industry over a wider range of market structures. Equivalently, a larger $\bar{R}$ raises $N_\ell(K, \bar{R})$: it does not merely increase the scale of intervention conditional on intervening; it makes intervention optimal for more values of N.

The comparative static with respect to $\bar{R}$ has a state-capacity interpretation. A low-capacity planner may optimally withhold resources because it cannot make the race safe enough to justify accelerating arrival. A high-capacity planner can support the industry over a wider range of market structures because, when $N < N_c$, larger resource provision makes the maximal-resource race safer. This does not imply that wealthy states should unconditionally subsidize national AI champions. The model abstracts from international externalities and geopolitical competition. The narrower point is that, for a planner that internalizes the relevant doom losses, greater public capacity expands the region in which intervention is welfare-improving.

This result qualifies simple calls for limiting industry resources. Resource controls are most valuable when the industry is highly competitive or when the planner places high weight on catastrophic losses suffered outside the firms. When the number of firms is small and the planner's doom weight is not too large, the second-best policy points in the opposite direction: consolidate the race into fewer firms and increase the resources available to the remaining firms. Fewer firms weaken business-stealing incentives, while greater resources allow those firms to devote a smaller share of their budgets to speed.

In frontier AI, this distinction maps naturally onto compute policy. A broad restriction on frontier compute, a targeted safety subsidy, and publicly provided compute need not have the same effect. Each changes a different object in the model: the resource constraint, the number of effective competitors, or the allocation of resources between speed and caution. Contemporary policy already distinguishes these margins. Export controls on advanced chips restrict access to a key input; the National AI Research Resource provides publicly supported compute and data access to researchers; and compute-governance proposals use reporting, monitoring, or threshold rules to condition access to large-scale training resources (Bureau of Industry and Security, 2024; National Science Foundation, 2024; U.S. Department of Energy, 2024; Sastry et al., 2024).

The result also gives the planner a useful diagnostic. The planner need not directly know N_ℓ or N_c to choose the right direction of policy. The key question is whether arrival is good or bad in expectation after accounting for the losses the planner cares about. If arrival has positive expected social value, increasing resources is desirable; if arrival has negative expected social value, the planner should withhold resources.

6.3 Demand for regulation as a commitment device

Clearly, the race has negative externalities that operate through both the business-stealing benefit and the doom risk. We now ask whether a firm might benefit from being able to commit to develop more slowly and safely. This is a more stringent question than whether firms would benefit from industry-wide regulation. The question is whether one firm might gain from binding its own behavior even if competitors remain formally unconstrained. Surprisingly, the answer is yes. By committing to move more slowly, a firm reduces its chance of winning, but that loss can be offset by a strategic benefit: rivals respond by moving more slowly as well.

Consider a two-stage variant of our model. In stage 1, firm 1 commits to a speed $x \in [0, R]$. In stage 2, firms $2, \ldots, N$ choose speeds simultaneously. For x near the simultaneous-move equilibrium speed s_N, let $a(x)$ be the common speed in the symmetric follower equilibrium, and define firm 1's payoff using (3) by

$$L(x) := \frac{xB(1 - q\left(\frac{R}{x}\right)) - D\big[xq\left(\frac{R}{x}\right) + (N-1)a(x)q\left(\frac{R}{a(x)}\right)\big]}{x + (N-1)a(x) + \rho}. \tag{9}$$

Proposition 9 (A firm can gain by committing to a slower speed)**.** *Let $N \geq 2$. There exists $\varepsilon > 0$ such that, for $x \in (s_N - \varepsilon, s_N + \varepsilon)$, the symmetric follower equilibrium $a(x)$ is well defined and satisfies $a'(s_N) > 0$. Moreover, there exists $\varepsilon' > 0$ such that*

$$L(x) > L(s_N) \quad \text{for all } x \in (s_N - \varepsilon', s_N).$$

Proposition 9 says that commitment power is valuable because it allows a firm to bind itself to greater caution and lower speed. The intuition comes from the doom-stealing term in (4). When firm 1 credibly moves more slowly, losing to firm 1 becomes less dangerous for the other firms because firm 1's arrival is less likely to result in doom. The other firms therefore also slow down. Firm 1 gives up some probability of arriving first, but it benefits from a slower and safer race overall. Commitment to caution is privately valuable because it induces caution by others.

This mechanism resembles a class of frontier-AI self-regulatory governance proposals in which developers publicly condition further scaling or deployment on meeting specified safety requirements. Anthropic (2023*a*)'s Responsible Scaling Policy is a prominent example. In its original form, the policy stated that the company would pause scaling or delay deployment of more capable models if its ability to scale outstripped its ability to comply with the corresponding safety procedures. Such commitments are often described as unilateral safety measures. Proposition 9 shows why they can also have strategic value: a credible commitment to move more slowly can induce rivals to move more slowly as well.

The next result shows that firm 1's commitment to go slower also benefits the other firms. If firm 1 commits to a slightly lower speed, each follower optimally chooses $a(x) < s_N$. Moreover, relative to the simultaneous-move equilibrium, each follower faces safer rivals and a less dangerous leader. Thus firm 1's commitment relaxes the race for all firms, not only for the committing firm.

Corollary 1 (Commitment to caution is Pareto-improving)**.** *All firms are better off when firm* 1 *can commit to a sufficiently small reduction in speed.*

The subsequent evolution of the Responsible Scaling Policy illustrates the limits of unilateral self-regulation. Anthropic later revised the policy and moved away from the original categorical commitment to pause scaling or delay deployment when safety procedures lagged, emphasizing instead a more conditional framework and greater transparency.[9] This is precisely the commitment problem highlighted by the model. A firm may want the industry to move more cautiously, but it may be unable to credibly bind itself unless rivals are similarly constrained.

The model studies unilateral commitment, but the same logic helps interpret industry calls for regulation.[10] A request for binding safety rules need not be read only as public-spirited restraint or as regulatory capture. In a race, common constraints can be privately valuable because they change rivals' incentives. A firm may prefer a regulated environment in which all firms move more cautiously to an unregulated environment in which unilateral caution mainly means falling behind.

6.4 A Publicly Owned Firm

The previous second-best analysis considered a planner who can regulate the resources available to private firms. A different instrument is public entry. Here, the planner cannot control private firms' speeds or resources, but can create a publicly owned firm and choose that firm's speed. The question is whether such a firm can improve welfare even when private firms continue to choose their speeds strategically.

This analysis is motivated by the fact that governments and public institutions are beginning to consider public AI systems as an alternative form of policy intervention. The Swiss AI Initiative's Apertus model is one example: it was developed by Swiss public research institutions, supported by national supercomputing infrastructure, and organized around civic principles including direct democracy, privacy protection, collective decision-making, sustainability, personal autonomy, and

[9] Anthropic's original RSP stated that following its AI Safety Level scheme implied a commitment to pause scaling and/or delay deployment when its ability to scale outstripped its ability to comply with the relevant safety procedures (Anthropic, 2023*a*). Anthropic's later RSP versions describe the framework as voluntary and introduce a more conditional structure (Anthropic, 2026). Contemporary coverage and governance commentary interpreted the 2026 revision as moving away from the earlier unilateral pause commitment in light of competitive pressures and the absence of binding industry-wide regulation (Perrigo, 2026; Griffiths, 2026; Sophie Williams and Jonas Freund, 2026).

[10] See, e.g., Cecilia Kang, "OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing," *New York Times*, May 16, 2023. Available at `https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html`.

human agency.[11] The point is not that existing public AI systems are close substitutes for frontier AGI projects. Rather, they illustrate the institutional possibility that a publicly governed AI developer may have objectives and constraints that differ from those of private frontier firms.

There are $N \geq 2$ private firms, each with budget $R > 0$. The planner introduces a public firm with resources $P \geq 0$ and chooses its speed $z \in [0, P]$. For $z > 0$, define the public firm's doom probability by

$$p\left(\frac{P}{z}\right) := \exp\left[-\alpha\left(\frac{P}{z} - 1\right)\right].$$

At $z = 0$, interpret the public entrant as inactive, so $zp\left(\frac{P}{z}\right) = 0$. Suppose the private firms treat the public firm as a Stackelberg leader. Let $s(z, P)$ denote the symmetric equilibrium speed of the private firms induced by the public entrant.

With social doom weight $K \geq N$, welfare (7) can be written as

$$W(z,P) := \frac{B(Ns(z,P) + z) - (B + KD)(Ns(z,P)q\left(\frac{R}{s(z,P)}\right) + zp\left(\frac{P}{z}\right))}{Ns(z,P) + z + \rho}. \tag{10}$$

The public entrant's arrival benefit and doom risk enter welfare just as the private firms' do. The doom-loss term remains KD: the public entrant is not itself counted as an additional doom victim.

Proposition 10 (A cautious public entrant improves welfare)**.** *Suppose $N \geq 3$ and $K \geq N$. Fix any private per-firm resource budget $R > 0$ and any public resource budget $P > 0$. Then there exists $\bar{z} \in (0, P]$ such that, for every $z \in (0, \bar{z})$,*

$$W(z, P) > W(0, 0).$$

Proposition 10 shows that public entry can be welfare-improving when the public project commits to a sufficiently cautious development policy. Introducing such a public entrant has two effects. The first is a direct effect: with positive probability, the public project wins the race, and because its arrival is very safe, replacing a private arrival with a public arrival raises expected welfare. The second is a strategic effect: public entry changes the equilibrium speed choices of the private firms.

This strategic effect is related to the commitment result above, in which one firm's commitment to slower development induces its rivals to slow down as well. In the proof of Proposition 10, we show that when the private race is already in the negative-arrival-value region, the public entrant has this same effect: it induces the private firms to slow down. When the private race is in the positive-arrival-value region, $N < N_c$, however, the strategic effect goes in the opposite direction: private firms speed up. Intuitively, adding the public project also increases competitive pressure, and greater competition strengthens firms' incentives to develop faster. Proposition 10 shows that, when the public project is sufficiently cautious, the direct welfare gain from adding a safe arrival channel dominates this strategic response.

The result in this section abstracts from the fiscal or opportunity cost of the public entrant's resources. It should therefore be read as a result about the strategic effect of public entry, not as a full cost-benefit analysis of public provision. If the public entrant requires resources P with social cost $c(P)$, the same argument applies whenever this cost is smaller than the equilibrium welfare gain generated by public entry. If public resources are sufficiently expensive, this cost can offset the benefit. The relevant implication is therefore not that public entry is always desirable, but that a sufficiently cautious public entrant can improve welfare when public resources are available at sufficiently low social cost.

[11] Nathan E. Sanders and Bruce Schneier, "Rewiring Democracy Now: Switzerland Shows Us an Alternative to Corporate AI," *The Renovator*, February 21, 2026.

7 Conclusion

This paper studies an AGI race in which firms allocate scarce resources between speed and safety. The central lesson is that catastrophic risk is not determined by technical progress alone. It depends on how market structure, resources, and policy shape the equilibrium tradeoff between arriving first and making arrival safe. Competition does not merely add more actors to the race. It changes the strategic value of speed and, therefore, what each actor chooses to do with its resources.

The model yields several implications. First, increased competition weakens safety incentives. When aggregate resources are fixed, fragmentation raises total speed and increases conditional doom risk by shifting the common resource pool toward speed. Second, there is a critical market size. When the race is relatively concentrated, firms have stronger incentives to make their own arrival safe, and the expected payoff from achieving AGI can be positive. When the race is crowded, each firm is less likely to win, the race is less safe, and firms may continue racing even though achieving AGI has negative expected value. Third, additional resources have ambiguous safety effects. More per-firm resources accelerate expected arrival, but they can either reduce or increase conditional doom risk depending on whether firms use the marginal resources disproportionately for safety or speed. Finally, policy affects risk by changing equilibrium incentives. Consolidation, resource controls, commitment devices, and public entry matter because they reshape the speed-safety margin.

The analysis is deliberately stylized. It abstracts from heterogeneous firms, endogenous investment in resources, uncertainty about safety technology, multi-stage development, and the opportunity cost of public resources. These are natural directions for future work. The core implication, however, is more general: the welfare effect of resources and competition cannot be evaluated separately from the equilibrium allocation of scarce inputs between speed and safety. This helps explain why the same policy debate can contain apparently opposed proposals—compute restrictions, public AI, consolidation, voluntary safety commitments, and industry demands for regulation. In the model, these proposals are not simply pro- or anti-innovation; they operate by changing the equilibrium allocation of scarce resources between speed and safety.

A Proofs

A.1 Notation and Preliminary Lemma

For the proofs, it is useful to rewrite the symmetric first-order condition (5) as $H(s) = 0$, where

$$H(s) := \left[B - (B + D)q\left(\frac{R}{s}\right)\left(1 + \frac{\alpha R}{s}\right)\right](Ns + \rho) - s\left[B - (B + ND)q\left(\frac{R}{s}\right)\right]. \tag{11}$$

For any $J > 0$, define

$$A_J(s) := B - (B + JD)q\left(\frac{R}{s}\right). \tag{12}$$

Note that

$$A_N(s) = B - (B + ND)q\left(\frac{R}{s}\right) \tag{13}$$

is simply N times the individual firm's expected payoff from achieving AGI when all firms choose speed s. Define

$$\Psi(s) := \left[B - (B + D)q\left(\frac{R}{s}\right)\left(1 + \frac{\alpha R}{s}\right)\right]. \tag{14}$$

Then the symmetric equilibrium condition (11) can be written as

$$H(s) = \Psi(s)(Ns + \rho) - sA_N(s) = 0. \tag{15}$$

Lemma 4. *When $N < N_c$, then $A_N(s_N) > 0$ and $\Psi(s_N) > 0$. When $N > N_c$, then $A_N(s_N) < 0$ and $\Psi(s_N) < 0$.*

Proof. This follows immediately from Proposition 3 and (15). Since this lemma is not used in the proof of Proposition 3, this completes the proof. □

A.2 Proof of the exponential-race lemma

Proof. If $\theta_i \sim \text{Exp}(\lambda)$ and $s_i > 0$, then $T_i = \theta_i/s_i \sim \text{Exp}(\lambda s_i)$. Independence is preserved across firms. The formulas follow from the standard properties of independent exponential random variables and

$$\mathbb{E}[e^{-\rho T}\mathbf{1}_{\{W=i\}}] = \mathbb{P}(W = i)\mathbb{E}[e^{-\rho T}] = \frac{\lambda s_i}{\lambda S + \rho}.$$

For any firm choosing $s_i = 0$, we adopt the standard convention that $T_i = +\infty$ almost surely. □

A.3 Proof of Lemma 3

Proof. Using Lemma 2 in a simple differentiation of (3) with respect to s_i yields (4). Setting $s_i = s$ makes the doom-stealing term vanish and yields (5). □

A.4 Proof of Proposition 1

Proof. We first show that the symmetric first-order condition has a unique solution. By continuity, the function H defined in (11) has at least one root in $(0, R)$ because

$$\lim_{s \downarrow 0} H(s) = B\rho > 0$$

and, since $q\left(\frac{R}{R}\right) = q(1) = 1$,

$$H(R) = -\alpha(B+D)(NR+\rho) - D\rho < 0.$$

It remains to show that this root is unique. Differentiating (11) gives

$$H'(s) = (N-1)B - (B+D)q\left(\frac{R}{s}\right)\left[N\left(1+\frac{\alpha R}{s}\right) + \frac{(\alpha R)^2}{s^3}(Ns+\rho)\right] + (B+ND)q\left(\frac{R}{s}\right)\left(1+\frac{\alpha R}{s}\right),$$

and

$$H''(s) = \frac{(\alpha R)^2}{s^5}q\left(\frac{R}{s}\right)\Big([B+ND+N(B+D)]s^2 + (B+D)(3\rho - N\alpha R)s - \alpha\rho R(B+D)\Big).$$

The bracket above is a convex quadratic in s and is negative at $s = 0$. Hence it has at most one positive root. Thus H'' changes sign at most once, and if it changes sign, it changes from negative to positive. Therefore H' is either everywhere decreasing or first decreasing and then increasing.

This shape rules out more than one zero of H. To see this, suppose toward a contradiction that H has two distinct roots in $(0, R)$. Since $H(0) > 0$ and $H(R) < 0$, H must cross from positive to negative at its first root. For H to have another root, it must subsequently turn upward and cross zero from below. But once H' has turned upward after the unique possible sign change in H'', H' cannot turn downward again. Thus any upward crossing would leave H weakly increasing thereafter, contradicting $H(R) < 0$. Hence H crosses zero exactly once. Let s_N denote the unique root.

We now verify that the symmetric first-order-condition solution is a Nash equilibrium. Fix a common speed s for the other $N-1$ firms, and let firm i choose speed s_i. Using the following notation

$$K \equiv (N-1)s + \rho$$

and

$$f(s_i; s) = s_i B - (B+D)s_i q\left(\frac{R}{s_i}\right) - D(N-1)sq\left(\frac{R}{s}\right),$$

we can write the utility as

$$U(s_i; s) = \frac{f(s_i; s)}{s_i + K}.$$

Thus

$$\frac{\partial U}{\partial s_i} = \frac{f_{s_i}(s_i; s)(s_i + K) - f(s_i; s)}{(s_i + K)^2}.$$

Since $(s_i + K)^2 > 0$, the sign of $\frac{\partial U}{\partial s_i}$ is the sign of the numerator. Differentiating the numerator we obtain that it is equal to $f_{s_i,s_i}(s_i; s)(s_i + K)$, and using Lemma 2 this is equal to

$$f_{s_i s_i}(s_i; s) = -(B+D)q\left(\frac{R}{s_i}\right)\frac{(\alpha R)^2}{s_i^3} < 0.$$

Thus $U(\cdot; s)$ is single-peaked: its derivative can cross zero at most once, and any crossing is from positive to negative. So for any symmetric strategy profile $\mathbf{s}$ of other firms, any root of the first-order condition is the unique best response for firm i. Since s_N is such a root when $\mathbf{s} = \mathbf{s}_N$, s_N is a Nash equilibrium.

Finally, any symmetric Nash equilibrium must solve the symmetric first-order condition $H(s) = 0$. Since H has exactly one root, the symmetric Nash equilibrium is unique. $\square$

A.5 Proof of Proposition 2

Proof. Fix $R > 0$ throughout, and index the symmetric first-order condition (11) by N. Let s_N be the unique root of $H_N(\cdot)$, and write

$$q_N := q\left(\frac{R}{s_N}\right), \qquad x_N := \frac{R}{s_N}.$$

For fixed s,

$$\begin{aligned} H_{N+1}(s) - H_N(s) &= s\left[B - (B+D)q\left(\frac{R}{s}\right)\left(1 + \alpha\frac{R}{s}\right)\right] + sDq\left(\frac{R}{s}\right) \\ &= s\left[B\left(1 - q\left(\frac{R}{s}\right)\right) - \alpha\frac{R}{s}(B+D)q\left(\frac{R}{s}\right)\right]. \end{aligned}$$

We claim that this expression is strictly positive at $s = s_N$. To see this, let

$$\omega_N := \frac{\rho}{Ns_N + \rho}$$

and define the doom-slope penalty

$$d_N := \alpha x_N(B+D)q_N.$$

The symmetric first-order condition (5) can be written as

$$d_N - \frac{N-1}{N}B(1-q_N) = \omega_N\left[\frac{B}{N}(1-q_N) - Dq_N\right].$$

Rearranging gives

$$\begin{aligned} B(1-q_N) - d_N &= \frac{B}{N}(1-q_N) - \omega_N\left[\frac{B}{N}(1-q_N) - Dq_N\right] \\ &= (1-\omega_N)\frac{B}{N}(1-q_N) + \omega_N Dq_N > 0. \end{aligned}$$

Therefore

$$H_{N+1}(s_N) - H_N(s_N) > 0.$$

Since $H_N(s_N) = 0$, this implies

$$H_{N+1}(s_N) > 0.$$

Also $H_{N+1}(R) < 0$, because $q(R/R) = q(1) = 1$. By uniqueness of the symmetric equilibrium root, the root of $H_{N+1}(\cdot)$ lies to the right of s_N. Hence

$$s_{N+1} > s_N.$$

It follows immediately that $S_N = Ns_N$ is strictly increasing in N, so expected arrival time $1/(\lambda S_N)$ is strictly decreasing. Since R is fixed, s_N/R is strictly increasing. Finally, $q(R/s)$ is strictly increasing in s for fixed R, so doom risk is strictly increasing in N. □

A.6 Proof of Proposition 3

Proof. Recall that the equilibrium condition can be written as (see (15)):

$$H(s) = \Psi(s)(Ns + \rho) - sA_N(s).$$

Let $\hat{s}_N$ solve $A_N(\hat{s}_N) = 0$, so

$$q\left(\frac{R}{\hat{s}_N}\right) = \frac{B}{B + ND}.$$

Using the definition of q,

$$\frac{\alpha R}{\hat{s}_N} = \alpha - \log\left(\frac{B}{B + ND}\right).$$

Define

$$\widehat{\Psi}(N) := \Psi(\hat{s}_N).$$

Substituting $\hat{s_N}$ into (14) gives

$$\widehat{\Psi}(N) = B\left(1 - \frac{B + D}{B + ND}\left[1 + \alpha - \log\left(\frac{B}{B + ND}\right)\right]\right).$$

The composed function $\widehat{\Psi}$ is single-crossing in N. Indeed,

$$\widehat{\Psi}'(N) = B\frac{D(B + D)}{(B + ND)^2}\left[\alpha - \log\left(\frac{B}{B + ND}\right)\right] > 0.$$

Moreover,

$$\widehat{\Psi}(1) = -B\left[\alpha + \log\left(\frac{B + D}{B}\right)\right] < 0,$$

while

$$\lim_{N\to\infty} \widehat{\Psi}(N) = B > 0.$$

Therefore there is a unique $N_c > 1$ such that

$$\widehat{\Psi}(N_c) = 0,$$

or equivalently $\Psi(\hat{s}_{N_c}) = 0$.

It remains to relate this break-even comparison to the equilibrium speed. At the break-even speed,

$$H(\hat{s}_N) = \Psi(\hat{s}_N)(N\hat{s}_N + \rho) = \widehat{\Psi}(N)(N\hat{s}_N + \rho),$$

because $A_N(\hat{s}_N) = 0$. Since $N\hat{s}_N + \rho > 0$, $H(\hat{s}_N)$ has the same sign as $\widehat{\Psi}(N)$.

If $N < N_c$, then $\widehat{\Psi}(N) < 0$, so $H(\hat{s}_N) < 0$. By Proposition 1, H has a unique root s_N and starts positive. Hence $s_N < \hat{s}_N$. Since q is decreasing in its argument,

$$\frac{R}{s_N} > \frac{R}{\hat{s}_N} \quad\Longrightarrow\quad q\left(\frac{R}{s_N}\right) < q\left(\frac{R}{\hat{s}_N}\right).$$

Together with $A_N(\hat{s}_N) = 0$, this implies $A_N(s_N) > 0$. Thus the expected payoff from achieving AGI is positive.

If $N > N_c$, then $\widehat{\Psi}(N) > 0$, so $H(\hat{s}_N) > 0$. Since H has a unique root and ends negative, this implies $s_N > \hat{s}_N$. Since q is decreasing in its argument,

$$\frac{R}{s_N} < \frac{R}{\hat{s}_N} \quad\Longrightarrow\quad q\left(\frac{R}{s_N}\right) > q\left(\frac{R}{\hat{s}_N}\right).$$

Together with $A_N(\hat{s}_N) = 0$, this implies $A_N(s_N) < 0$. Thus the expected payoff from achieving AGI is negative.

Holding R fixed, the expected payoff from achieving AGI is strictly decreasing in N. To see this, take any $N' > N$. By Proposition 2, $s_{N'} > s_N$. Therefore

$$\frac{R}{s_{N'}} < \frac{R}{s_N}.$$

Since q is strictly decreasing in its argument,

$$q\left(\frac{R}{s_{N'}}\right) > q\left(\frac{R}{s_N}\right).$$

Therefore

$$\frac{B}{N'}\left(1 - q\left(\frac{R}{s_{N'}}\right)\right) - Dq\left(\frac{R}{s_{N'}}\right) < \frac{B}{N'}\left(1 - q\left(\frac{R}{s_N}\right)\right) - Dq\left(\frac{R}{s_N}\right).$$

Moreover, since $s_N \in (0, R)$, we have $q\left(\frac{R}{s_N}\right) < 1$, and hence

$$\frac{B}{N'}\left(1 - q\left(\frac{R}{s_N}\right)\right) - Dq\left(\frac{R}{s_N}\right) < \frac{B}{N}\left(1 - q\left(\frac{R}{s_N}\right)\right) - Dq\left(\frac{R}{s_N}\right).$$

Combining the two inequalities gives

$$\frac{B}{N'}\left(1 - q\left(\frac{R}{s_{N'}}\right)\right) - Dq\left(\frac{R}{s_{N'}}\right) < \frac{B}{N}\left(1 - q\left(\frac{R}{s_N}\right)\right) - Dq\left(\frac{R}{s_N}\right),$$

which proves strict monotonicity.

Finally, using $\widehat{\Psi}(N_c) = 0$ and rearranging yields

$$(B + DN_c)\exp\left(-\frac{D(N_c - 1)}{B + D}\right) = Be^{-\alpha}.$$

The expression contains neither R nor ρ, which proves the final statement. □

A.7 Proof of Proposition 4

Proof. Write R for the per-firm resource endowment, and write $H(s; R)$ for the left-hand side of the symmetric first-order condition when the per-firm budget is R. Thus

$$H(s; R) = (Ns + \rho)\left[B - (B + D)q\left(\frac{R}{s}\right)\left(1 + \alpha\frac{R}{s}\right)\right] - s\left[B - (B + ND)q\left(\frac{R}{s}\right)\right].$$

First, equilibrium speed is increasing in R. Fix $R_2 > R_1$, and let $s_1 = s_N(R_1)$. For fixed s_1,

$$H_R(s_1; R) = \frac{\alpha q\left(\frac{R}{s_1}\right)}{s_1^2}\left[(B + D)\alpha R(Ns_1 + \rho) - s_1^2(B + ND)\right].$$

Since $q\left(\frac{R}{s_1}\right) > 0$, the sign of $H_R(s_1; R)$ is the sign of the linear term in brackets. Hence, as a function of R, $H(s_1; R)$ first decreases and then increases.

Moreover,

$$H(s_1; s_1) = -(Ns_1 + \rho)[D + \alpha(B + D)] + Ns_1 D = -\rho[D + \alpha(B + D)] - Ns_1\alpha(B + D) < 0,$$

while

$$\lim_{R\to\infty} H(s_1;R) = B\big[(N-1)s_1 + \rho\big] > 0.$$

Since $H(s_1;R_1) = 0$ and $R_1 > s_1$, the preceding single-trough shape implies that $H(s_1;R) > 0$ for every $R > R_1$. In particular,

$$H(s_1;R_2) > 0.$$

From Proposition 1, $H(\cdot;R_2)$ has a unique root and crosses zero from above. Therefore s_1 lies to the left of the root under R_2, so

$$s_N(R_2) > s_N(R_1).$$

Thus equilibrium speed is strictly increasing in the per-firm resource endowment.

It remains to determine how the speed share changes with resources. Let

$$x(R) := \frac{R}{s_N(R)}$$

denote the equilibrium safety ratio. Define

$$m(x) := B - (B+D)q(x)(1+\alpha x)$$

and

$$\ell_N(x) := B - (B+ND)q(x).$$

The symmetric first-order condition can then be written as

$$H(s;R) = (Ns+\rho)m\left(\frac{R}{s}\right) - s\ell_N\left(\frac{R}{s}\right) = 0.$$

Substituting $s = R/x$ and multiplying by x, the equilibrium condition becomes

$$F(x,R) := xH(R/x;R) = \rho x m(x) - R(\ell_N(x) - Nm(x)) = 0.$$

Equivalently, along the equilibrium path,

$$\rho x m(x) = R\big(\ell_N(x) - Nm(x)\big). \tag{16}$$

By the implicit function theorem,

$$x'(R) = -\frac{F_R(x(R),R)}{F_x(x(R),R)} = \frac{\ell_N(x(R)) - Nm(x(R))}{F_x(x(R),R)}.$$

At the equilibrium root,

$$F_x(x(R),R) = -s_N(R)H_s(s_N(R);R) > 0,$$

where the inequality follows from the strict single crossing established in Proposition 1. Hence

$$\operatorname{sgn} x'(R) = \operatorname{sgn}\left(\ell_N(x(R)) - Nm(x(R))\right).$$

Using (16), the term $\ell_N(x(R)) - Nm(x(R))$ has the same sign as $m(x(R))$. The original first-order condition also implies

$$(Ns_N(R)+\rho)m(x(R)) = s_N(R)\ell_N(x(R)),$$

so $m(x(R))$ and $\ell_N(x(R))$ have the same sign. Therefore

$$\operatorname{sgn} x'(R) = \operatorname{sgn} \ell_N(x(R)).$$

Finally,

$$\ell_N(x) = B - (B + ND)q(x) = N\left[\frac{B}{N}(1 - q(x)) - Dq(x)\right].$$

The term in brackets is the individual firm's expected payoff from achieving AGI. By Proposition 3, this payoff is positive when $N < N_c$ and negative when $N > N_c$. Hence

$$x'(R) > 0 \quad \text{if } N < N_c,$$

and

$$x'(R) < 0 \quad \text{if } N > N_c.$$

Since

$$\frac{s_N(R)}{R} = \frac{1}{x(R)},$$

the equilibrium speed share moves in the opposite direction from $x(R)$. Thus the speed share decreases when $N < N_c$ and increases when $N > N_c$. Finally, because doom risk strictly decreases in the safety ratio $x(R)$, doom risk also decreases when $N < N_c$ and increases when $N > N_c$. This proves the result. □

A.8 Proof of Proposition 5

Proof. Using R as the total resource, setting individual resources at R/N and individual speed at $s = S/N$, the equilibrium condition (11) can be written in terms of total speed:

$$\widetilde{H}(S) := H\left(\frac{S}{N}\right) = \left[B - (B+D)q\left(\frac{R}{S}\right)\left(1 + \frac{\alpha R}{S}\right)\right](S+\rho) - \frac{S}{N}[B - (B+ND)q\left(\frac{R}{S}\right)] = 0. \quad (17)$$

Hence a unique total speed $S_N \in (0, R)$ solving (17). For any fixed $S \in (0, R)$,

$$\widetilde{H}_{N+1}(S) - \widetilde{H}_N(S) = \frac{S}{N(N+1)}B(1 - q\left(\frac{R}{S}\right)) > 0.$$

Evaluating at the root S_N gives $\widetilde{H}_{N+1}(S_N) > 0$. Also $\widetilde{H}_{N+1}(R) < 0$ because $q\left(\frac{R}{R}\right) = 1$. Since $\widetilde{H}_{N+1}$ has a unique root, that root must lie to the right of S_N. Thus $S_{N+1} > S_N$. The claims about doom risk and speed share follow because q and S/R are strictly increasing in S. □

A.9 Proof of Proposition 6

Proof. At the unique root, Proposition 1 implies $H'(s_N) < 0$. Hence, for any parameter θ,

$$\frac{ds_N}{d\theta} = -\frac{\partial_\theta H(s_N)}{H'(s_N)},$$

so $ds_N/d\theta$ has the same sign as $\partial_\theta H(s_N)$.

To prove the results for D and B, we use the linearity of H in these parameters. Hence for D,

$$H_D(s) = -q\left(\frac{R}{s}\right)\left[\rho + (Ns + \rho)\frac{\alpha R}{s}\right] < 0,$$

so $ds_N/dD < 0$. For B, using $H(s_N) = 0$,

$$BH_B(s_N) = Dq\left(\frac{R}{s_N}\right)\left[\rho + (Ns_N + \rho)\frac{\alpha R}{s_N}\right] > 0,$$

so $ds_N/dB > 0$. Since $q\left(\frac{R}{s}\right)$ is increasing in s for fixed R, doom risk moves in the same direction as s_N, and the expected arrival date moves in the opposite direction.

For ρ,

$$H_\rho(s) = B - (B + D)q\left(\frac{R}{s}\right)\left(1 + \frac{\alpha R}{s}\right).$$

At the equilibrium root,

$$(Ns_N + \rho)H_\rho(s_N) = s_N[B - (B + ND)q\left(\frac{R}{s_N}\right)],$$

whose sign is the sign of $A_N(s_N)$. Proposition 3 gives the sign in terms of N_c. Finally, the unnormalized payoff depends on ρ and λ only through ρ/λ, so increasing λ has the opposite effect of increasing ρ.

Proof of part (c) We will prove here only that the expected arrival time decreases since the remaining results follow directly from the comparative statics with respect to ρ and fact that the equilibrium quantities depend on ρ and λ only via $\frac{\rho}{\lambda}$.

When λ is not normalized to one, the symmetric equilibrium condition (11) becomes

$$G(s; \lambda, R, \rho) := \left[B - (B + D)q\left(\frac{R}{s}\right)\left(1 + \frac{\alpha R}{s}\right)\right]\left(Ns + \frac{\rho}{\lambda}\right) - s[B - (B + ND)q\left(\frac{R}{s}\right)] = 0. \quad (18)$$

Let $s(\lambda, R, \rho)$ denote the unique solution to (18).

Consider the expected arrival date $T(\lambda) := \frac{1}{\lambda N s(\lambda, R, \rho)}$. Differentiating $T(\lambda)$, we obtain

$$T'(\lambda) = -\frac{s(\lambda, R, \rho) + \lambda s_\lambda(\lambda, R, \rho)}{N\lambda^2 s(\lambda, R, \rho)^2}. \quad (19)$$

To sign the numerator, note that the equilibrium equation (18) is homogeneous of degree one in (s, R, ρ):

$$G(cs; \lambda, cR, c\rho) = cG(s; \lambda, R, \rho).$$

Hence the unique root satisfies

$$s(\lambda, cR, c\rho) = c\, s(\lambda, R, \rho).$$

Differentiating with respect to c at $c = 1$ delivers

$$Rs_R(\lambda, R, \rho) + \rho s_\rho(\lambda, R, \rho) = s(\lambda, R, \rho).$$

Moreover, since s depends on ρ and λ only through ρ/λ, we have

$$s_\lambda = -\frac{\rho}{\lambda} s_\rho(\lambda, R, \rho).$$

Therefore the numerator of (19) is

$$s(\lambda, R, \rho) + \lambda s_\lambda(\lambda, R, \rho) = s(\lambda, R, \rho) - \rho s_\rho(\lambda, R, \rho) = Rs_R(\lambda, R, \rho).$$

By the resource comparative static established in Proposition 4, increasing per-firm resources strictly increases absolute equilibrium speed, $s_R(\lambda, R, \rho) > 0$. So the numerator of (19) is positive, establishing that $T'(\lambda) < 0$. □

A.10 Proof of Proposition 7

Proof. Fix total speed $S = \sum_i s_i > 0$. For active projects, define weights $\omega_i = s_i/S$ and safety ratios $y_i = R_i/s_i$. Then $\sum_i \omega_i = 1$ and

$$\sum_i \omega_i y_i = \frac{1}{S}\sum_i R_i = \frac{R}{S}.$$

Since $y \mapsto e^{-\alpha(y-1)}$ is convex,

$$\begin{aligned}\sum_i s_i e^{-\alpha(R_i/s_i-1)} &= S\sum_i \omega_i e^{-\alpha(y_i-1)}\\ &\geq Se^{-\alpha(R/S-1)} = Sq\left(\frac{R}{S}\right),\end{aligned}$$

with equality if and only if all active projects have the same ratio $R_i/s_i = R/S$. Substitution gives the one-dimensional problem (8). Any allocation with this common ratio implements the same value, including a one-project allocation.

For the speed comparison, let

$$V_L(S) := \frac{S[B-(B+LD)q\left(\frac{R}{S}\right)]}{S+\rho}.$$

The monopolist maximizes V_1 and the planner maximizes V_K. The first-order condition for V_L can be written as

$$P_L(S) := \rho[B-(B+LD)q\left(\frac{R}{S}\right)] - (S+\rho)(B+LD)q\left(\frac{R}{S}\right)\frac{\alpha R}{S} = 0.$$

Moreover,

$$\partial_L P_L(S) = -Dq\left(\frac{R}{S}\right)\left[\rho + (S+\rho)\frac{\alpha R}{S}\right] < 0.$$

Thus, at the planner's solution S_K^{SP}, $P_1(S_K^{\mathrm{SP}}) > P_K(S_K^{\mathrm{SP}}) = 0$. Since the monopolist problem has a unique interior maximizer and its first-order condition crosses from positive to negative, $S_K^{\mathrm{SP}} < s_M$.

Finally, by Proposition 5, with fixed total resources the competitive market's total speed is at least the monopolist speed and is strictly larger when $N > 1$. Hence S_K^{SP} is below the competitive market speed. □

A.11 Proof of Proposition 8

Proof. Let $S_N(R)$ denote the aggregate equilibrium speed induced by total resources R, and define the equilibrium safety ratio

$$x_N(R) := \frac{R}{S_N(R)}.$$

For any $J > 0$, define

$$\ell_J(x) := B - (B+JD)e^{-\alpha(x-1)}.$$

Thus $\ell_N(x_N(R))$ is N times the individual firm's expected payoff from achieving AGI. Moreover,

$$\ell_J'(x) = \alpha(B+JD)e^{-\alpha(x-1)} > 0,$$

so ℓ_J is increasing in the safety ratio.

In a symmetric equilibrium, planner welfare can be written as

$$W_K(R,N) = \frac{S_N(R)}{S_N(R)+\rho}\ell_K(x_N(R)). \tag{20}$$

Since total speed goes to zero as $R \to 0$, define

$$W_K(0,N) := \lim_{R\to 0} W_K(R,N) = 0.$$

First suppose $N > N_c$. By Proposition 3,

$$\ell_N(x_N(R)) < 0$$

for every $R > 0$. Since $K \geq N$,

$$\ell_K(x_N(R)) = \ell_N(x_N(R)) - (K-N)De^{-\alpha(x_N(R)-1)} < 0.$$

By Proposition 4, increasing resources raises aggregate speed $S_N(R)$ and, when $N > N_c$, lowers the safety ratio $x_N(R)$. Hence the multiplier $S_N(R)/(S_N(R)+\rho)$ increases in R, while $\ell_K(x_N(R))$ decreases in R and is negative. Therefore $W_K(R,N)$ is strictly decreasing in R. Since $W_K(0,N) = 0$, the planner chooses

$$R^{SP} = 0.$$

Now suppose $N < N_c$. By Proposition 4, increasing resources raises aggregate speed $S_N(R)$ and raises the safety ratio $x_N(R)$. Thus both

$$\frac{S_N(R)}{S_N(R)+\rho} \qquad \text{and} \qquad \ell_K(x_N(R))$$

are increasing in R.

If $W_K(\bar{R},N) \leq 0$, then

$$\ell_K(x_N(\bar{R})) \leq 0.$$

For any $R < \bar{R}$, $x_N(R) < x_N(\bar{R})$, so

$$\ell_K(x_N(R)) < \ell_K(x_N(\bar{R})) \leq 0.$$

Hence $W_K(R,N) < 0$ for all $R \in [0,\bar{R})$. Since $W_K(0,N) = 0$, the planner chooses $R^{SP} = 0$ (or is indifferent between $R^{SP} = 0$ and $R^{SP} = \bar{R}$ if $W_K(\bar{R},N) = 0$).

If $W_K(\bar{R},N) > 0$, then

$$\ell_K(x_N(\bar{R})) > 0.$$

For any $R < \bar{R}$, either $\ell_K(x_N(R)) \leq 0$, in which case $W_K(R,N) \leq 0 < W_K(\bar{R},N)$, or $\ell_K(x_N(R)) > 0$, in which case both the timing multiplier and $\ell_K(x_N(R))$ are strictly smaller than at $\bar{R}$. Thus

$$W_K(R,N) < W_K(\bar{R},N)$$

for all $R < \bar{R}$. The planner therefore chooses

$$R^{SP} = \bar{R}.$$

It remains to characterize when $W_K(\bar{R},N) > 0$. From (20),

$$W_K(\bar{R},N) > 0 \quad \Longleftrightarrow \quad \ell_K(x_N(\bar{R})) > 0.$$

Equivalently,

$$e^{-\alpha(x_N(\bar{R})-1)} < \frac{B}{B+KD},$$

or

$$x_N(\bar{R}) > 1 + \frac{1}{\alpha}\log\left(\frac{B+KD}{B}\right) =: \bar{x}_K.$$

By Proposition 5, $S_N(\bar{R})$ is strictly increasing in N under fixed aggregate resources. Hence

$$x_N(\bar{R}) = \frac{\bar{R}}{S_N(\bar{R})}$$

is strictly decreasing in N. Therefore there is a cutoff $N_\ell(K,\bar{R})$, possibly at the boundary, such that

$$x_N(\bar{R}) > \bar{x}_K \iff N < N_\ell(K,\bar{R}).$$

This gives the cutoff characterization.

Since $\bar{x}_K$ is increasing in K, a larger social doom weight makes the condition $x_N(\bar{R}) > \bar{x}_K$ harder to satisfy. Thus $N_\ell(K,\bar{R})$ is decreasing in K, strictly whenever the cutoff is interior. Since $x_N(\bar{R})$ is increasing in $\bar{R}$ when $N < N_c$, a larger resource ceiling makes the condition easier to satisfy. Thus $N_\ell(K,\bar{R})$ is increasing in $\bar{R}$, strictly whenever the cutoff is interior.

Finally, consider the benchmark in which the planner counts the doom losses borne by the N firms, so $K = N$. In that case, $\ell_N(x_N(R))$ has the same sign as the firms' aggregate expected payoff from achieving AGI. By Proposition 3, this sign is positive if $N < N_c$ and negative if $N > N_c$, independently of R. Hence the resource-regulation cutoff coincides with N_c. □

A.12 Proof of Corollary 1

Proof. Proposition 9 already shows that firm 1 is strictly better off for all $x < s_N$ sufficiently close to s_N. It remains to show that the followers are also better off.

Let

$$M(x) := U_F(a(x); a(x), x)$$

denote the payoff of a representative follower along the symmetric follower equilibrium. At $x = s_N$, we have $a(s_N) = s_N$, so $M(s_N)$ is the payoff of each firm in the simultaneous-move equilibrium.

Differentiating $M(x)$ at $x = s_N$ gives

$$M'(s_N) = U_y(s_N; s_N, s_N)a'(s_N) + (N-2)U_a(s_N; s_N, s_N)a'(s_N) + U_x(s_N; s_N, s_N),$$

where U_y is the derivative with respect to the follower's own speed, U_a is the derivative with respect to the common speed of another follower, and U_x is the derivative with respect to the leader's speed. The first term is zero by the follower's first-order condition. By symmetry, the marginal effect of any rival's speed on a firm's payoff at the symmetric profile is the same. From the proof of Proposition 9,

$$U_x(s_N; s_N, s_N) = U_a(s_N; s_N, s_N) = -\frac{F_x(s_N, s_N)}{Ns_N + \rho} < 0.$$

Therefore

$$M'(s_N) = -\frac{F_x(s_N, s_N)}{Ns_N + \rho}\left[1 + (N-2)a'(s_N)\right] < 0,$$

since $F_x(s_N, s_N) > 0$ and $a'(s_N) > 0$.

Thus, as x decreases below s_N, each follower's payoff increases locally. Hence there exists $\bar{\varepsilon} > 0$ such that $M(x) > M(s_N)$ for all $x \in (s_N - \bar{\varepsilon}, s_N)$. Combining this with Proposition 9, firm 1 and every follower are strictly better off than in the simultaneous-move equilibrium for all such x. □

A.13 Proof of Proposition 9

Proof. Fix a leader speed x and suppose firms $3, \dots, N$ choose common speed a. If a representative follower chooses y, its payoff is

$$U_F(y; a, x) = \frac{yB(1 - q\left(\frac{R}{y}\right)) - D[xq\left(\frac{R}{x}\right) + (N-2)aq\left(\frac{R}{a}\right) + yq\left(\frac{R}{y}\right)]}{x + (N-2)a + y + \rho}.$$

So $a(x)$ solves the following symmetric follower first-order condition at $y = a$:

$$\begin{aligned} F(a, x) := &\left[B - (B + D)q\left(\frac{R}{a}\right)\left(1 + \frac{\alpha R}{a}\right)\right](x + (N-1)a + \rho) \\ &- aB(1 - q\left(\frac{R}{a}\right)) + D[xq\left(\frac{R}{x}\right) + (N-1)aq\left(\frac{R}{a}\right)] = 0. \end{aligned} \tag{21}$$

At $x = a = s_N$, this is the symmetric equilibrium condition: $F(s_N, s_N) = H(s_N) = 0$. The implicit function theorem yields a unique continuously differentiable branch $a(x)$ near s_N, with $a(s_N) = s_N$, and

$$a'(s_N) = -\frac{F_x(s_N, s_N)}{F_a(s_N, s_N)} \tag{22}$$

Denote $q_N := q\left(\frac{R}{s_N}\right)$. Differentiate (21) with respect to x and evaluate it at $x = a = s_N$:

$$F_x(s_N, s_N) = B\left[1 - q_N\left(1 + \frac{\alpha R}{s_N}\right)\right].$$

Using $H(s_N) = 0$, this can be rewritten as

$$F_x(s_N, s_N) = \frac{s_N B(1 - q_N) + Dq_N\left[\rho + (Ns_N + \rho)\frac{\alpha R}{s_N}\right]}{Ns_N + \rho} > 0. \tag{23}$$

Since $H(s) = F(s, s)$,

$$H'(s_N) = F_a(s_N, s_N) + F_x(s_N, s_N).$$

Proof of Proposition 1 gives $H'(s_N) < 0$, which with (23) delivers

$$F_a(s_N, s_N) = H'(s_N) - F_x(s_N, s_N) < 0. \tag{24}$$

Using (23) and (24) in (22) delivers that $a'(s_N) > 0$.

Totally differentiating the leader payoff (9) along the equilibrium path gives

$$L'(s_N) = \frac{\partial U_1}{\partial s_1}(s_N, \dots, s_N) + (N-1)\frac{\partial U_1}{\partial s_2}(s_N, \dots, s_N)a'(s_N).$$

The first term is zero by the simultaneous-move first-order condition. A direct calculation gives

$$\frac{\partial U_1}{\partial s_2}(s_N, \dots, s_N) = -\frac{F_x(s_N, s_N)}{Ns_N + \rho} < 0.$$

Therefore

$$L'(s_N) = -\frac{N-1}{Ns_N + \rho}F_x(s_N, s_N)a'(s_N) < 0.$$

Continuity then implies the strict inequalities for speeds just below and just above s_N. □

A.14 Proof of Proposition 10

We first prove the following lemma, which establishes a useful property of the equilibrium speed when $z = 0$.

Lemma 5. *Suppose $3 \le N < N_c$. Fix $P > 0$ and let $s(z)$ denote the symmetric private response when the public entrant has speed z and resources P. Let $s_N := s(0)$ denote the equilibrium symmetric private speed from the model without public entry. Then*

$$q\left(\frac{R}{s_N}\right)(B+D)\frac{\alpha R}{s_N}\left(N+\frac{\rho}{s_N}\right) > B. \tag{25}$$

Proof. The equilibrium speed without the entrant satisfies (15). Using this condition, (25) is equivalent to

$$q\left(\frac{R}{s_N}\right)(B+D)\frac{\alpha R}{s_N}A_N(s_N) - B\Psi(s_N) > 0.$$

Define

$$E(s) := q\left(\frac{R}{s}\right)(B+D)\frac{\alpha R}{s}A_N(s) - B\Psi(s).$$

We need to show that $E(s_N) > 0$. To do so, we first implicitly define s_L and s_H such that $s_N \in (s_L, s_H)$ (Step 1). Moreover, we show that for any interval, $E(s)$ attains its minimum at one of the endpoints of this interval (Step 2), hence it is sufficient to show that $E(s_L) > 0$ and $E(s_H) > 0$, which we do in Step 3.

Step 1 From (13), $A_N(s)$ is positive at $s = 0$, decreases in s, and is positive at s_N by Lemma 4. So there exists a set (s_L, s_H) such that $s_N \in (s_L, s_H)$ and for all $s \in (s_L, s_H)$, $A_N(s) > 0$. Define s_H by $\Psi(s_H) = 0$ and s_L by $A_N(s_L) = N\Psi(s_L)$. Then by the equilibrium condition (15), $s_N \in (s_L, s_H)$, which is what we need.

Step 2 Differentiating $E(s)$, we obtain

$$E'(s) = \frac{\alpha R}{s^2}q\left(\frac{R}{s}\right)(B+D)\left(2\frac{\alpha R}{s}-1\right)A_N(s).$$

So for any s for which $A_N(s) > 0$, the sign of $E'(s)$ is the sign of $2\frac{\alpha R}{s} - 1$. Therefore E is either monotone in s or it rises first and then falls. In either case, for any interval (s_L, s_H), it attains the minimum at one of its endpoints.

Step 3 From Step 1, $\Psi(s_H) = 0$, and so

$$E(s_H) = q\left(\frac{R}{s_H}\right)(B+D)\frac{\alpha R}{s_H}A_N(s_H) > 0.$$

Also from Step 1, $A_N(s_L) = N\Psi(s_L)$, and hence

$$E(s_L) = q\left(\frac{R}{s_L}\right)(B+D)\frac{\alpha R}{s_L}N\Psi(s_L) - B\Psi(s_L) = \Psi(s_L)\left[Nq\left(\frac{R}{s_L}\right)(B+D)\frac{\alpha R}{s_L} - B\right].$$

Using $A_N(s_L) = N\Psi(s_L)$ again and (13) and (14), s_L satisfies

$$q\left(\frac{R}{s_L}\right)\left[(N-1)B + N(B+D)\frac{\alpha R}{s_L}\right] = (N-1)B. \tag{26}$$

Isolating $Nq\left(\frac{R}{s_L}\right)(B+D)\frac{\alpha R}{s_L}$ from that equation and using it in $E(s_L)$, we obtain

$$E(s_L) = \Psi(s_L)[N-2-(N-1)q\left(\frac{R}{s_L}\right)]B$$

So $E(s_L) > 0$ if and only if

$$q\left(\frac{R}{s_L}\right) < \frac{N-2}{N-1}.$$

It remains to show that this inequality holds when $N \geq 3$.

Let $\hat{s}$ be such that $q\left(\frac{R}{\hat{s}}\right) = \frac{N-2}{N-1}$. We will show that the definition of s_L, $A_N(s_L) = N\Psi(s_L)$, implies that $s_L < \hat{s}$, which then implies that $q\left(\frac{R}{s_L}\right) < \frac{N-2}{N-1}$ as needed. The way we show this is by evaluating the left-hand side of (26) at $\hat{s}$ and showing that it is greater than the right-hand side, and then arguing that this implies $s_L < \hat{s}$.

By the formula for q, $q\left(\frac{R}{\hat{s}}\right) = \frac{N-2}{N-1}$ implies that $\frac{\alpha R}{\hat{s}} = \alpha - \ln\frac{N-2}{N-1}$. Plugging these into the left-hand side of (26) we obtain

$$\begin{aligned}\frac{N-2}{N-1}\left[(N-1)B + N(B+D)(\alpha - \ln\frac{N-2}{N-1})\right] &> (N-2)B - \frac{N(N-2)}{N-1}(B+D)\ln\frac{N-2}{N-1} \\ &> (N-2)B - B\frac{N(N-2)}{N-1}\ln\frac{N-2}{N-1} > (N-1)B,\end{aligned}$$

where the last inequality holds for $N \geq 3$. Moreover, the left-hand side of (26) is 0 at $s = 0$, is single-peaked and is greater than $(N-1)B$ at $s = R$. So the fact that it is above $(N-1)B$ at $\hat{s}$ implies that $s_L < \hat{s}$ as needed. $\square$

Proof of Proposition 10. For fixed $P > 0$, note that

$$p\left(\frac{P}{0}\right) := 0, \qquad \frac{d}{dz}[zp\left(\frac{P}{z}\right)]\Big|_{z=0+} = 0. \tag{27}$$

Define

$$A^g(z) := B - (B+KD)p\left(\frac{P}{z}\right). \tag{28}$$

Using (28) and (12) in (10), welfare can be written as

$$W(z,P) = \frac{Ns(z,P)A_K(s(z,P)) + zA^g(z)}{Ns(z,P) + z + \rho}. \tag{29}$$

Let $s(z)$ denote the symmetric private response when the public entrant has speed z and resources P. Let $s_N := s(0)$ denote the equilibrium symmetric private speed from the model without public entry.

From (29),

$$\bar{W} = W(0,0) = \frac{Ns_N A_K(s_N)}{Ns_N + \rho}.$$

Using (3), the utility of firm i when facing $N-1$ other firms at speed s and one public entrant at speed z is

$$U_i(s_i; s, z) = \frac{s_i(B - q\left(\frac{R}{s_i}\right)(B+D)) - (N-1)sq\left(\frac{R}{s}\right)D - zp\left(\frac{P}{z}\right)D}{(N-1)s + z + s_i + \rho}.$$

Differentiating with respect to s_i and evaluating at $s_i = s$, we obtain the symmetric equilibrium condition for the speed of the private firms in the presence of a public entrant with speed z and resources P:

$$F(s,z) := \Psi(s)(Ns + z + \rho) - sA_N(s) + Dzp\left(\frac{P}{z}\right) = 0.$$

At $z = 0$, this condition reduces to the baseline equilibrium condition (15), or equivalently,

$$\Psi(s_N)(Ns_N + \rho) = s_N A_N(s_N). \tag{30}$$

By the local uniqueness and stability of the symmetric equilibrium without public entrant,

$$F_s(s_N, 0) < 0.$$

Moreover, using (27),

$$F_z(s_N, 0) = \Psi(s_N).$$

Hence the implicit function theorem gives

$$s_z(0) = -\frac{F_z(s_N,0)}{F_s(s_N,0)} = -\frac{\Psi(s_N)}{F_s(s_N,0)}.$$

Now define

$$\mathcal{W}(s,z) := \frac{NsA_K(s) + zA^g(z)}{Ns + z + \rho}. \tag{31}$$

Then

$$W(z,P) = \mathcal{W}(s(z,P), z).$$

At the baseline,

$$\bar{W} = \mathcal{W}(s_N, 0).$$

Partially differentiating (31) with respect to z, the direct effect of introducing an infinitesimally slow, and hence infinitesimally safe, public entrant is

$$\mathcal{W}_z(s_N, 0) = \frac{B - \bar{W}}{Ns_N + \rho}.$$

This is strictly positive because $\bar{W} < B$. Partially differentiating (31) with respect to s, we obtain

$$\mathcal{W}_s(s_N, 0) = \frac{N\left[\rho A_K(s_N) + s_N A_K'(s_N)(Ns_N + \rho)\right]}{(Ns_N + \rho)^2}.$$

Case with $N > N_c$ Since $K \geq N \geq N_c$, Lemma 4 implies that $A_N(s_N) < 0$ and $\Psi(s_N) < 0$, so $s_z(0) \leq 0$. Moreover,

$$A_K(s_N) \leq A_N(s_N) \leq 0.$$

Also,

$$A_K'(s) = -(B + KD)q'(s) < 0.$$

Therefore

$$\mathcal{W}_s(s_N, 0) < 0.$$

Combining this with $s_z(0) \leq 0$, we obtain

$$\mathcal{W}_s(s_N, 0)s_z(0) \geq 0.$$

Thus

$$\frac{d}{dz}W(z,P)\Big|_{z=0+} = \mathcal{W}_z(s_N,0) + \mathcal{W}_s(s_N,0)s_z(0) > 0.$$

Hence, for every fixed $P > 0$, all sufficiently small positive $z < P$ increase welfare.

Case with $3 \le N < N_c$ Since $N < N_c$, Lemma 4 implies that $A_N(s_N) > 0$ and $\Psi(s_N) > 0$, so $s_z(0) \ge 0$. Thus the previous argument does not apply, because the new entrant increases the equilibrium speed of the private firms. However, we show below that as long as $3 \le N$, $\frac{d}{dz}W(z,P)\Big|_{z=0+} > 0$.

Totally differentiating (31), we obtain

$$\begin{aligned}\frac{d}{dz}\mathcal{W}(z,P)\Big|_{z=0+} &= \frac{\partial \mathcal{W}(s,z)}{\partial z}\Big|_{(s,z)=(s_N,0)} + \frac{\partial \mathcal{W}(s,z)}{\partial s}\Big|_{(s,z)=(s_N,0)} s_z(0)|_{z=0} \\ &= \frac{B - \frac{N s_N A_K(s_N)}{N s_N + \rho}}{N s_N + \rho} - \frac{N\left[\rho A_K(s_N) + s_N A_K'(s_N)(N s_N + \rho)\right]}{(N s_N + \rho)^2}\frac{\Psi(s_N)}{F_s(s_N,0)}. \qquad (32)\end{aligned}$$

We first sign this derivative at $K = N$. Using (13), we have

$$\begin{aligned}&\rho A_N(s_N) + s_N A_N'(s_N)(N s_N + \rho) \\ &\quad = (N s_N + \rho)\left[\frac{\rho}{s_N}\Psi(s_N) - (B + ND)q\left(\frac{R}{s_N}\right)\frac{\alpha R}{s_N}\right].\end{aligned}$$

Substituting this expression into (32) and using (30), we obtain

$$\begin{aligned}\frac{d}{dz}W(z,P)\Big|_{z=0+,\,K=N} &= \frac{1}{N s_N + \rho}\Bigg[B - N\Psi(s_N) \qquad (33)\\ &\quad + \frac{N\Psi(s_N)}{F_s(s_N,0)}\left((B+ND)q\left(\frac{R}{s_N}\right)\frac{\alpha R}{s_N} - \frac{\rho}{s_N}\Psi(s_N)\right)\Bigg]. \qquad (34)\end{aligned}$$

Now use the formula for $F_s(s_N, 0)$. A direct differentiation of $F(s, z)$ gives

$$\begin{aligned}F_s(s_N,0) &= -\frac{\rho}{s_N}\Psi(s_N) + (B+ND)q\left(\frac{R}{s_N}\right)\frac{\alpha R}{s_N} \\ &\quad - (B+D)q\left(\frac{R}{s_N}\right)\left(\frac{\alpha R}{s_N}\right)^2\left(N + \frac{\rho}{s_N}\right). \qquad (35)\end{aligned}$$

Rearranging,

$$\begin{aligned}&(B+ND)q\left(\frac{R}{s_N}\right)\frac{\alpha R}{s_N} - \frac{\rho}{s_N}\Psi(s_N) \\ &\quad = F_s(s_N,0) + (B+D)q\left(\frac{R}{s_N}\right)\left(\frac{\alpha R}{s_N}\right)^2\left(N + \frac{\rho}{s_N}\right).\end{aligned}$$

Substituting this into (33) yields

$$\begin{aligned}\frac{d}{dz}W(z,P)\Big|_{z=0+,\,K=N} &= \frac{1}{N s_N + \rho}\Bigg[B - N\Psi(s_N) \\ &\quad + \frac{N\Psi(s_N)}{F_s(s_N,0)}\left(F_s(s_N,0) + (B+D)q\left(\frac{R}{s_N}\right)\left(\frac{\alpha R}{s_N}\right)^2\left(N + \frac{\rho}{s_N}\right)\right)\Bigg] \\ &= \frac{1}{N s_N + \rho}\left[B + \frac{N\Psi(s_N)(B+D)q\left(\frac{R}{s_N}\right)\left(\frac{\alpha R}{s_N}\right)^2\left(N + \frac{\rho}{s_N}\right)}{F_s(s_N,0)}\right].\end{aligned}$$

Since $F_s(s_N,0)<0$, $\frac{d}{dz}W(z,P)\Big|_{z=0+,\,K=N}>0$ if and only if

$$B\big(-F_s(s_N,0)\big)-N\Psi(s_N)(B+D)q\left(\frac{R}{s_N}\right)\left(\frac{\alpha R}{s_N}\right)^2\left(N+\frac{\rho}{s_N}\right)>0. \tag{36}$$

Using (35), the left-hand side of (36) equals

$$\begin{aligned} &B\left[\frac{\rho}{s_N}\Psi(s_N)-(B+ND)q\left(\frac{R}{s_N}\right)\frac{\alpha R}{s_N}\right]\\ &\quad+(B+D)q\left(\frac{R}{s_N}\right)\left(\frac{\alpha R}{s_N}\right)^2\left(N+\frac{\rho}{s_N}\right)\left[B-N\Psi(s_N)\right]. \end{aligned}\tag{37}$$

Also,

$$B-N\Psi(s_N)>0,$$

because

$$N\Psi(s_N)=\frac{NA_N(s_N)}{N+\frac{\rho}{s_N}}<A_N(s_N)<B.$$

Using (25) in (37), we obtain

$$\begin{aligned} &B\left[\frac{\rho}{s_N}\Psi(s_N)-(B+ND)q\left(\frac{R}{s_N}\right)\frac{\alpha R}{s_N}\right]\\ &\quad+(B+D)q\left(\frac{R}{s_N}\right)\left(\frac{\alpha R}{s_N}\right)^2\left(N+\frac{\rho}{s_N}\right)\left[B-N\Psi(s_N)\right]\\ &>B\left[\frac{\rho}{s_N}\Psi(s_N)-(B+ND)q\left(\frac{R}{s_N}\right)\frac{\alpha R}{s_N}+\frac{\alpha R}{s_N}\left(B-N\Psi(s_N)\right)\right]. \end{aligned}$$

Since

$$B-(B+ND)q\left(\frac{R}{s_N}\right)=A_N(s_N)=\Psi(s_N)\left(N+\frac{\rho}{s_N}\right),$$

the expression in square brackets becomes

$$\frac{\rho}{s_N}\Psi(s_N)+\frac{\alpha R}{s_N}\left[A_N(s_N)-N\Psi(s_N)\right]=\frac{\rho}{s_N}\Psi(s_N)\left(1+\frac{\alpha R}{s_N}\right)>0.$$

Therefore (36) holds, which establishes that $\frac{d}{dz}W(z,P)\Big|_{z=0+,\,K=N}>0$.

Finally, we extend the result from $K=N$ to every $K\geq N$. Differentiating (32) with respect to K, holding the baseline private equilibrium fixed, gives

$$\frac{\partial}{\partial K}\,\frac{d}{dz}W(z,P)\Big|_{z=0+}=\frac{NDq\left(\frac{R}{s_N}\right)}{(Ns_N+\rho)^2}\left[s_N-\left(\rho+\frac{\alpha R}{s_N}(Ns_N+\rho)\right)s_z(0)\right]. \tag{38}$$

Substituting

$$s_z(0)=-\frac{\Psi(s_N)}{F_s(s_N,0)},$$

the term in square brackets in (38) becomes

$$\frac{s_NF_s(s_N,0)+\Psi(s_N)\left(\rho+\frac{\alpha R}{s_N}(Ns_N+\rho)\right)}{F_s(s_N,0)}.$$

Using (35) and the baseline condition (30), the numerator simplifies to

$$\alpha R\left[B - q\left(\frac{R}{s_N}\right)(B+D)\frac{\alpha R}{s_N}\left(N + \frac{\rho}{s_N}\right)\right].$$

By the equivalent form of the lemma above, this numerator is negative. Since $F_s(s_N, 0) < 0$, the entire square bracket in (38) is positive. Therefore

$$\frac{\partial}{\partial K}\left.\frac{d}{dz}W(z,P)\right|_{z=0+} > 0.$$

Since the derivative is already positive at $K = N$, it is positive for every $K \geq N$:

$$\left.\frac{d}{dz}W(z,P)\right|_{z=0+} > 0.$$

Case with $N = N_c$ is omitted for brevity as it follows similar logic and N_c is generically not an integer. □

B Endogenous entry

The analysis thus far has either treated the number of firms as fixed or considered comparative static changes to competitiveness. Here we consider a short extension with endogenous entry, asking the number of firms that will enter the race in equilibrium.

Suppose there are $\mathcal{N}$ potential firms. Each potential firm begins with resources R. Before the race starts, each firm chooses whether to enter or to take an outside option worth $\bar{u} > 0$. If $n \leq \mathcal{N}$ firms enter, the entrants play the symmetric equilibrium of the n-firm race with per-firm resources R. Non-entrants receive the outside option, but they remain exposed to the doom risk created by the entrants.

Let s_n denote the symmetric equilibrium speed when n firms enter, and write

$$q_n := q\left(\frac{R}{s_n}\right), \qquad S_n := n s_n, \qquad \delta_n := \frac{S_n}{S_n + \rho}.$$

If $n-1$ other firms are entering, the payoff to firm i of entering is

$$E_n = \delta_n\left[\frac{B}{n}(1 - q_n) - D q_n\right].$$

It is useful to separate this payoff into its prize and doom components:

$$E_n = P_n - C_n,$$

where

$$P_n := \delta_n \frac{B}{n}(1 - q_n), \qquad C_n := \delta_n D q_n,$$

with $C_0 := 0$. The payoff to firm i of not entering is

$$\bar{u} - C_{n-1}.$$

Thus entry is privately optimal when $n-1$ other firms are entering if

$$P_n - C_n \geq \bar{u} - C_{n-1},$$

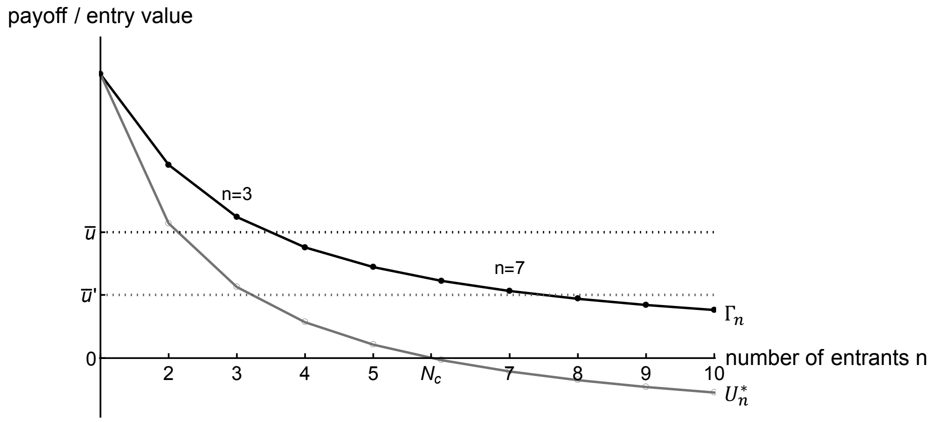

Figure 1: Endogenous entry can result in market sizes on either side of N_c. The black curve plots the private entry value Γ_n, and the gray curve plots the payoff from being in the race U_n^*, which crosses zero at $N_c \in (5,6)$. The dotted horizontal lines show two values of the outside option: $\bar{u}$ results in entry $n = 3 < N_c$, while $\bar{u}'$ results in entry $n = 7 > N_c$.

or equivalently when

$$\Gamma_n \geq \bar{u}, \qquad \Gamma_n := P_n - (C_n - C_{n-1}).$$

The relevant doom cost of entry is the incremental doom exposure $C_n - C_{n-1}$, not the full doom exposure C_n. The firm would already bear C_{n-1} if it stayed out.

A profile with $n \in \{1, \ldots, \mathcal{N} - 1\}$ entrants is a pure-strategy equilibrium if entrants do not want to exit and non-entrants do not want to enter. Using the definition of Γ_n, these two one-deviation conditions are

$$\Gamma_n \geq \bar{u} \qquad \text{and} \qquad \bar{u} \geq \Gamma_{n+1}.$$

With the conventions

$$\Gamma_0 = +\infty, \qquad \Gamma_{\mathcal{N}+1} = -\infty,$$

to cover no entry and full entry. Thus, in the entry extension, $n \in \{0, \ldots, \mathcal{N}\}$ is a pure-strategy equilibrium number of entrants if and only if

$$\Gamma_n \geq \bar{u} \geq \Gamma_{n+1}. \tag{39}$$

Condition (39) shows why endogenous entry need not prevent races with negative expected payoffs from arriving at AGI. Since non-entrants remain exposed to doom risk, a firm may prefer entering even when the continuation race has negative arrival payoffs.

Importantly, then, market sizes with $N < N_c$ or $N > N_c$ are consistent with endogenous entry. This is illustrated in Figure 1, which computes the value of Γ_n and U_n^* for the parameter values $(B = 1, D = 1, \alpha = 0.5, R = 1, \rho = 0.1)$. The point where U_n^* crosses zero is $N_c \in (5,6)$. When the outside option is $\bar{u}$, entry is $n = 3 < N_c$. When the outside option is $\bar{u}' < \bar{u}$, entry is $n = 7 > N_c$.

References

Abraham, Lisa, Joshua Kavner and Alvin Moon. 2025. A Prisoner's Dilemma in the Race to Artificial General Intelligence. Technical Report RR-A4245-1 RAND Corporation.

Acemoglu, Daron and Pascual Restrepo. 2020. "The wrong kind of AI? Artificial intelligence and the future of labour demand." *Cambridge Journal of Regions, Economy and Society* 13(1):25–35.

Aghion, Philippe, Christopher Harris and John Vickers. 1997. "Competition and Growth with Step-by-Step Innovation: An Example." *European Economic Review* 41(3–5):771–782.

Aghion, Philippe, Christopher Harris, Peter Howitt and John Vickers. 2001. "Competition, Imitation and Growth with Step-by-Step Innovation." *Review of Economic Studies* 68(3):467–492.

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith and Peter Howitt. 2005. "Competition and Innovation: An Inverted-U Relationship." *Quarterly Journal of Economics* 120(2):701–728.

Aghion, Philippe and Peter Howitt. 1992. "A Model of Growth Through Creative Destruction." *Econometrica* 60(2):323–351.

AI 2027. 2025. "AI 2027." Scenario and roadmap initiative.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs et al. 2023. "Frontier AI Regulation: Managing Emerging Risks to Public Safety." arXiv preprint arXiv:2307.03718.
URL: https://arxiv.org/abs/2307.03718

Anthropic. 2023*a*. "Anthropic's Responsible Scaling Policy." https://www.anthropic.com/news/anthropics-responsible-scaling-policy. Accessed May 2026.

Anthropic. 2023*b*. "Core Views on AI Safety: When, Why, What, and How.". Accessed: 2026-05-22.
URL: https://www.anthropic.com/news/core-views-on-ai-safety

Anthropic. 2026. "Anthropic's Responsible Scaling Policy: Version 3.0." https://www.anthropic.com/news/responsible-scaling-policy-v3. Accessed May 2026.

Armstrong, Stuart, Nick Bostrom and Carl Shulman. 2016. "Racing to the Precipice: A Model of Artificial Intelligence Development." *AI & Society* 31(2):201–206.

Askell, Amanda, Miles Brundage and Gillian Hadfield. 2019. "The Role of Cooperation in Responsible AI Development." arXiv preprint arXiv:1907.04534.
URL: https://arxiv.org/abs/1907.04534

Bengio, Yoshua, Geoffrey Hinton, Yuhuai Wu, Dawn Song, Pieter Abbeel, Yuval Noah Harari et al. 2024. "Managing Extreme AI Risks Amid Rapid Progress." *Science* 384(6703):842–845.

Black, James, Mattias Eken, Jacob Parakilas, Stuart Dee, Conlan Ellis, Kiran Suman-Chauhan, Ryan Bain, Harper Fine, Maria Chiara Aquilino, Mélusine Lebret and Ondrej Palicka. 2024. Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence. Technical Report RRA3295-1 RAND Corporation.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong et al. 2020. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." arXiv preprint arXiv:2004.07213.

Bryan, Kevin A., Jorge Lemus and Guillermo Marshall. 2022. "R&D Competition and the Direction of Innovation." *International Journal of Industrial Organization* 82:102841.

Bureau of Industry and Security. 2024. "Commerce Strengthens Export Controls to Restrict China's Capability to Produce Advanced Semiconductors for Military Applications." U.S. Department of Commerce press release. Accessed May 2026.

Cave, Stephen and Seán S ÓhÉigeartaigh. 2018. An AI race for strategic advantage: rhetoric and risks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. pp. 36–40.

Center for AI Safety. 2023. "Statement on AI Risk.". Accessed: 2026-05-22.
URL: https://www.safe.ai/statement-on-ai-risk

Christiano, Paul, Buck Shlegeris and Dario Amodei. 2018. "Supervising Strong Learners by Amplifying Weak Experts." *arXiv preprint arXiv:1810.08575* .

Cohen, Michael K., Nathan Kolt, Yoshua Bengio and Stuart Russell. 2024. "Regulating Advanced Artificial Agents." *Science* 384(6692):36–38.

Dasgupta, Partha and Joseph Stiglitz. 1980. "Industrial Structure and the Nature of Innovative Activity." *Economic Journal* 90(358):266–293.

Dechenaux, Emmanuel, Dan Kovenock and Roman M Sheremeta. 2015. "A survey of experimental research on contests, all-pay auctions and tournaments." *Experimental Economics* 18(4):609–669.

Dragan, Anca, Helen King and Allan Dafoe. 2024. "Introducing the Frontier Safety Framework.". Google DeepMind. Accessed: 2026-05-22.
URL: https://deepmind.google/blog/introducing-the-frontier-safety-framework/

Eloundou, Tyna, Sam Manning, Pamela Mishkin and Daniel Rock. 2024. "GPTs are GPTs: Labor market impact potential of LLMs." *Science* 384(6702):1306–1308.

Emery-Xu, Nicholas, Andrew Park and Robert Trager. 2024. "Uncertainty, Information, and Risk in International Technology Races." *Journal of Conflict Resolution* 68(10):2019–2047.

Figueroa, Nicolas, Carla Guadalupi and Jorge Lemus. 2026. "Regulation and responsible innovation." *Journal of Law, Economics, and Organization* .

Frelinger, David R. and Karl P. Mueller. 2026. The AGI Rideout Strategy for Reducing Strategic Risk and Promoting Stability in the Transition to Artificial General Intelligence. Technical Report PE-A4347-1 RAND Corporation.

Gans, Joshua. 2025. "Self-regulating artificial general intelligence." *AI & SOCIETY* pp. 1–10.

Giffiths, Brent D. 2025. "OpenAI Says This Chart Explains Its Need for More Compute." *Business Insider*. https://www.businessinsider.com/openai-chart-compute-future-plans-profitability-2025-12. Accessed May 2026.

Good, Irving John. 1966. Speculations concerning the first ultraintelligent machine. In *Advances in computers*. Vol. 6 Elsevier pp. 31–88.

Griffiths, Brent D. 2026. “Anthropic Is Dropping Its Signature Safety Pledge Amid a Heated AI Race.” *Business Insider*. Accessed May 2026.

Han, The Anh, Luís Moniz Pereira, Tom Lenaerts and Francisco C. Santos. 2021. “Mediating Artificial Intelligence Developments Through Negative and Positive Incentives.” PLOS ONE / arXiv preprint arXiv:2010.00403.
URL: https://arxiv.org/abs/2010.00403

Harris, Christopher and John Vickers. 1985. “Perfect Equilibrium in a Model of a Race.” *Review of Economic Studies* 52(2):193–209.

Harris, Christopher and John Vickers. 1987. “Racing with Uncertainty.” *Review of Economic Studies* 54(1):1–21.

Hendrycks, Dan, Mantas Mazeika and Thomas Woodside. 2023. “An overview of catastrophic AI risks.” *arXiv preprint arXiv:2306.12001* .

Hirshleifer, Jack. 1989. “Conflict and Rent-Seeking Success Functions: Ratio vs. Difference Models of Relative Success.” *Public Choice* 63(2):101–112.

Hua, Shin-Shin and Haydn Belfield. 2021. “AI & Antitrust: Reconciling Tensions Between Competition Law and Cooperative AI Development.” *Yale Journal of Law & Technology* 23:415–537.

International Energy Agency. 2025. “Energy and AI.” International Energy Agency report. Accessed May 2026.

Jensen, Mckay, Nicholas Emery-Xu and Robert Trager. 2023. “Industrial Policy for Advanced AI: Compute Pricing and the Safety Tax.” arXiv preprint arXiv:2302.11436.
URL: https://arxiv.org/abs/2302.11436

Kamien, Morton I. and Nancy L. Schwartz. 1972. “Timing of Innovations Under Rivalry.” *Econometrica* 40(1):43–60.

Konrad, Kai A. 2009. *Strategy and Dynamics in Contests*. Oxford: Oxford University Press.

Lee, Tom and Louis L. Wilde. 1980. “Market Structure and Innovation: A Reformulation.” *Quarterly Journal of Economics* 94(2):429–436.

Loury, Glenn C. 1979. “Market Structure and Innovation.” *Quarterly Journal of Economics* 93(3):395–410.

Miyazaki, Sho and Andrew B Hall. 2026. “Why Do AI Models Tell Left-Wing Voters to Support the Communist Party? AI Voting Advice in Japan’s 2026 General Election.”.

National Science Foundation. 2024. “Democratizing the Future of AI R&D: NSF to Launch National AI Research Resource Pilot.” National Science Foundation news release. Accessed May 2026.

OpenAI. 2018. “OpenAI Charter.” https://openai.com/charter/. Accessed May 2026.

OpenAI. 2023. “Planning for AGI and beyond.”. Updated October 28, 2025. Accessed: 2026-05-22.
URL: https://openai.com/index/planning-for-agi-and-beyond/

Perrigo, Billy. 2026. "Exclusive: Anthropic Drops Flagship Safety Pledge." *Time*. Accessed May 2026.

Polborn, Mattias. 2025. "Competitive Development of Dangerous Technologies." Working paper, April 27.

Powell, Robert. 2015. "Nuclear Brinkmanship, Limited War, and Military Power." *International Organization* 69(3):589–626.

Reinganum, Jennifer F. 1981. "Dynamic Games of Innovation." *Journal of Economic Theory* 25(1):21–41.

Reinganum, Jennifer F. 1982. "A Dynamic Game of R and D: Patent Protection and Competitive Behavior." *Econometrica* 50(3):671–688.

Reinganum, Jennifer F. 1983. "Uncertain Innovation and the Persistence of Monopoly." *American Economic Review* 73(4):741–748.

Reinganum, Jennifer F. 1985. "Innovation and Industry Evolution." *Quarterly Journal of Economics* 100(1):81–99.

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O'Keefe, Gillian K. Hadfield, Richard Ngo, Konstantin Pilz et al. 2024. "Computing Power and the Governance of Artificial Intelligence." arXiv preprint arXiv:2402.08797.
URL: https://arxiv.org/abs/2402.08797

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schelling, Thomas C. 1966. *Arms and Influence*. New Haven, CT: Yale University Press.

Segal, Ilya and Michael D. Whinston. 2007. "Antitrust in Innovative Industries." *American Economic Review* 97(5):1703–1730.

Shavit, Yonadav. 2023. "What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring." arXiv preprint arXiv:2303.11341.
URL: https://arxiv.org/abs/2303.11341

Shevlane, Toby, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt et al. 2023. "Model Evaluation for Extreme Risks." arXiv preprint arXiv:2305.15324.
URL: https://arxiv.org/abs/2305.15324

Skaperdas, Stergios. 1996. "Contest Success Functions." *Economic Theory* 7(2):283–290.

Sophie Williams and Jonas Freund. 2026. "Anthropic's RSP v3.0: How It Works, What's Changed, and Some Reflections." GovAI. https://www.governance.ai/analysis/anthropics-rsp-v3-0-how-it-works-whats-changed-and-some-reflections. Accessed May 2026.

Spirlet, Thibault. 2026. "OpenAI's CFO Says the Company Is Passing on Opportunities Because It Does Not Have Enough Compute." *Business Insider*. https://www.businessinsider.com/openai-cfo-says-compute-crunch-is-forcing-tough-trade-offs-2026-4. Accessed May 2026.

Tegmark, Max. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.

Trueman, Charlotte. 2025. "Microsoft Has AI GPUs Sitting in Inventory Because It Lacks the Power Necessary to Install Them." *Data Center Dynamics*. https://www.datacenterdynamics.com/en/news/microsoft-has-ai-gpus-sitting-in-inventory-because-it-lacks-the-power-necessary-to-install- Accessed May 2026.

Tullock, Gordon. 1980. Efficient Rent Seeking. In *Toward a Theory of the Rent-Seeking Society*, ed. James M. Buchanan, Robert D. Tollison and Gordon Tullock. College Station: Texas A&M University Press pp. 97–112.

U.S. Department of Energy. 2024. "DOE Advancing Safe and Secure AI Research Infrastructure Through National Artificial Intelligence Research Resource Pilot." U.S. Department of Energy news release. Accessed May 2026.

Webb, Michael. 2019. "The impact of artificial intelligence on the labor market." *Available at SSRN 3482150* .

Westwood, Sean J, Justin Grimmer and Andrew B Hall. 2025. "Measuring perceived slant in large language models through user evaluations." *Standford Business School* .

Yudkowsky, Eliezer. 2023. "Pausing AI Developments Isn't Enough. We Need to Shut It All Down." *Time*.