Preprints.org

Article

Not peer-reviewed version

Optimal Release Timing of AI Systems: A Strategic Analysis with Safety Externalities

Yijiashun Qi *

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2470.v1

Keywords: artificial intelligence; release timing; preemption game; safety regulation; technology policy

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Optimal Release Timing of AI Systems: A Strategic Analysis with Safety Externalities †

Yijiashun Qi

University of Michigan; elijahqi@umich.edu

† Preliminary and incomplete. Comments welcome.

Abstract

We study the strategic release timing of frontier AI systems by competing firms. Each firm develops a model whose quality improves with development time, but faces incentives to release early to capture first-mover advantages. Premature release imposes safety externalities on society that firms do not fully internalize. We characterize the symmetric Nash equilibrium in a preemption game and show that equilibrium release occurs strictly before the social optimum. We analyze four policy interventions: (i) minimum quality standards, which can implement the first-best; (ii) mandatory release delays, which paradoxically *reduce* deployed model quality by shifting preemption to the announcement stage, where quality locks in before the mandated waiting period; (iii) voluntary safety commitments, which can sustain cooperative outcomes when observable and credible; and (iv) Pigouvian safety taxes, which partially correct the externality but cannot eliminate the preemption distortion alone. Our results speak to ongoing policy debates about frontier AI regulation and generalize to other technologies with safety externalities and first-mover advantages.

Keywords: artificial intelligence; release timing; preemption game; safety regulation; technology policy

JEL Codes: L13; L50; O31; O38

---

1. Introduction

The development and release of frontier artificial intelligence (AI) systems has become one of the most consequential technology races of the 21st century. Firms such as OpenAI, Google DeepMind, Anthropic, and Meta invest billions of dollars in developing increasingly capable AI models, and face a fundamental strategic tension: releasing early captures market share and mindshare, but releasing later produces higher-quality, safer systems. This tension is not merely private—premature release of powerful AI systems can impose significant externalities on society through misuse, misinformation, labor displacement, and unforeseen risks.

A vivid illustration: in February 2026, Anthropic—the firm most publicly committed to AI safety—revised its Responsible Scaling Policy to drop its flagship commitment to "pause the scaling and/or delay the deployment of new models" when safety measures lagged behind capabilities. Chief scientist Jared Kaplan explained that unilateral restraint "wouldn't actually help anyone" if rival labs continued advancing. Meanwhile, in December 2025, OpenAI CEO Sam Altman declared "Code Red" internally in response to Google's Gemini surpassing ChatGPT on key benchmarks, redirecting resources from planned products to accelerate competitive response. These episodes are not anomalies—they are the predictable equilibrium outcome of a race with strong first-mover advantages and safety externalities that firms cannot unilaterally internalize.

Despite the centrality of release timing to the AI industry and to AI governance, no formal economic model captures this strategic problem. Jones (2024) studies the planner's tradeoff between

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

AI-driven growth and existential risk, but abstracts from strategic interaction between firms. We fill this gap by embedding the planner's problem in a competitive game.

We develop a continuous-time preemption game between two firms developing frontier AI systems. Model quality improves deterministically with development time, while the safety risk associated with deployment decreases. The first firm to release captures a disproportionate share of the market due to developer ecosystem lock-in, user switching costs, and reputational advantages. We make three main contributions.

First, we characterize the symmetric equilibrium of the release timing game and show that it involves *socially premature release*. The gap between equilibrium timing and the social optimum is increasing in the first-mover advantage and decreasing in the degree to which firms internalize safety costs. This result formalizes the widespread intuition that "AI races" lead to rushed deployment.

Second, we compare four regulatory instruments:

(i) *Minimum quality standards* (e.g., mandatory safety evaluations before release) can implement the social optimum by directly constraining the release decision.

(ii) *Mandatory release delays* (e.g., a required waiting period after announcing completion) paradoxically *reduce* the quality and safety of deployed models. Because development ceases upon announcement (quality lock-in), firms preempt on announcements, compressing the development window and deploying inferior systems. Any positive delay strictly worsens welfare relative to laissez-faire. This is our most striking result.

(iii) *Voluntary safety commitments* (e.g., Anthropic's Responsible Scaling Policy) can sustain cooperative delay as a subgame-perfect equilibrium when commitments are publicly observable and deviations are detectable.

(iv) *Pigouvian safety taxes*, set equal to the marginal social harm, partially correct the externality by making firms internalize safety costs. However, they cannot eliminate the preemption distortion—only a combination of the tax and a minimum standard achieves the first-best.

Third, we identify a *concentration paradox*: increasing competition among AI developers exacerbates the race and reduces welfare, because each additional firm makes preemption more attractive. This creates a genuine tension between antitrust goals and safety objectives that is unique to technologies with large safety externalities.

Fourth, we show that the model generates empirically testable predictions about the clustering of release dates, the relationship between capability jumps and release gaps, and the strategic use of staged rollouts.

Related Literature. This paper contributes to several literatures. The core model builds on the technology adoption timing literature initiated by Fudenberg and Tirole (1985) and the general innovation timing framework of Hoppe and Lehmann-Grube (2005), as well as the patent race literature (Loury 1979; Reinganum 1981). Our safety externality extension connects to the literature on technology regulation under uncertainty (Weitzman 1974) and the economics of catastrophic risk (Posner 2004; Weitzman 2009).

Most closely related within the AI economics literature is Jones (2024), who studies the tradeoff between AI-driven growth and existential risk in a social planner's framework. Our paper can be viewed as embedding his planner's problem inside a *strategic* environment: we show that competition between firms distorts the planner's optimum through preemption, and we characterize which policy instruments can restore it. Aschenbrenner and Trammell (2024) extend the growth-risk analysis to dynamic settings; our contribution is complementary, adding the firm-level strategic interaction that these planner models abstract from.

Several papers model AI race dynamics using different game-theoretic frameworks. Armstrong et al. (2016) provide the foundational conceptual model of AI development as a "race to the precipice," but their static 2×2 game does not capture timing or quality dynamics. Han et al. (2020) and Han et al. (2021) study AI race regulation using evolutionary game theory on finite populations; their strategic variable is the choice between safe and unsafe development, not the continuous-time release timing

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

decision that is our focus. Naudé and Dimitri (2020) model the AGI race as an all-pay contest with policy implications, but without safety externalities tied to release quality. LaCroix and Mohseni (2022) frame AI safety as a tragedy of the commons. We contribute to this literature by providing the first continuous-time preemption game of AI *release timing* with safety externalities, yielding closed-form equilibrium characterization and a complete policy ranking.

The broader AI economics context relates to (Acemoglu 2024; Agrawal et al. 2018; Korinek 2024) and AI governance (Askell et al. 2019; Dafoe 2018; Hendrycks et al. 2023). The concentration paradox connects to Vives (2008) on competition and innovation incentives. The result that mandatory delays backfire echoes findings in the financial regulation literature on how timing constraints can accelerate rather than slow strategic actions (Abreu and Brunnermeier 2003), and the pharmaceutical regulation literature where pre-market review freezes product quality at submission time (Peltzman 1973).

Outline. Section 2 presents the model. Section 3 characterizes the equilibrium. Section 4 analyzes welfare. Section 5 evaluates policy interventions. Section 6 considers extensions. Section 7 discusses empirical implications. Section 8 concludes.

2. Model

2.1. Environment

Time is continuous, $t \in [0, \infty)$. There are two firms, $i \in \{1, 2\}$, each developing a frontier AI system. Both firms begin development at $t = 0$ and choose a release time $t_i \geq 0$. The strategic variable is *when* to release—not *whether* to invest in safety (as in Armstrong et al. 2016) or *which strategy* to adopt (as in the evolutionary dynamics of Han et al. 2020). This distinction matters because release timing is the margin on which competitive pressure most visibly operates in practice.

Assumption 1 (Quality Technology)**.** *Model quality $q(t)$ is a function of development time alone, with $q : [0, \infty) \to [0, \bar{Q}]$ satisfying:*

(i) $q(0) = q_0 > 0$ *(initial quality),*
(ii) $q'(t) > 0$ *for all t (quality improves with development),*
(iii) $q''(t) < 0$ *for all t (diminishing returns),*
(iv) $\lim_{t\to\infty} q(t) = \bar{Q} < \infty$ *(bounded quality).*

The key substantive content of Assumption 1 is that quality improvement is *deterministic* and *concave*. This captures the empirical regularity that additional training compute, RLHF iterations, and red-teaming reliably improve model quality, but with diminishing returns.

Assumption 2 (Safety Risk)**.** *The social harm from deploying a model of quality q is $h(q)$, with $h : [0, \bar{Q}] \to [0, \infty)$ satisfying:*

(i) $h'(q) < 0$ *(higher quality ⇒ lower harm),*
(ii) $h''(q) > 0$ *(harm is convex in quality* deficiency*: releasing a very low-quality model is disproportionately harmful),*
(iii) $h(\bar{Q}) = 0$ *(a perfectly developed model causes no harm).*

Composing with the quality function, safety harm as a function of release time is $H(t) \equiv h(q(t))$, which satisfies $H'(t) < 0$ and $H''(t) > 0$ under our assumptions—harm falls with development time, and the marginal safety benefit of delay is greatest early on.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

Assumption 3 (Market Structure). *Consumer value from a model of quality q is $v(q)$, with $v'(q) > 0$ and $v''(q) \leq 0$. Market shares depend on release order:*

$$s_i(t_i, t_j) = \begin{cases} \alpha & \text{if } t_i < t_j \\ \frac{1}{2} & \text{if } t_i = t_j \\ 1 - \alpha & \text{if } t_i > t_j \end{cases}$$

where $\alpha \in (1/2, 1)$ parameterizes the first-mover advantage.

The parameter α captures the degree of developer ecosystem lock-in, user switching costs, and brand recognition advantages. When α is close to $1/2$, first-mover advantage is negligible; when α is close to 1, the market is winner-take-all.

Assumption 4 (Discounting). *Both firms and the social planner discount at rate $r > 0$, so the discount factor at time t is e^{-rt}.*

2.2. Firm's Problem

Firm i chooses release time t_i to maximize discounted profits:

$$\pi_i(t_i, t_j) = e^{-rt_i} \cdot v(q(t_i)) \cdot s_i(t_i, t_j) \tag{1}$$

This formulation embeds three forces:

- **Discounting** (e^{-rt_i}): Delay is costly because future profits are worth less.
- **Quality** ($v(q(t_i))$): Delay is beneficial because the model improves.
- **Preemption** (s_i): Releasing before the rival captures a larger market share.

2.3. Social Planner's Problem

The social planner chooses release times (t_1, t_2) to maximize total surplus net of safety harm:

$$W(t_1, t_2) = \sum_{i=1}^{2} e^{-rt_i} \left[v(q(t_i)) \cdot \frac{1}{2} - \lambda \cdot h(q(t_i)) \right] \tag{2}$$

where $\lambda > 0$ is the weight on safety harm. Note that the planner assigns equal market shares (no wasteful duplication of first-mover rents) and internalizes the safety externality.

3. Equilibrium Analysis

3.1. Monopolist Benchmark

As a benchmark, consider a single firm choosing its release time to maximize:

$$\pi^M(t) = e^{-rt} \cdot v(q(t))$$

The first-order condition yields:

$$\frac{v'(q(t^M)) \cdot q'(t^M)}{v(q(t^M))} = r \tag{3}$$

The monopolist releases when the *proportional* rate of quality improvement equals the discount rate. This is the classic innovation timing result: the firm balances the marginal value of waiting (better quality) against the cost of delay (discounting).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

3.2. Duopoly Equilibrium

In the duopoly, we look for a symmetric Nash equilibrium in release times. The key insight is that preemption creates a discontinuity in payoffs at $t_i = t_j$: a firm gains discretely by releasing an instant before its rival.

Lemma 1 (No Pure-Strategy Equilibrium with Interior Timing). *If $\alpha > 1/2$, there is no pure-strategy Nash equilibrium with $t_1 = t_2 = t > 0$ and $t < t^M$.*

Proof. Suppose both firms release at $t > 0$. Each earns $e^{-rt}v(q(t))/2$. By releasing at $t - \varepsilon$, firm i earns approximately $e^{-rt}v(q(t))\alpha > e^{-rt}v(q(t))/2$ for small ε, a profitable deviation. □

Following Fudenberg and Tirole (1985), we characterize the equilibrium using the concept of a *preemption time*: the earliest time at which a firm is willing to release if it expects the rival to release an instant later.

Definition 1 (Preemption Time). *The preemption time t^* is the earliest t such that:*

$$e^{-rt}v(q(t))\alpha \geq e^{-r\tilde{t}}v(q(\tilde{t}))(1-\alpha) \tag{4}$$

for all $\tilde{t} > t$, where the left side is the payoff from leading at t and the right side is the payoff from optimally following.

Proposition 2 (Equilibrium Release Timing). *In the unique symmetric equilibrium:*

(i) Both firms release at t^* defined by

$$e^{-rt^*}v(q(t^*))\alpha = \max_{\tilde{t}>t^*} e^{-r\tilde{t}}v(q(\tilde{t}))(1-\alpha) \tag{5}$$

(ii) t^* is strictly less than t^M (the monopolist's release time).

(iii) t^* is decreasing in α: stronger first-mover advantage leads to earlier release.

Proof. See Appendix A. The argument adapts the construction in Fudenberg and Tirole (1985) to our setting with continuous quality improvement. The key is that the leader's payoff $L(t) = e^{-rt}v(q(t))\alpha$ crosses the follower's optimized payoff $F(t) = \max_{\tilde{t}>t} e^{-r\tilde{t}}v(q(\tilde{t}))(1-\alpha)$ exactly once from below, defining t^*. □

Interpretation. The equilibrium has a "war of attrition" flavor: both firms would prefer to wait longer, but neither can credibly commit to waiting because the rival could preempt. The result is a race to the bottom in release timing.

3.3. Parametric Example

To build intuition, consider the following parametric specification:

$$q(t) = 1 - e^{-\gamma t}, \quad \gamma > 0 \tag{6}$$

$$v(q) = q \tag{7}$$

$$h(q) = (1-q)^2 \tag{8}$$

Under this specification, the monopolist's optimal release time is:

$$t^M = \frac{1}{\gamma}\ln\left(\frac{\gamma + r}{r}\right) \tag{9}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

The duopoly equilibrium time t^* solves:

$$e^{-rt^*}(1-e^{-\gamma t^*})\alpha = \max_{\tilde{t}>t^*} e^{-r\tilde{t}}(1-e^{-\gamma\tilde{t}})(1-\alpha) \tag{10}$$

The follower's optimal response given leader release at t^* is to release at t^M (since the follower faces no further preemption threat), yielding:

$$e^{-rt^*}(1-e^{-\gamma t^*})\alpha = e^{-rt^M}(1-e^{-\gamma t^M})(1-\alpha) \tag{11}$$

4. Welfare Analysis

4.1. Social Optimum

The social planner solves:

$$t^{**} = \arg\max_t \; e^{-rt}\left[\frac{v(q(t))}{2} - \lambda h(q(t))\right] \tag{12}$$

The first-order condition is:

$$\frac{v'(q(t^{**}))q'(t^{**})/2 - \lambda h'(q(t^{**}))q'(t^{**})}{v(q(t^{**}))/2 - \lambda h(q(t^{**}))} = r \tag{13}$$

Proposition 3 (Socially Premature Release)**.** *In equilibrium, release occurs before the social optimum: $t^* < t^{**}$. The welfare gap $t^{**} - t^*$ is:*

(i) Increasing in α (first-mover advantage),
(ii) Increasing in λ (social cost of safety failures),
(iii) Decreasing in r (discount rate).

Proof. The social planner's timing accounts for the safety externality ($-\lambda h'(q)q'(t) > 0$ pushes toward later release) and uses equal market shares (eliminating preemption incentives). Both forces imply $t^{**} > t^*$. The comparative statics follow from the implicit function theorem applied to the equilibrium and social planner FOCs. See Appendix A. □

4.2. Decomposing the Welfare Loss

The welfare loss from equilibrium play relative to the social optimum has two components:

$$\underbrace{W(t^{**}) - W(t^*)}_{\text{Total welfare loss}} = \underbrace{[W(t^{**}) - W(t^M)]}_{\text{Safety externality}} + \underbrace{[W(t^M) - W(t^*)]}_{\text{Preemption distortion}} \tag{14}$$

The first term captures the loss from ignoring safety (present even for a monopolist). The second captures the additional loss from competitive preemption. Both are strictly positive under our assumptions.

5. Policy Interventions

We now analyze three regulatory instruments.

5.1. Minimum Quality Standards

Suppose the regulator imposes a minimum quality standard $\bar{q}$, requiring $q(t_i) \geq \bar{q}$ before release. This is equivalent to a minimum development time $\bar{t} = q^{-1}(\bar{q})$.

Proposition 4 (Optimal Minimum Standard)**.** *There exists a minimum standard $\bar{q}^* = q(t^{**})$ that implements the social optimum as the unique equilibrium. Moreover:*

(i) Any $\bar{q} \in (q(t^*), q(t^{**}))$ strictly improves welfare over laissez-faire.
(ii) Standards set above $q(t^{**})$ reduce welfare (excessive delay).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

(iii) The optimal standard $\bar{q}^*$ is increasing in λ and α.

Proof. With binding standard $\bar{q} \geq q(t^*)$, the earliest feasible release time is $\bar{t} = q^{-1}(\bar{q}) \geq t^*$. If $\bar{t} \geq t^M$, preemption incentives vanish and both firms release at $\bar{t}$. If $t^* < \bar{t} < t^M$, the preemption game restarts at $\bar{t}$ but with reduced scope. Setting $\bar{q} = q(t^{**})$ eliminates the gap entirely. □

Policy implication. This result supports mandatory safety evaluations as a welfare-improving instrument, provided the standard is calibrated correctly.

Real-world analogue: the EU AI Act. The EU AI Act (effective August 2024, phased implementation through 2027) is the world's most comprehensive minimum quality standard regime for AI. It requires conformity assessments for high-risk AI systems before market placement, and imposes additional obligations on general-purpose AI models with systemic risk (those trained above 10^{25} FLOP), including adversarial evaluations and incident reporting. This maps directly to a binding quality standard $\bar{q}$ that prevents release until the standard is met. Two practical challenges illuminate the limits of Proposition 4: first, the EU has struggled to specify harmonized technical standards (the CEN/CENELEC process has faced repeated delays), and a standard that cannot be precisely defined may function as a *de facto* delay rather than a quality gate. Second, the standard applies only within the EU—firms may release earlier in unregulated jurisdictions, generating regulatory arbitrage. California's SB 1047, which would have imposed safety requirements on frontier AI developers, was vetoed in September 2024; the replacement SB 53 (signed September 2025) requires only transparency disclosures, not quality testing. This trajectory illustrates that optimal standards face political economy constraints not captured in our model.

5.2. Mandatory Release Delays

Now suppose the regulator mandates that a firm must wait $\Delta > 0$ periods after *announcing* completion before releasing. The intended effect is to slow the race. We show the opposite occurs.

Assumption 5 (Announcement Game with Quality Lock-In)**.** *Under mandatory delay Δ, the game has two stages:*

(i) Firm i chooses announcement time $a_i \geq 0$ (publicly declaring model ready).
(ii) Upon announcement, the firm enters a regulatory review period. Development ceases and model quality is locked *at $q(a_i)$.*
(iii) Release occurs at $t_i = a_i + \Delta$ with quality $q(a_i)$.
(iv) Market share priority is determined by release time (equivalently, announcement time, since Δ is common).

The quality lock-in assumption captures a realistic feature of regulatory review: once a firm submits a model for evaluation, the evaluated artifact is frozen. The firm cannot continue improving the model during the review period. This is analogous to pharmaceutical regulation, where the drug submitted for FDA review cannot be reformulated during the approval process.

Proposition 5 (Mandatory Delays Backfire)**.** *Under mandatory delay Δ with quality lock-in:*

(i) The equilibrium announcement time is $a^*(\Delta) = \max\{0,\, t^* - \Delta\}$.
(ii) Actual release time is $\max\{t^*, \Delta\}$, but deployed quality is $q(\max\{0,\, t^* - \Delta\})$, which is strictly lower *than laissez-faire quality $q(t^*)$ for all $\Delta \in (0, t^*)$.*
(iii) For $\Delta < t^*$: release timing is unchanged at t^*, but quality drops *from $q(t^*)$ to $q(t^* - \Delta)$. Welfare is strictly lower than laissez-faire.*
(iv) For $\Delta \geq t^*$: firms announce immediately ($a^* = 0$), locking in minimal quality $q(0) = q_0$. Release occurs at Δ, but with the worst possible model.
(v) Welfare under mandatory delay is strictly lower than under laissez-faire for all $\Delta > 0$, and strictly lower than under the optimal minimum standard for all Δ.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

Proof. The firm's problem under mandatory delay is:

$$\max_{a_i} \; e^{-r(a_i+\Delta)} \cdot v\big(q(a_i)\big) \cdot s_i(a_i, a_j)$$

where quality is $q(a_i)$ (locked at announcement) but discounting and market timing use the release date $a_i + \Delta$.

The preemption logic applies to announcements: each firm wants to announce just before its rival. The announcement preemption time satisfies $a^* + \Delta = t^*$, i.e., $a^* = t^* - \Delta$ when $\Delta < t^*$.

Case 1 ($\Delta < t^*$): Announcement at $a^* = t^* - \Delta$, release at t^*. Release timing is unchanged, but quality drops from $q(t^*)$ to $q(t^* - \Delta) < q(t^*)$. Social welfare is:

$$W(\Delta) = 2e^{-rt^*}\left[\frac{v(q(t^* - \Delta))}{2} - \lambda h(q(t^* - \Delta))\right]$$

Since $q(t^* - \Delta) < q(t^*)$, both $v(\cdot)$ falls and $h(\cdot)$ rises, so $W(\Delta) < W(0)$: welfare is strictly below laissez-faire.

Case 2 ($\Delta \geq t^*$): Firms announce at $a^* = 0$, locking in quality $q(0) = q_0$. Release occurs at Δ. Welfare is:

$$W(\Delta) = 2e^{-r\Delta}\left[\frac{v(q_0)}{2} - \lambda h(q_0)\right]$$

which is negative for typical parameters (high harm from minimal quality) and unambiguously worse than both laissez-faire and minimum standards. □

Key insight. Mandatory delays regulate the wrong margin. They constrain the time between announcement and release, but the strategic variable is the *announcement* itself. Firms shift their competitive behavior upstream to the announcement stage. Critically, because development ceases upon announcement (quality lock-in), earlier announcement directly degrades the deployed model. The delay intended to buy time for safety instead *compresses* the development window, producing lower-quality, less safe systems. This mechanism is robust: any $\Delta > 0$ strictly reduces welfare relative to laissez-faire. By contrast, minimum quality standards directly constrain the relevant variable (model quality at release) and therefore cannot be gamed in this way.

Remark 1 (When Could Delays Work?)**.** *The delay backfire result depends critically on quality lock-in (Assumption 5): development ceases upon announcement. If instead firms could continue improving models during the waiting period (no lock-in), the delay would not compress the development window. Formally, without lock-in, a delay Δ shifts release from t^* to $t^* + \Delta$ with quality $q(t^* + \Delta) > q(t^*)$, unambiguously improving quality.*

Three conditions must hold simultaneously for delays to backfire: (i) quality lock-in during the waiting period, as in regulatory review processes that freeze the evaluated artifact (analogous to FDA drug review, where reformulation during review is prohibited; see Peltzman 1973); (ii) intense competition, so preemption incentives are strong enough to drive announcement forward; and (iii) first-mover benefits accrue at announcement rather than release, so firms cannot simply announce without consequence. When any of these conditions fails—for instance, if the regulator allows continued development during review, or if competition is weak (α close to $1/2$)—delays may improve or have no effect on welfare. In addition, if the delay serves a screening *function that differentially filters high-quality from low-quality submissions (cf. Weitzman 1974), the informational benefit of delay could outweigh the quality lock-in cost. Our result is therefore best understood as identifying a specific but empirically relevant failure mode, rather than as a universal indictment of waiting periods.*

Real-world analogue: China's algorithm filing system. China's Cyberspace Administration (CAC) operates the only functioning mandatory pre-release approval regime for AI models: since August 2023, generative AI services must pass algorithm filing and security assessment before deployment. By October 2025, thousands of algorithm filings had been approved. The Chinese system

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

embodies precisely the mandatory delay mechanism analyzed here—firms must submit models for review, during which the evaluated artifact is frozen. Whether the resulting quality lock-in leads to the welfare losses predicted by Proposition 5 is an empirical question, but the mechanism design concern applies in principle. Notably, the Biden administration's Executive Order 14110 (October 2023) imposed compute-threshold reporting requirements for frontier models—a partial delay mechanism—before being revoked on January 20, 2025 by the incoming administration, illustrating the political fragility of delay-based instruments.

5.3. Voluntary Safety Commitments

Finally, we consider whether firms can sustain cooperative delay through voluntary commitments, as exemplified by Anthropic's Responsible Scaling Policy (RSP) or industry pledges at the 2023 AI Safety Summit.

Definition 2 (Safety Commitment). *A safety commitment is a publicly announced minimum release time $\hat{t}_i \geq 0$. Commitment is* credible *if deviating (releasing before $\hat{t}_i$) is observable and triggers a punishment.*

Proposition 6 (Self-Regulation as Coordination). *Consider a two-stage game: (1) firms simultaneously announce commitments $\hat{t}_i$; (2) firms play the release timing game subject to $t_i \geq \hat{t}_i$.*

(i) If commitments are not credible ($\hat{t}_i$ is cheap talk), the unique equilibrium is t^* (commitments are ignored).

(ii) If commitments are credible, there exists a subgame-perfect equilibrium with $\hat{t}_1 = \hat{t}_2 = t^C$ where $t^* < t^C \leq t^M$, improving welfare over laissez-faire.

(iii) The cooperative release time t^C can approach t^{**} if firms partially internalize safety costs (e.g., through reputational concerns or liability).

(iv) Cooperation is easier to sustain when α is smaller (less first-mover advantage) and when firms are more patient (lower r).

Proof. Part (i) follows because cheap talk does not change the payoff structure. For part (ii), with credible commitments to $\hat{t} \geq t^M$, both firms release at $\hat{t}$ and earn $e^{-r\hat{t}}v(q(\hat{t}))/2$. This exceeds the preemption equilibrium payoff $e^{-rt^*}v(q(t^*))/2$ (since at t^* both firms split equally) when $\hat{t}$ is not too far above t^M. The commitment eliminates the preemption incentive because deviating to release early is no longer feasible. See Appendix A for the full construction. □

5.4. Pigouvian Safety Tax

The textbook solution to a negative externality is a Pigouvian tax. Suppose the regulator imposes a per-release tax $\tau(q)$ that is decreasing in quality: deploying a lower-quality model incurs a higher tax.

Proposition 7 (Optimal Pigouvian Tax). *Consider a safety tax $\tau(q) = \lambda h(q)$ levied at release:*

(i) Each firm's problem becomes

$$\max_{t_i} \; e^{-rt_i}[v(q(t_i)) - \lambda h(q(t_i))]s_i(t_i, t_j)$$

which internalizes the safety externality.

(ii) The tax shifts the monopolist's optimal release time from t^M to $t^M_\tau > t^M$, where t^M_τ solves

$$\frac{[v'(q) - \lambda h'(q)]q'(t)}{v(q) - \lambda h(q)} = r$$

(iii) In the duopoly, the preemption time shifts to $t^*_\tau > t^*$, but $t^*_\tau < t^{**}$ unless the tax also addresses the preemption distortion. Specifically,

$$t^* < t^*_\tau < t^{**} = t^M_\tau$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

(iv) The Pigouvian tax partially corrects the externality but cannot fully implement the social optimum because it does not eliminate the preemption incentive. A combination of the Pigouvian tax (addressing safety) and a minimum standard (addressing preemption) can achieve the first-best.

Proof. Parts (i)–(ii): With the tax, the net value function becomes $\hat{v}(q) = v(q) - \lambda h(q)$, which satisfies $\hat{v}'(q) > 0$, $\hat{v}''(q) \leq 0$ under our assumptions (since $v'' \leq 0$ and $h'' > 0$). The monopolist's timing condition now accounts for the marginal safety benefit of delay ($-\lambda h'(q)q'(t) > 0$), pushing release later.

Part (iii): The preemption game with the tax is equivalent to the baseline game with $\hat{v}$ replacing v. Since $\hat{v}(q)/v(q)$ is increasing in q (the tax penalizes low quality disproportionately), the leader payoff $\hat{L}(t) = e^{-rt}\hat{v}(q(t))\alpha$ peaks later, and the preemption crossing occurs at $t_\tau^* > t^*$. However, the preemption distortion ($\alpha > 1/2$) remains, so $t_\tau^* < t_\tau^M = t^{**}$.

Part (iv): The remaining gap $t^{**} - t_\tau^*$ is due purely to preemption. A minimum standard $\bar{q} \geq q(t^{**})$ closes this gap. The two instruments are complementary: the tax handles the externality, the standard handles preemption. □

Quantitative illustration. Under our baseline parameterization ($\alpha = 0.7$, $\gamma = 1.0$, $r = 0.1$, $\lambda = 1.0$), the Pigouvian tax shifts the equilibrium release time from $t^* = 0.38$ to $t_\tau^* = 0.78$, closing 54% of the welfare gap between laissez-faire and the social optimum. The remaining gap reflects the preemption distortion that the tax cannot address. See Figure 1.

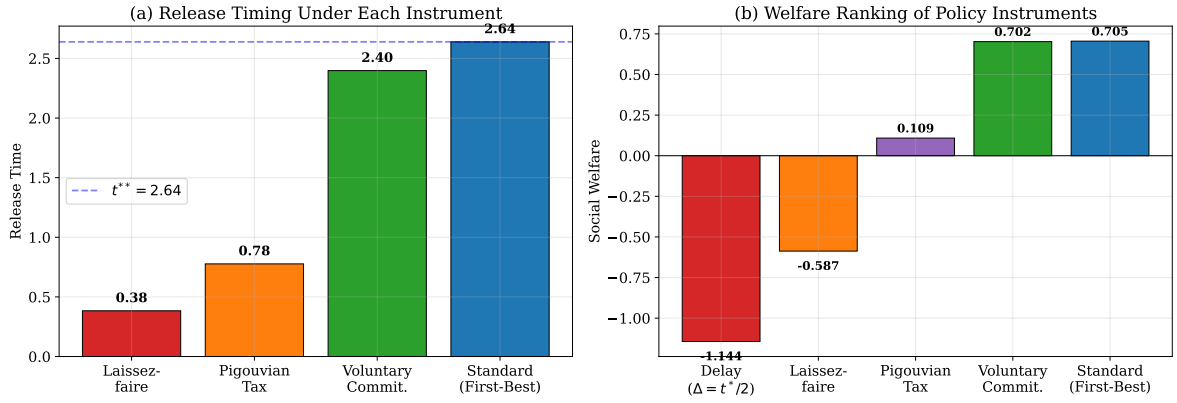

Figure 1. Pigouvian Tax Analysis. (a) Release timing under each policy instrument. (b) Welfare ranking: standards achieve the first-best; the Pigouvian tax closes 54% of the gap; mandatory delays make welfare worse than laissez-faire.

Policy synthesis. Table 1 summarizes the four instruments.

Table 1. Comparison of Policy Instruments.

Instrument	Addresses	Implements First-Best?	Welfare vs. Laissez-faire
Minimum standard $\bar{q}^*$	Preemption + safety	Yes	+
Pigouvian tax $\tau(q)$	Safety only	No	+
Voluntary commitment	Preemption (if credible)	Partially	+
Mandatory delay Δ	Neither (wrong margin)	No	−

6. Extensions

6.1. Asymmetric Firms

We relax the symmetry assumption by allowing firms to differ in development speed: firm i has quality $q_i(t) = q(\gamma_i t)$ where $\gamma_1 > \gamma_2$ (firm 1 develops faster).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

Proposition 8 (Asymmetric Equilibrium). *When $\gamma_1 > \gamma_2$:*

(i) The faster firm releases first in the unique equilibrium. Firm 1 releases at t_1^* defined by

$$e^{-rt_1^*} v(q(\gamma_1 t_1^*))\alpha = \max_{\tilde{t} > t_1^*} e^{-r\tilde{t}} v(q(\gamma_2 \tilde{t}))(1-\alpha)$$

and firm 2 follows at its monopolist-optimal time $t_2^* = t_2^M > t_1^*$.

(ii) The welfare loss is larger than in the symmetric case with $\gamma = \gamma_1$, because firm 2 releases with quality $q(\gamma_2 t_2^M) < q(\gamma_1 t_2^M)$.

(iii) There exists a threshold $\underline{\gamma}$ such that if $\gamma_2 < \underline{\gamma}$, firm 2's optimal follower profit is below its outside option, and it exits.

Proof. Define $L_i(t) = e^{-rt} v(q(\gamma_i t))\alpha$ and $F_i(t) = \max_{\tilde{t} > t} e^{-r\tilde{t}} v(q(\gamma_i \tilde{t}))(1-\alpha)$. Since $\gamma_1 > \gamma_2$, firm 1's leader payoff rises faster and peaks earlier than firm 2's. Firm 1's preemption time $t_1^* < t_2^*$, so firm 1 leads. The follower (firm 2) then faces no preemption threat and optimizes over its own timing. Part (iii): the follower payoff $F_2^* = e^{-rt_2^M} v(q(\gamma_2 t_2^M))(1-\alpha)$ is continuous and decreasing in γ_1/γ_2; when this ratio is large enough, F_2^* falls below any fixed outside option. □

Interpretation. Asymmetry intensifies the race for the leader but paradoxically *relaxes* pressure on the follower. The faster firm races harder (lower t_1^*) because it knows the follower cannot credibly preempt. This matches the industry pattern where a clear frontrunner (OpenAI in 2023) faces more release pressure than followers who can take a "fast follower" approach.

6.2. N Firms

With $N > 2$ firms competing, the first releaser captures share α, while each of the $N-1$ followers receives $(1-\alpha)/(N-1)$.

Proposition 9 (More Firms, Earlier Release). *As N increases:*

(i) Equilibrium release time $t^*(N)$ is strictly decreasing in N.

(ii) The welfare gap $t^{**} - t^*(N)$ is strictly increasing in N.

(iii) The marginal acceleration diminishes: $t^*(N) - t^*(N+1)$ is decreasing in N.

Proof. The preemption condition equates the leader payoff $L(t) = e^{-rt} v(q(t))\alpha$ to the follower payoff $F_N(t) = \max_{\tilde{t} > t} e^{-r\tilde{t}} v(q(\tilde{t})) \cdot (1-\alpha)/(N-1)$. Increasing N reduces $F_N(t)$ for every t without affecting $L(t)$. Since L crosses F_N from below, and F_N shifts down, the crossing point $t^*(N)$ moves left. Part (ii) follows directly. For part (iii), $F_N(t) = F_2(t)/(N-1)$, so the shift from N to $N+1$ is $F_2(t)/(N-1) - F_2(t)/N = F_2(t)/[N(N-1)]$, which is decreasing in N, implying diminishing marginal acceleration. □

This result generates a striking policy implication:

Corollary 10 (The Concentration Paradox). *Market concentration in AI development—holding other factors constant—reduces the pressure for premature release and improves safety outcomes. Specifically, total welfare under a duopoly ($N = 2$) exceeds welfare under an oligopoly ($N > 2$) because the duopoly releases later with higher quality.*

This creates a genuine policy tension: antitrust authorities may seek to promote competition in AI markets on traditional efficiency grounds, but doing so exacerbates the safety externality from premature release. The "right" number of AI firms balances static efficiency gains from competition against the dynamic welfare losses from accelerated release. Proposition 9 shows that the safety cost of competition can be substantial, particularly when α is large (strong lock-in) and λ is large (high safety stakes).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

6.3. Open-Source as Strategic Release

We extend the model by allowing firms to choose release *mode* in addition to timing. At release time t_i, firm i chooses $m_i \in \{C, O\}$ (closed or open). Closed release yields the standard payoff. Open release yields zero direct revenue but *commoditizes* the market: it reduces the first-mover advantage from α to $\hat{\alpha} < \alpha$ for all subsequent releases, as open-source alternatives reduce switching costs.

Proposition 11 (Strategic Open-Sourcing)**.** *Consider an asymmetric duopoly with $\gamma_1 > \gamma_2$ (firm 1 faster):*

(i) If both firms are closed, the equilibrium is as in Proposition 8: firm 1 leads at t_1^*.

(ii) The slower firm (firm 2) prefers open-source release if the commoditization effect is sufficiently strong:

$$\underbrace{e^{-rt_2^O} v(q(\gamma_2 t_2^O)) \cdot 0}_{\text{OS revenue}=0} + \underbrace{\delta_2(\hat{\alpha})}_{\text{future value under commoditization}} > \underbrace{e^{-rt_2^M} v(q(\gamma_2 t_2^M))(1-\alpha)}_{\text{closed follower payoff}}$$

where $\delta_2(\hat{\alpha})$ captures firm 2's continuation value in a commoditized market (e.g., via complementary services, data advantages, or future product differentiation).

(iii) Open-sourcing by the slower firm accelerates *the leader's release: $t_1^*(O) < t_1^*(C)$, because the leader anticipates commoditization and rushes to capture rents before the market is disrupted.*

(iv) The welfare effect of open-sourcing is ambiguous: it reduces first-mover rents (good) but accelerates the race (bad).

Proof. Part (ii): Firm 2's closed follower payoff is bounded by $(1-\alpha) \cdot e^{-rt_2^M} v(q(\gamma_2 t_2^M))$. If α is large (winner-take-most), this is small. Open-sourcing yields $\delta_2(\hat{\alpha})$ which can exceed the closed payoff when commoditization reduces α substantially and firm 2 has complementary assets. Part (iii): The threat of commoditization effectively reduces the leader's *future* payoff from delay, making the leader's value function steeper and the preemption time earlier. Part (iv): The welfare effect depends on whether the quality loss from accelerated release outweighs the gain from reduced monopoly distortion. □

Application. This rationalizes Meta's strategy of open-sourcing Llama while trailing OpenAI and Google DeepMind in capability. Meta's complementary assets (social media data, advertising revenue, device ecosystem) give it a large δ_2 even with zero direct model revenue. Meanwhile, open-sourcing pressures closed competitors to release faster—consistent with the observed acceleration of release cadence after Llama's release.

6.4. Staged Release and Information Revelation

Firms may release in stages: research paper, limited API, broad deployment. We model this as a choice of release *breadth* $b \in [0, 1]$, where $b = 0$ is a paper-only release and $b = 1$ is full deployment. A release of breadth b at time t yields:

- Market share: $s_i(t, b) = b \cdot s_i(t)$ (proportional to breadth).
- Safety information: reveals a signal $\sigma \sim G(q, b)$ about true risk, with informativeness increasing in b.
- Harm: $h(q, b) = b \cdot h(q)$ (harm proportional to deployment scale).

Proposition 12 (Staged Release as Screening)**.** *Consider a two-period extension where a firm can release at breadth b_1 in period 1 and update to $b_2 \leq 1$ in period 2 after observing the safety signal σ:*

(i) Safety screening*: The optimal staged release sets $b_1^* < 1$, deploying narrowly first to learn about risks. The value of staging is increasing in the variance of the harm distribution and in λ (safety weight).*

(ii) Strategic signaling*: A firm with quality $q > q^*$ (above a threshold) releases a research paper ($b_1 \approx 0$) as a credible signal, because only high-quality firms can afford to reveal capabilities without capturing market share immediately.*

(iii) Competition erodes staging*: In equilibrium, the number of stages and the duration of limited deployment decrease with the first-mover advantage α, because the opportunity cost of delayed full deployment rises.*

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

Proof sketch. Part (i): With staged release, expected harm is $b_1 h(q) + (1-b_1)\mathbb{E}[\max\{0, b_2 h(q) \mid \sigma\}]$. The option to withdraw after observing σ is valuable, and this value increases in the variance of h and in λ. Part (ii): Releasing a paper reveals q to the market, increasing demand for the eventual product. Low-quality firms prefer not to reveal, creating a separating equilibrium above threshold q^*. Part (iii): The cost of staging is the forgone first-mover advantage $(1 - b_1)\alpha e^{-rt} v(q(t))$, which increases in α. □

Application. The model explains why OpenAI published GPT-4's system card before broad deployment, why Anthropic releases Claude models with initial rate limits that gradually expand, and why Google's Gemini had a staged rollout from API to consumer product.

6.5. Stochastic Quality

Our baseline assumes deterministic quality improvement. We now sketch an extension with stochastic breakthroughs that strengthens the delay backfire result.

Suppose quality evolves as $dq = \mu(q)\,dt + \sigma(q)\,dB_t$, where B_t is a standard Brownian motion. Quality is no longer deterministic—there is genuine uncertainty about when a model will be "ready."

Proposition 13 (Stochastic Quality Strengthens Delay Backfire)**.** *Under stochastic quality:*

(i) Firms possess an option value *of continued development: the option to delay release if a bad quality shock occurs. The option value is increasing in σ (quality uncertainty).*

(ii) Mandatory delays with quality lock-in destroy *this option value, because firms must commit at announcement time (before the delay period) and cannot respond to subsequent quality realizations.*

(iii) The welfare loss from mandatory delays is strictly larger under stochastic quality than under deterministic quality: $W^{delay}_{stoch} < W^{delay}_{det}$ for all $\Delta > 0$, $\sigma > 0$.

(iv) Minimum quality standards preserve *the option value, since firms can continue developing until the standard is met, regardless of the quality path.*

Intuition. Under uncertainty, flexibility has value. Minimum standards are compatible with flexibility (keep developing until ready); mandatory delays with quality lock-in are not (commit before the waiting period, losing the ability to respond to new information). This deepens the distinction between the two instruments beyond the deterministic case.

7. Empirical Implications

Our model generates several testable predictions:

1. **Clustering of releases.** Equilibrium involves simultaneous release. Empirically, we should observe clustering of major AI releases within narrow time windows. Early evidence includes GPT-4 (March 2023), Gemini (December 2023), and Claude 3 (March 2024). The most striking confirmation came in November 2025, when four frontier models were released within six days: xAI's Grok 4.1 (November 17), Google's Gemini 3 Pro (November 18), OpenAI's GPT-5.1 (November 12–24), and Anthropic's Claude Opus 4.5 (November 24). This unprecedented clustering is precisely the equilibrium prediction: once one firm's release becomes imminent, all rivals release near-simultaneously at t^*.
2. **Release gaps shrink with competition.** As more firms reach the capability frontier, the time between successive releases should decrease. The gap between GPT-3 (June 2020) and GPT-3.5 (November 2022) was 29 months; between GPT-3.5 and GPT-4, 4 months. By 2025, internal reports confirmed the mechanism: OpenAI CEO Sam Altman declared "Code Red" in December 2025 after Google's Gemini 3 surpassed ChatGPT on major benchmarks, redirecting resources to accelerate the next release. Industry sources described models being "released to the public without a lot of holding time" due to competitive pressure, with employees requesting delays overruled by strategic urgency.
3. **First-mover advantage predicts timing.** Markets with stronger lock-in effects (enterprise API adoption) should see earlier release than markets with weaker lock-in (consumer chatbots).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

4. **Voluntary commitments correlate with patience—but unravel under pressure.** Firms with longer time horizons (e.g., well-capitalized labs) should be more likely to adopt voluntary safety commitments. Consistent with Proposition 6(iv), commitments should be harder to sustain when α is large. Anthropic's February 2026 revision of its Responsible Scaling Policy—dropping the commitment to pause scaling when safety lagged—is a direct empirical confirmation: even the most safety-committed firm concluded that unilateral restraint was untenable under intensifying competition.
5. **Capability-gap releases.** A firm that achieves a significant capability jump should delay release *more* (to maximize the quality advantage), while a firm with a marginal improvement should release *faster* (before the gap closes).
6. **Concentration and safety.** Periods or segments with fewer frontier competitors should exhibit longer development cycles and higher model quality at release. The entry of new frontier labs (e.g., xAI in 2023, DeepSeek in 2025) should be followed by accelerated release schedules across all incumbents. DeepSeek's January 2025 release of R1—matching OpenAI's o1 on key benchmarks at a fraction of the cost—triggered a market-wide acceleration, with NVIDIA shares falling 20% and rival labs scrambling to respond. By mid-2025, Epoch AI counted over 30 models trained above 10^{25} FLOP from 12 developers, consistent with the prediction that $t^*(N)$ is decreasing in N.
7. **Regulatory arbitrage in delays.** If jurisdictions impose mandatory pre-release waiting periods, firms headquartered there should announce *earlier* (not later), and the quality of models released under delay mandates should be measurably lower than comparable models released without such mandates.

8. Conclusion

We have presented a formal model of AI release timing as a preemption game with safety externalities. Our central results are: (1) competition leads to socially premature release; (2) minimum quality standards can implement the first-best, but mandatory delays strictly worsen outcomes due to quality lock-in at the announcement stage; (3) Pigouvian safety taxes partially correct the externality but cannot eliminate preemption; (4) voluntary safety commitments can sustain cooperative outcomes when credible; and (5) more competition exacerbates the race, creating a tension between antitrust goals and safety objectives.

The broader message is that the "AI race" is not merely a metaphor—it is a well-defined strategic interaction with predictable welfare consequences. Designing effective AI governance requires understanding these strategic incentives, not merely cataloging potential harms. In particular, the quality lock-in result suggests that well-intentioned regulatory delays can be counterproductive: if development ceases during mandatory review periods, firms will rush to announce earlier, deploying less developed systems. The policy implication is clear: regulate quality directly, not timing. Pigouvian taxes can complement standards by making firms internalize safety costs, but the preemption distortion requires quantity-based instruments (standards) rather than price-based instruments (taxes) alone—an AI-specific echo of the Weitzman (1974) prices-vs-quantities insight.

Several directions merit further investigation. First, a full dynamic model with stochastic breakthroughs would capture the option-value aspects of AI development more richly. Second, incorporating asymmetric information about capabilities would formalize the signaling aspects of staged releases. Third, an empirical analysis using the growing database of AI model releases could test our predictions and calibrate the model. Fourth, the concentration paradox deserves deeper investigation: characterizing the socially optimal market structure for AI development when safety externalities interact with scale economies and innovation incentives is a first-order policy question.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

Appendix A. Proofs

Appendix A.1. Proof of Proposition 2

We adapt the argument in Fudenberg and Tirole (1985). Define:

$$L(t) = e^{-rt} v(q(t))\alpha \tag{A1}$$

$$F(t) = \max_{\tilde{t} \geq t} e^{-r\tilde{t}} v(q(\tilde{t}))(1-\alpha) \tag{A2}$$

$L(t)$ is the payoff from leading at t; $F(t)$ is the follower's optimized payoff given the leader released at t.

Properties of $L(t)$:

- $L(0) = v(q_0)\alpha > 0$
- $L(t)$ is single-peaked: $L'(t) = e^{-rt}[v'(q(t))q'(t) - rv(q(t))]\alpha$, which is positive for small t and negative for large t.
- $L(t) \to 0$ as $t \to \infty$.

Properties of $F(t)$:

- $F(t)$ is weakly decreasing in t (earlier leader release gives the follower more time to optimize).
- For t small, $F(t) = e^{-rt^M} v(q(t^M))(1-\alpha)$ (follower optimally waits until t^M).
- For $t > t^M$, $F(t) = e^{-rt} v(q(t))(1-\alpha)$ (follower releases immediately).

At $t = 0$: $L(0) = v(q_0)\alpha$ versus $F(0) = e^{-rt^M} v(q(t^M))(1-\alpha)$. For α not too close to 1 and q_0 small, $L(0) < F(0)$: it is better to follow than to lead at $t = 0$.

At $t = t^M$: $L(t^M) = e^{-rt^M} v(q(t^M))\alpha > e^{-rt^M} v(q(t^M))(1-\alpha) = F(t^M)$.

By continuity, there exists $t^* \in (0, t^M)$ where $L(t^*) = F(t^*)$. This is the preemption time. At t^*, each firm is indifferent between leading and following. For $t < t^*$, following is preferred; for $t > t^*$, leading is preferred. The unique equilibrium has both firms releasing at t^* (with mixing to handle the simultaneous case).

The comparative static $\partial t^*/\partial \alpha < 0$ follows because increasing α raises $L(t)$ and lowers $F(t)$, shifting the crossing point to the left.

Appendix A.2. Proof of Proposition 3

The social planner's FOC (13) implies that t^{**} equates the rate of return to waiting (quality improvement net of safety improvement) to the discount rate. Since the planner:

(a) Faces no preemption incentive ($s_i = 1/2$ always),

(b) Internalizes safety ($\lambda h(q(t))$ term),

t^{**} is determined by a strictly higher marginal benefit of waiting than is faced by either firm in equilibrium. Hence $t^{**} > t^*$.

For the comparative statics: (i) Higher α decreases t^* (Proposition 2) without affecting t^{**}. (ii) Higher λ increases t^{**} without affecting t^*. (iii) Higher r decreases both t^* and t^{**}, but decreases t^{**} more because the safety externality term is also discounted.

Appendix A.3. Proof of Proposition 5

Under mandatory delay Δ with quality lock-in (Assumption 5), the firm's problem is:

$$\max_{a_i} \; e^{-r(a_i+\Delta)} v\big(q(a_i)\big) s_i(a_i, a_j)$$

Note the crucial distinction from the baseline: the value function evaluates quality at a_i (announcement time, when development ceases), while discounting uses $a_i + \Delta$ (release time).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

The preemption structure in announcements is identical to the baseline game in release times. Define the announcement leader payoff:

$$\hat{L}(a) = e^{-r(a+\Delta)} v(q(a))\alpha$$

and announcement follower payoff (follower optimally announces at $a^F = t^M - \Delta$ if leader announced at $a < a^F$):

$$\hat{F}(a) = \max_{\tilde{a}>a} e^{-r(\tilde{a}+\Delta)} v(q(\tilde{a}))(1-\alpha) = e^{-r\Delta} F_0(a)$$

where $F_0(a)$ is the follower payoff from the baseline game. Similarly, $\hat{L}(a) = e^{-r\Delta} e^{-ra} v(q(a))\alpha$.

The preemption announcement time a^* satisfies $\hat{L}(a^*) = \hat{F}(a^*)$, which (dividing by $e^{-r\Delta}$) reduces to the baseline preemption condition at a^*. Hence $a^* = t^* - \Delta$ when $\Delta < t^*$, and $a^* = 0$ when $\Delta \geq t^*$.

Case 1 ($\Delta < t^*$): Release at $a^* + \Delta = t^*$ (unchanged), but quality is $q(t^* - \Delta) < q(t^*)$. This *strictly reduces* both consumer value and safety relative to laissez-faire, since $v'(q) > 0$ and $h'(q) < 0$.

Case 2 ($\Delta \geq t^*$): Announcement at $a^* = 0$, quality $q(0) = q_0$ (minimum), release at Δ. Welfare is $2e^{-r\Delta}[v(q_0)/2 - \lambda h(q_0)]$, which is strictly negative for typical parameters and strictly below both laissez-faire and any minimum standard.

The strict welfare dominance of minimum standards follows because standards constrain quality directly ($q(t_i) \geq \bar{q}$ at release), ensuring development continues until the standard is met. Delays create a wedge between the quality determination date (announcement) and the release date, allowing preemption to erode quality while release timing appears unchanged.

Appendix A.4. Proof of Proposition 6

Part (i): If commitments are cheap talk (not enforceable), the release subgame is identical to the baseline. Each firm announces $\hat{t}_i = t^{**}$ but deviates to t^* in the release stage, since $L(t^*) = F(t^*)$ and deviating yields the preemption payoff. Cheap talk does not alter the incentive-compatibility constraints of the release game.

Part (ii): With credible commitments, the binding constraint is $t_i \geq \hat{t}_i$. Consider symmetric commitments $\hat{t}_1 = \hat{t}_2 = t^C$. If $t^C \geq t^M$, both firms release at t^C and earn $\pi^C = e^{-rt^C} v(q(t^C))/2$. This exceeds the preemption payoff $\pi^* = e^{-rt^*} v(q(t^*))/2$ when

$$e^{-r(t^C - t^*)} \frac{v(q(t^C))}{v(q(t^*))} > 1$$

which holds for t^C near t^M because t^M maximizes the monopolist's discounted value. The commitment eliminates preemption by making early release infeasible, converting the preemption game into a coordination game with a Pareto-superior equilibrium at t^C.

Part (iii): If firms internalize fraction $\phi \in [0, 1]$ of safety costs (e.g., through reputational liability), the cooperative optimum shifts toward t^{**}. As $\phi \to \lambda/(\lambda + 1)$, the cooperative t^C converges to t^{**}.

Part (iv): Sustainability requires that the one-shot deviation gain from releasing at t^* given the other firm commits to t^C does not exceed the punishment cost. The deviation gain is $e^{-rt^*} v(q(t^*))\alpha - e^{-rt^C} v(q(t^C))/2$, which increases in α (larger first-mover advantage makes deviation more tempting). A lower discount rate r increases the weight on future punishment, making cooperation easier to sustain in repeated interactions.

Appendix A.5. Proof of Proposition 7

Parts (i)–(ii): The tax $\tau(q) = \lambda h(q)$ adds a cost at release, making firm i's payoff $e^{-rt_i}[v(q(t_i)) - \lambda h(q(t_i))]s_i$. Define $\hat{v}(q) = v(q) - \lambda h(q)$. Under our assumptions ($v' > 0$, $h' < 0$), we have $\hat{v}'(q) = v'(q) - \lambda h'(q) > 0$. The monopolist maximizes $e^{-rt}\hat{v}(q(t))$, yielding the FOC $\hat{v}'(q(t))q'(t)/\hat{v}(q(t)) = r$. Since $\hat{v}'(q)/\hat{v}(q) > v'(q)/v(q)$ for $q < \bar{Q}$ (the tax penalizes low quality disproportionately), the LHS at any given t is larger under the tax, so the FOC is satisfied at a later time: $t^M_\tau > t^M$.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

Part (iii): The preemption game with the tax replaces v with $\hat{v}$. The leader payoff $\hat{L}(t) = e^{-rt}\hat{v}(q(t))\alpha$ peaks at $t_\tau^M > t^M$, while the follower payoff shifts commensurately. The preemption crossing $\hat{L}(t_\tau^*) = \hat{F}(t_\tau^*)$ occurs at $t_\tau^* > t^*$. However, the structural preemption distortion ($\alpha > 1/2$) persists: the leader still has disproportionate incentive to release early. Hence $t_\tau^* < t_\tau^M = t^{**}$.

Part (iv): The gap $t^{**} - t_\tau^*$ is due entirely to the preemption distortion, which the tax cannot address (it only adjusts the payoff level, not the share structure). A minimum standard $\bar{q} \geq q(t^{**})$ binds the release timing directly, closing the remaining gap. Together, the tax internalizes the externality and the standard eliminates preemption.

Appendix A.6. Proof of Proposition 13 (Sketch)

Under stochastic quality $dq = \mu(q)\,dt + \sigma(q)\,dB_t$, the firm's release problem becomes an optimal stopping problem:

$$V(q) = \sup_\tau \mathbb{E}\left[e^{-r\tau} v(q_\tau) s_i \mid q_0 = q\right]$$

where τ is a stopping time adapted to the filtration generated by q_t.

Part (i): By standard real options theory, the optimal stopping boundary q^*_{stop} satisfies $q^*_{\text{stop}} > q^*_{\text{det}}$ (the threshold exceeds the deterministic optimum) because waiting preserves the option to stop at a higher quality realization. The option value $V(q) - v(q)$ is positive and increasing in σ.

Part (ii): Under mandatory delay with lock-in, the firm must announce at time τ_a and deploy quality q_{τ_a} at time $\tau_a + \Delta$. The announcement decision at τ_a must be made without knowing the quality path during $[\tau_a, \tau_a + \Delta]$. Even though quality continues evolving, the deployed model is frozen at q_{τ_a}. This eliminates the option to "undo" a bad draw—the firm cannot withdraw after announcement even if quality deteriorates, and cannot benefit from improvements during the delay period.

Part (iii): The welfare gap between deterministic and stochastic cases under mandatory delays is $\mathbb{E}[\text{option value destroyed}] > 0$ for all $\sigma > 0$. Under deterministic quality, mandatory delays cause welfare loss $W_{\text{det}}^{\text{delay}} < W^{\text{LF}}$. Under stochastic quality, the additional loss from destroyed option value implies $W_{\text{stoch}}^{\text{delay}} < W_{\text{det}}^{\text{delay}}$.

Part (iv): Standards set a threshold $\bar{q}$ that the firm must exceed *at release time*. The firm's stopping rule becomes $\tau = \inf\{t : q_t \geq \bar{q}\}$, which preserves full flexibility to continue developing if quality is temporarily below the threshold due to a bad shock. The option value under standards is $V^{\text{std}}(q) \geq V^{\text{det,std}}(q)$ (weakly greater than deterministic because upside shocks are captured).

References

Dilip Abreu and Markus K. Brunnermeier. Bubbles and crashes. *Econometrica*, 71(1):173–204, 2003.

Daron Acemoglu. The simple macroeconomics of AI. *NBER Working Paper*, (32487), 2024.

Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press, 2018.

Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the precipice: A model of artificial intelligence development. *AI & Society*, 31(2):201–206, 2016.

Leopold Aschenbrenner and Philip Trammell. Existential risk and growth. Global Priorities Institute Working Paper 13-2024, 2024.

Amanda Askell, Miles Brundage, and Gillian Hadfield. The role of cooperation in responsible AI development. *arXiv preprint arXiv:1907.04534*, 2019.

Allan Dafoe. AI governance: A research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford*, 2018.

Drew Fudenberg and Jean Tirole. Preemption and rent equalization in the adoption of new technology. *Review of Economic Studies*, 52(3):383–401, 1985.

The Anh Han, Luís Moniz Pereira, Francisco C. Santos, and Tom Lenaerts. To regulate or not: A social dynamics analysis of an idealised AI race. *Journal of Artificial Intelligence Research*, 69:881–921, 2020.

The Anh Han, Luís Moniz Pereira, Tom Lenaerts, and Francisco C. Santos. Mediating artificial intelligence developments through negative and positive incentives. *PLOS ONE*, 16(1):e0244592, 2021.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2470.v1

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

Heidrun C. Hoppe and Ulrich Lehmann-Grube. Innovation timing games: A general framework with applications. *Journal of Economic Theory*, 121(1):30–50, 2005.

Charles I. Jones. The AI dilemma: Growth versus existential risk. *AER: Insights*, 6(4):575–590, 2024.

Anton Korinek. Scenarios for the transition to AGI. *NBER Working Paper*, (32255), 2024.

Travis LaCroix and Aydin Mohseni. The tragedy of the AI commons. *Synthese*, 200:289, 2022.

Glenn C. Loury. Market structure and innovation. *Quarterly Journal of Economics*, 93(3):395–410, 1979.

Wim Naudé and Nicola Dimitri. The race for an artificial general intelligence: Implications for public policy. *AI & Society*, 35:367–379, 2020.

Sam Peltzman. An evaluation of consumer protection legislation: The 1962 drug amendments. *Journal of Political Economy*, 81(5):1049–1091, 1973.

Richard A. Posner. *Catastrophe: Risk and Response*. Oxford University Press, 2004.

Jennifer F. Reinganum. On the diffusion of new technology: A game theoretic approach. *Review of Economic Studies*, 48(3):395–405, 1981.

Xavier Vives. Innovation and competitive pressure. *Journal of Industrial Economics*, 56(3):419–469, 2008.

Martin L. Weitzman. Prices vs. quantities. *Review of Economic Studies*, 41(4):477–491, 1974.

Martin L. Weitzman. On modeling and interpreting the economics of catastrophic climate change. *Review of Economics and Statistics*, 91(1):1–19, 2009.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.