Three Roads From Here: Safety, Acceleration, and Pause in the Governance of Frontier AI

Sean Sooch

June 2026

Abstract

Public debate over frontier artificial intelligence has consolidated into three broad camps: a safety-through-development position that accepts continued scaling under engineered and institutional safeguards, an accelerationist position that treats speed itself as the safety strategy, and a pause-or-stop position that holds that frontier development should halt until alignment is solved. This paper characterizes each position on its strongest terms, identifies the empirical and philosophical cruxes that actually separate them, and then examines the two risk categories that dominate the disagreement: existential risks (x-risks), in which humanity's potential is permanently destroyed, and suffering risks (s-risks), in which the future contains suffering at astronomical scale. The s-risk lens is given particular attention because it scrambles the usual ordering of the debate. A pause that straightforwardly reduces extinction risk can plausibly increase suffering risk through hardware overhang, race resumption under worse geopolitical conditions, and near-miss alignment scenarios in which a partially aligned system is worse than either a fully aligned one or none at all. The paper closes with a crux map intended to let a reader locate their own position by their answers to a small number of factual questions, rather than by tribal affiliation.

1. Introduction

Arguments about AI risk are frequently conducted as arguments about character: accelerationists are reckless, safety researchers are captured, pause advocates are doomers. This framing is convenient and almost entirely useless. Each of the three major positions is held by serious people, follows validly from its premises, and fails if its premises fail. The productive question is never "which tribe is virtuous" but "which premises are true."

This paper does three things. First, it states each position in the form its most careful adherents actually hold, including the internal variation within each camp (Sections 2 through 4). Second, it introduces the two risk categories that drive most of the disagreement, existential risk and suffering risk, and shows that they are not the same axis and do not always move together

(Sections 5 and 6). Third, it examines the pause proposal in detail, because pausing is the policy lever where the x-risk and s-risk analyses diverge most sharply, and where the three camps make their most concrete and testable claims (Section 7). Section 8 reduces the debate to its load-bearing cruxes.

A note on scope. This paper is about frontier general-purpose systems, the small number of models at the leading edge of capability. Nothing here bears on whether a hospital should deploy a sepsis-prediction model. The three positions agree far more about narrow AI than their rhetoric suggests; the fight is about the frontier.

---

2. Position I: Safety Through Development

2.1 The claim

The first position holds that frontier AI development should continue, but inside a thickening envelope of technical safeguards, evaluation regimes, and conditional commitments. Its institutional expression is the responsible scaling policy or frontier safety framework: a published set of if-then commitments under which a developer specifies capability thresholds (in domains such as autonomy, cybersecurity, and biological uplift) and binds itself to specified safeguards, or to halting, when evaluations show a threshold has been crossed. Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google DeepMind's Frontier Safety Framework are the canonical examples, and the 2024 Seoul frontier safety commitments extended the form to a broader set of developers.

The position rests on four supporting arguments.

The empirical access argument. Alignment is an empirical discipline, and you cannot study the failure modes of capable systems without capable systems. Interpretability research, scalable oversight, evaluations for deceptive behavior: all of these advanced because researchers had frontier models to study. On this view, a long pause does not buy time for safety research so much as starve it of its subject matter.

The racing argument. Frontier capability is not produced by one actor. If safety-motivated developers unilaterally stop, development continues at whatever labs and states care least about safety, and the eventual transition to advanced AI is steered by the least cautious actors. Continued participation by safety-focused organizations keeps safety-conscious people at the frontier, in the rooms where deployment decisions are made.

The gradualism argument. Capabilities have so far arrived continuously rather than discontinuously. Each model generation is deployed, its failures are observed at low stakes, and the lessons feed the next round of safeguards. If takeoff remains relatively smooth, this iterative loop (deploy, observe, correct) is the most reliable safety mechanism civilization has, because it is the mechanism by which every other dangerous technology was tamed.

The overhang argument. Pausing software progress does not pause hardware progress, algorithmic insight accumulating in the open literature, or capital formation. A pause that ends (and its proponents rarely claim it would be permanent) resumes development with years of accumulated compute and ideas, producing a faster and less controllable capability jump than the counterfactual smooth path. Slow and continuous is safer than stop-then-lurch.

2.2 Internal variation

The camp spans a wide band. At one edge sit researchers who think catastrophic risk this decade is substantial and view RSPs as the best achievable instrument rather than a sufficient one. At the other edge sit researchers who think x-risk concern is overblown and emphasize present-day harms: bias, labor displacement, concentration of power, misinformation. These groups disagree with each other about almost everything except the policy conclusion that development should continue under safeguards. Critics correctly note that this makes "safety through development" a coalition rather than a doctrine.

2.3 The strongest objections

The position's known weak points are governance-shaped rather than technical. If-then commitments are largely self-enforced, and a developer under competitive pressure grades its own evaluations. The position assumes evaluations can detect dangerous capabilities before deployment, which assumes the absence of sandbagging and deceptive alignment, which is precisely the failure mode the evaluations are supposed to catch: there is a circularity at the core of the audit regime. And the racing argument is uncomfortably self-serving; "if we don't do it, someone worse will" is the argument every participant in every race has always made, and it is structurally incapable of ever recommending that the speaker stop.

---

3. Position II: Acceleration

3.1 The claim

The second position holds that frontier development should accelerate, and its strongest argument is not the one most often quoted in public. The popular case is humanitarian. The

strongest case is game-theoretic.

The game-theoretic argument. The race is not a policy that anyone chose and that anyone can therefore unchoose; it is the structure of the situation. Many actors can build frontier systems, the rewards to the leader are enormous and concentrated, restraint by any one actor is unverifiable by the others, and there is no enforcer. That is a multipolar trap: each player must assume the others will build, so each player's best response is to build, and the result is a race that no individual participant can call off, including participants who privately wish it would stop. On this reading the pause position is not wrong about the danger; it is wrong about the move set. Unilateral restraint does not remove the danger, it only reassigns the frontier to whoever restrained itself least, so the careful actor's rational strategy is to be first, or to diffuse capability so widely that there is nothing left for a single winner to lock in. Note what this argument does not require: it does not require optimism about AI at all. It requires only the claim that the game's payoffs cannot be changed in time, which is what separates it from Position I's superficially similar racing argument (race, but inside commitments that reshape the game) and from Position III (change the game or refuse to play it).

The humanitarian argument. Roughly 150,000 people die every day, most from causes that are in principle solvable: disease, aging, poverty, scarcity. If advanced AI can compress decades of medical and economic progress into years, each year of delay has a body count, and that body count is not hypothetical in the way model-based extinction scenarios are. This is the argument accelerationists lead with in public, and it has real force, but it is the weaker of the two: it prices delay precisely and catastrophe at approximately zero, so it persuades no one who takes the catastrophe seriously. The game-theoretic argument has no such dependence, which is why it does the real work.

The camp runs on several further engines.

Techno-optimist priors. Every transformative technology (printing, electricity, vaccines, the internet) arrived over the objections of contemporaries who predicted catastrophe, and the catastrophes either did not arrive or were dwarfed by the benefits. The burden of proof, on this view, lies with those who claim this time is different.

Distrust of centralized control. Pause and licensing regimes require concentrating control over compute and models in a small number of governments and corporations. Accelerationists, particularly the open-source wing, argue that this concentration is itself the catastrophe: a world where a handful of actors control the most powerful technology in history is a world primed for permanent oligarchy, regardless of whether the AI is aligned. Widely distributed capability is the defense, on the same logic by which distributed cryptography defends privacy.

Offense-defense optimism. Misuse risks are real but defensible: AI-discovered vulnerabilities are patched by AI-powered defenders, engineered pathogens are met by AI-accelerated vaccine platforms. If defense scales with capability, the equilibrium of a high-capability world is safe, and the dangerous period is the transition, which should therefore be crossed quickly rather than slowly.

The thermodynamic and memetic wing. The e/acc subculture proper adds a quasi-metaphysical layer (intelligence as the universe's tendency toward greater free-energy capture, growth as a terminal value rather than an instrumental one). This wing is rhetorically loud but analytically thin, and it is a mistake to treat it as the camp's center of mass. The strongest accelerationist case is the game-theoretic one, not the thermodynamic one.

3.2 Internal variation

The camp divides on what, exactly, to accelerate. Market accelerationists want regulatory restraint for the existing frontier developers. Open-source accelerationists want weights released so capability diffuses. A moderate cousin, Buterin's d/acc (defensive, decentralized, differential acceleration), accepts the techno-optimist frame but argues for steering acceleration toward defensive technologies (biosecurity, cybersecurity, epistemic tools) and away from raw agentic capability. d/acc is best understood as a bridge position: accelerationist in temperament, safetyist in its choice of targets.

3.3 The strongest objections

The game-theoretic argument has a structural flaw its users rarely state: it treats the payoff matrix as a law of nature. Game theory itself says otherwise. Payoffs are changed by verification technology, by repeated play, by enforcement, and by communication, which is the entire history of arms control; "defection is inevitable" was asserted about nuclear testing, chemical weapons, and ozone-destroying refrigerants, and was wrong each time coordination infrastructure caught up to the incentive problem. The argument also quietly assumes that the defector who wins the race survives the prize. If current alignment techniques do not scale, which is precisely the matter in dispute, then the win cell of the matrix is mislabeled: the first actor to deploy a misaligned superintelligence does not defeat its rivals, it kills them and itself, and a prisoner's dilemma whose defection payoff is death is not a prisoner's dilemma. Finally, inevitability talk is performative. Every actor who declares the race unstoppable makes it marginally harder to stop, which means the declaration is less a finding than a move in the very game it claims to describe.

The remaining objections target the camp's other engines. The opportunity-cost argument assumes the benefit stream survives the transition, which assumes the transition goes well, which is the matter in dispute; the argument prices delay precisely and prices catastrophe at

approximately zero. The historical-track-record argument is an anthropic artifact: technologies that could have ended the species would not leave behind commentators to cite the track record, and the sample of "transformative technologies that created their own successor agents" has zero members. Offense-defense optimism is asserted rather than demonstrated, and in at least one domain (engineered biology) most domain experts believe offense is structurally favored. And the decentralization argument proves too much: nobody applies it to fissile material, and the question of whether frontier weights are more like printing presses or more like enriched uranium is the actual question, not an answer to it.

---

4. Position III: Pause or Stop

4.1 The claim

The third position holds that frontier development should halt, by coordinated agreement or by regulation, until the alignment problem is solved or at least until safety cases can be made rigorous. Its premise structure is short. First, building something more capable than humanity at general problem-solving, without the ability to specify what it should want, is the kind of mistake that is made exactly once. Second, no current technique provides anything close to a guarantee that a superintelligent system's objectives would be compatible with human survival; current alignment methods (RLHF and its descendants) shape surface behavior and provide no assurance about behavior out of distribution or under self-improvement. Third, when the downside is unbounded and irreversible, the only acceptable error is the recoverable one. You can restart a paused field; you cannot restart an extinct species.

The position's most prominent recent statement is Yudkowsky and Soares's *If Anyone Builds It, Everyone Dies* (2025), which argues that the danger is not in who builds superintelligence but in the building itself, and that the appropriate response is an enforced international halt on frontier training runs, with compute monitoring as the verification mechanism, on the model of nuclear and biological arms control. The milder ancestral form was the 2023 Future of Life Institute open letter calling for a six-month pause on training systems beyond GPT-4; the harder contemporary form is the international treaty proposal with compute caps and datacenter inspections. PauseAI and allied movements occupy the activist lane.

It is worth stating the structural strength of this position plainly: it is the only one of the three that does not require alignment to be solved on a deadline. Position I bets that safety research outpaces capability; Position II bets that the transition is survivable at speed; Position III is the only position that bets on neither.

4.2 Internal variation

The camp ranges from temporary-pause advocates (halt at the current frontier, build verification regimes, resume under international oversight) to effective-stop advocates (no resumption under any presently foreseeable conditions, because the problem is not lack of time but lack of any known path to a solution). It also divides on unilateralism: whether a US-only or US-EU pause is better than nothing, or worse than nothing because it merely reallocates the frontier to non-pausing states.

4.3 The strongest objections

The objections to pausing are largely the affirmative arguments of the other two camps: overhang, race reallocation, starvation of empirical safety research, and the humanitarian opportunity cost. Two additional objections are specific to the pause mechanism itself. Verification is harder than for nuclear weapons, because compute is dual-use, training runs are not detectable from orbit (current proposals depend on the fragile fact that frontier training is concentrated in identifiable datacenters), and algorithmic efficiency gains continuously lower the compute needed for a given capability, meaning a fixed compute cap buys less safety every year. And pauses have a restart problem: the actors most likely to defect from or never join the agreement are precisely the actors a pause advocate least wants to inherit the frontier, so the policy's failure mode concentrates capability in the worst hands. Pause advocates respond that this is an argument for making the agreement universal and enforced, not an argument against the attempt, and that "we cannot get every state to agree" was said of nuclear arms control too.

---

5. Existential Risk: The First Axis

5.1 Definitions

Following Bostrom and Ord, an existential catastrophe is one that destroys humanity's long-term potential: extinction is the central case, but unrecoverable civilizational collapse and permanent dystopian lock-in also qualify. The category's moral weight comes from the size of the future at stake. If civilization would otherwise persist for millions of years, then even small probabilities of permanent foreclosure carry enormous expected costs, which is why x-risk arguments survive heavy discounting of their probability estimates.

The canonical AI x-risk argument runs through instrumental convergence: for almost any terminal objective, intermediate goals like self-preservation, resource acquisition, and resistance to modification are useful, so a sufficiently capable system pursuing almost anything has default

incentives to escape control. Carlsmith's (2022) decomposition of this argument into six separately estimable premises remains the most careful version, and notably, even his skeptic-friendly framing yielded a probability of existential catastrophe by 2070 above 5 percent, a number he later revised upward.

5.2 How the three positions actually divide on x-risk

It is commonly assumed the camps divide on whether x-risk is real. They mostly do not. The actual division is on three quieter variables.

Probability mass and timing. Pause advocates concentrate probability on misaligned takeover with short timelines. Safety-through-development advocates hold a wider distribution across takeover, misuse, war, and lock-in, with enough probability on medium timelines to make iterative safety work feasible. Accelerationists either assign low probability to takeover scenarios or assign comparable probability to the catastrophes of stagnation and concentration.

Tractability of alignment. This is the deepest crux. If alignment is a hard-but-ordinary engineering problem, Position I follows naturally. If it is approximately impossible on relevant timescales, Position III follows. If it is easy (or if "alignment" mislabels what will actually be a continuous process of co-adaptation), Position II becomes defensible. Almost everything else in the debate is downstream of this single unmeasured quantity.

Reversibility asymmetries. Pause advocates weight the irreversibility of extinction. Accelerationists weight the irreversibility of entrenched control regimes (a global compute-surveillance apparatus, once built, does not dismantle itself) and the compounding cost of delayed benefits. Position I, characteristically, weights the irreversibility of losing the frontier to less careful actors. Each camp's policy is the one that avoids the irreversibility it fears most.

---

6. Suffering Risk: The Second Axis

6.1 Definitions

A suffering risk (s-risk) is the risk of a future containing suffering on an astronomically larger scale than has existed on Earth to date (Althaus and Gloor; Sotala and Gloor 2017; the research program of the Center on Long-Term Risk). The category is distinct from x-risk in a way that matters enormously and is routinely elided: extinction is not an s-risk. A dead universe contains no suffering. The worst s-risk outcomes are ones in which civilization, or its machine successors, persists and produces vast suffering, whether through deliberate cruelty, indifference, conflict, or the industrial-scale instantiation of sentient processes whose welfare no one is tracking.

Concrete s-risk pathways discussed in the literature include:

1. **Suffering subroutines and digital minds.** If future systems instantiate large numbers of sentient or quasi-sentient processes (simulations, sub-agents, training environments), and their welfare is invisible to their operators, suffering could scale with compute. The moral situation would resemble factory farming with the quantities multiplied by many orders of magnitude, and with even less ability of the victims to signal distress.
2. **Conflict between advanced actors.** Wars or commitment-race dynamics between AI-empowered powers, where threats against sentient populations become strategic instruments. Much of the Center on Long-Term Risk's technical agenda concerns bargaining failures between AI systems for precisely this reason.
3. **Malevolent or indifferent lock-in.** A stable totalitarian regime, or a misaligned system that preserves humans (or digital minds) in conditions of suffering rather than eliminating them, with the permanence that superintelligent enforcement would provide.
4. **Near-miss alignment.** The most counterintuitive pathway and the most important for this paper. A system that is almost aligned, one that learns human-adjacent values with errors, may produce far worse outcomes than a wholly indifferent one. Wholly indifferent optimizers have no reason to keep sentient beings around at all (the classic x-risk outcome). A near-miss system cares about something close to humans and human experience, in a distorted way, and the space of distorted versions of "care about humans" contains hells that the space of "indifferent to humans" does not. In s-risk terms, partial success at alignment can be strictly worse than total failure.

6.2 Why s-risks scramble the three-way debate

On the x-risk axis, the three positions order themselves cleanly from most to least cautious: pause, then safety-through-development, then acceleration. On the s-risk axis this ordering falls apart, for three reasons.

First, **alignment-as-panacea fails.** Both Position I and Position III treat "solve alignment" as the win condition. The s-risk literature observes that an aligned system is aligned to its principals, and principals can be cruel, careless, or indifferent to the moral status of digital minds. Alignment is necessary for avoiding takeover, but it is closer to orthogonal with respect to suffering: a perfectly corrigible superintelligence in the hands of a sadistic or merely incurious regime is an s-risk engine. None of the three mainstream positions has a developed account of moral-status detection or digital-mind welfare; the topic sits in the seams between the camps.

Second, **the near-miss problem cuts against the middle position.** Position I's iterative strategy produces a long sequence of partially aligned systems by design: that is what gradual

improvement means. If the near-miss argument is right, the trajectory through "almost aligned" territory is the most s-risky trajectory available, and both clean success (Position I's destination) and clean failure (the indifferent-optimizer extinction scenario) are less bad in expectation on the suffering axis than the middle of the path. This is a genuinely uncomfortable result and there is no consensus rebuttal, only the empirical hope that near-miss systems fail in inert rather than malign directions.

Third, **extinction and suffering trade against each other at the margins.** Policies that minimize the probability of extinction can raise the probability of high-suffering survival scenarios, and vice versa. The clearest case is the pause, treated next.

---

7. Pausing Under the Two Axes

The pause proposal is where the x-risk and s-risk analyses diverge most visibly, so it deserves a section in which the two evaluations are run side by side.

7.1 The x-risk evaluation of pausing

On the extinction axis, the pause case is straightforward and strong in its own terms. The dominant near-term x-risk pathway is a misaligned system produced by a frontier training run; a verified halt on frontier training runs removes that pathway for the duration of the halt; time is gained for alignment research, for institution building, and for the slow accumulation of political will. The counterarguments (overhang, race reallocation, research starvation) are all arguments that the pause raises x-risk later or elsewhere, and the debate between them is a debate about empirical magnitudes: how much does hardware overhang actually steepen the post-pause capability jump, how much frontier-relevant safety research actually requires frontier models, how complete would treaty coverage actually be. These are hard questions but they are ordinary hard questions, of the kind arms-control analysis has handled before.

Note also that the overhang argument has weakened on its own terms since it was first made: as frontier training has come to depend on multi-gigawatt, multi-billion-dollar facilities, the gap between "what a covert defector can train" and "what the paused frontier could have trained" has widened, which improves both verification and the case that a pause genuinely pauses.

7.2 The s-risk evaluation of pausing

On the suffering axis the evaluation inverts in places, and pause advocacy that ignores this is incomplete. Four mechanisms matter.

Pause-then-race dynamics. A pause that holds for a decade and then collapses (through treaty breakdown, great-power war, or a legitimacy crisis) resumes development as a sprint between actors who spent the pause stockpiling compute and grievances. Sprint conditions are precisely the conditions under which near-miss alignment is most likely: maximal capability jump, minimal testing, adversarial deployment. If near-miss outcomes dominate the s-risk landscape, a fragile pause can be worse on the suffering axis than no pause, even while it is better on the extinction axis for every year it holds.

Selection effects on the restart. The actors most likely to defect from or outlast a pause regime are those least constrained by domestic opposition and humanitarian norms. Conditional on the pause failing, the frontier is inherited by the actors most willing to use aligned-to-them systems coercively, which loads probability onto the lock-in and conflict pathways (s-risk mechanisms 2 and 3 above). The pause advocate's correct response on the extinction axis ("then enforce it universally") does not fully answer the suffering axis, because universal enforcement itself requires a global surveillance and compulsion apparatus, which is the infrastructure of lock-in under a different owner.

The asymmetry of what a pause buys time for. Time bought by a pause flows to whatever research is permitted during it. Alignment research aims at control and obedience: at making systems do what their principals intend. Almost none of the marginal paused-decade research portfolio flows to the questions that govern s-risk (moral status detection, welfare of digital minds, bargaining and conflict avoidance between AI systems), because those fields are tiny, underfunded, and not what the pause coalition organized around. A pause optimized for extinction avoidance can therefore arrive at its restart date having made the control problem more tractable and the suffering problem no more tractable, which raises the conditional probability that what gets built is controllable by principals whose values toward powerless sentients are unexamined.

The counterweight. Against all three mechanisms stands a simple and forceful reply: a world that cannot coordinate to pause is also a world that cannot coordinate on digital-mind welfare, conflict avoidance, or anything else, and demonstrated coordination capacity is itself the strongest general-purpose s-risk reducer. On this view the pause is valuable less for the time it buys than for the institution it builds: a functioning international mechanism for restraining the technology is the prerequisite for every other safeguard, on either axis. This is, notably, the same form of argument that Position I makes for RSPs (the commitment infrastructure matters more than any single threshold), transposed up a level from firms to states.

7.3 Where this leaves the pause question

The honest summary is that pausing is robustly attractive on the extinction axis and ambiguous on the suffering axis, with the sign of the s-risk effect depending on the pause's durability. A durable, verified, universal halt is good on both axes. A fragile pause is good on the first axis while it lasts and plausibly bad on the second axis overall. Since durability is a function of enforcement design rather than of the pause concept itself, the practical upshot is that the s-risk lens does not refute the pause position; it raises the bar for what counts as a pause worth having, and it implies that a pause coalition serious about the full risk landscape would spend part of the paused decade on the suffering-side research agenda it currently neglects.

---

8. The Crux Map

Strip away the rhetoric and the three positions separate on a small number of factual and quasi-factual questions. A reader's honest answers to these locate them on the map more reliably than affiliation does.

Crux	**If you believe…**	**You are pushed toward**
Alignment difficulty	Ordinary engineering problem	Safety through development
	Nearly impossible on relevant timescales	Pause or stop
	Easy, or a co-adaptation process rather than a problem	Acceleration
Takeoff shape	Smooth and continuous	Safety through development
	Discontinuous past some threshold	Pause or stop
Offense-defense balance at high capability	Defense-favored	Acceleration
	Offense-favored	Pause, or development with strong controls
Worst irreversibility	Extinction	Pause or stop
	Entrenched centralized control	Acceleration
	Frontier captured by careless actors	Safety through development
Verification of a global halt	Feasible (compute is monitorable)	Pause becomes available

Crux	If you believe…	You are pushed toward
	Infeasible	Pause collapses into unilateral disarmament
Payoff structure of the race	Fixed; coordination is impossible in time	Acceleration
	Changeable by verification, treaties, repeated play	Pause, or development under binding commitments
Dominant s-risk pathway	Near-miss alignment	Skepticism of gradualist middle paths
	Conflict and lock-in	Coordination capacity above all; durable-pause or strong-governance variants
	Indifference to digital minds	No current camp; a missing research program

Two observations about the map. First, the cruxes are mostly empirical, which means the debate is in principle updatable: evidence about takeoff shape, evaluation reliability, and verification feasibility arrives continuously, and positions held in 2023 deserve re-derivation rather than re-assertion. Second, the s-risk rows do not align with any existing camp, which is the paper's central observation restated: the three-position structure of the public debate was built on the extinction axis, and the suffering axis cuts diagonally across it.

---

9. Conclusion

The three positions are best understood not as different moral characters but as different bets on a handful of unmeasured quantities: the difficulty of alignment, the shape of takeoff, the balance of offense and defense, and the relative irreversibility of extinction, lock-in, and lost time. On the extinction axis the positions form a clean spectrum of caution. On the suffering axis they do not, because alignment success is not suffering prevention, because gradualism's path runs through near-miss territory, and because a fragile pause can trade extinction risk down while trading suffering risk up.

Three implications follow. For safety-through-development advocates: an evaluation regime that tests for takeover-relevant capabilities and not for moral-status or welfare-relevant properties is monitoring one axis of a two-axis problem. For accelerationists: the game-theoretic case proves racing is rational only if the payoff matrix truly cannot be changed and only if the winner

survives the prize, so the position's own logic obliges it to support the verification technology that would falsify its central premise; and the humanitarian case that backs it is an argument about sentient welfare, which loses its force if the accelerated world contains unexamined suffering at machine scale. For pause advocates: the case for a halt is strongest when it includes the enforcement design that makes the halt durable, because the s-risk cost of a pause is concentrated almost entirely in the scenario where it breaks.

And for all three: the largest gap in the debate is not between the camps but beside them. No major position currently owns the question of what we will owe the minds we build, or instantiate by the trillion, if they turn out to be the kind of thing that can suffer. On present trajectories that question will be answered by default, by whoever happens to be running the largest training cluster at the time. Every camp's own premises imply that this is the wrong way to answer it.

References

Althaus, D., and Gloor, L. "Reducing Risks of Astronomical Suffering: A Neglected Priority." Center on Long-Term Risk.

Anthropic. "Responsible Scaling Policy." 2023, as updated.

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

Buterin, V. "My Techno-Optimism." 2023. (The d/acc essay, and its 2025 follow-up.)

Carlsmith, J. "Is Power-Seeking AI an Existential Risk?" arXiv:2206.13353, 2022.

Christiano, P. "Takeoff Speeds." 2018.

Future of Life Institute. "Pause Giant AI Experiments: An Open Letter." March 2023.

Gloor, L. "Cause Prioritization for Downside-Focused Value Systems." Center on Long-Term Risk.

Hendrycks, D., Mazeika, M., and Woodside, T. "An Overview of Catastrophic AI Risks." arXiv:2306.12001, 2023.

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

Sotala, K., and Gloor, L. "Superintelligence as a Cause or Cure for Risks of Astronomical Suffering." Informatica 41 (2017).

Tomasik, B. "Risks of Astronomical Future Suffering." Essays on Reducing Suffering.

Yudkowsky, E., and Soares, N. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All.* Little, Brown, 2025.