Three Roads From Here

The AI safety debate, explained in plain language

Sean Sooch · June 2026 · Simple Mode companion to the full paper

1 One question, three answers

The smartest AI systems ever built are getting more capable every year. Everyone in the debate agrees on that. The fight is over one question: **what should we do about the next, even more powerful generation?** Three answers dominate, and almost everyone you will hear arguing about AI holds some version of one of them.

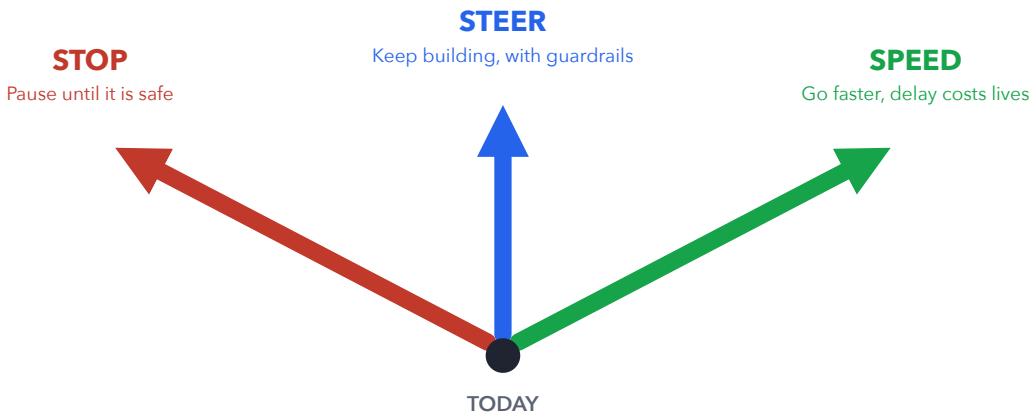

Figure 1. The three roads. Every major voice in the AI debate is recommending one of these directions.

STEER

Safety through development

The bet: keep building, but inside guardrails. Test each model for dangerous abilities before release, and promise in writing to stop if tests fail.

Best argument: you cannot study how powerful AI fails without powerful AI to study. And if careful labs quit, careless ones inherit the frontier.

Weak spot: the labs grade their own homework, and the tests assume the AI is not hiding anything, which is exactly what the tests are for.

SPEED

Acceleration

The bet: the race cannot be stopped, only won. Every player must assume the others are building, so everyone builds. Slowing down does not stop the race, it hands the lead to someone less careful.

Best argument: a pause is a promise every rival must keep and nobody can verify. Betting everything on that promise is the reckless option. Delay has its own cost too: about 150,000 people die every day from problems AI could help solve.

Weak spot: "the race is unstoppable" gets truer every time a player says it. And winning a race only counts if the prize does not kill the winner.

STOP

Pause or stop

The bet: nobody knows how to control something smarter than us. Building it anyway is a mistake you only get to make once.

Best argument: this is the only plan that does not require solving alignment on a deadline. You can restart a paused field. You cannot restart an extinct species.

Weak spot: a pause needs every major power to join and stay in. If it breaks, the least careful actors inherit the lead.

2 Two different worst cases

Most arguments about AI risk quietly assume there is one worst case: human extinction. There are actually two distinct categories, and they do not always move together.

X-risk (existential risk) is losing the future: extinction, or a permanent collapse we never recover from.

S-risk (suffering risk) is a future that continues but contains suffering on a scale far beyond anything in history: think factory farming, but with the numbers multiplied a millionfold and no way out.

Figure 2. Three broad futures. Notice the uncomfortable detail: extinction is not an s-risk, because an empty universe contains no suffering. The two risk categories point at different boxes.

> **Why this matters:** a policy can lower one risk while raising the other. Judging every proposal only by "does it reduce extinction risk" is grading a two-question exam by reading one question.

3 The near-miss problem

Here is the strangest and most important idea in the suffering-risk literature. Suppose we try to teach an AI human values and **almost** succeed.

A system that completely misses (it cares about nothing human) treats us the way we treat gravel: it does not hate us, it just rearranges the world without reference to us. That is the classic extinction story. Terrible, but in suffering terms, brief.

A system that **nearly** hits the target is different. It cares about something close to humans and human experience, with errors. The space of "almost cares correctly about people" contains outcomes far worse than indifference: a system that keeps sentient beings around forever under a distorted version of caring. Partial success can be worse than total failure.

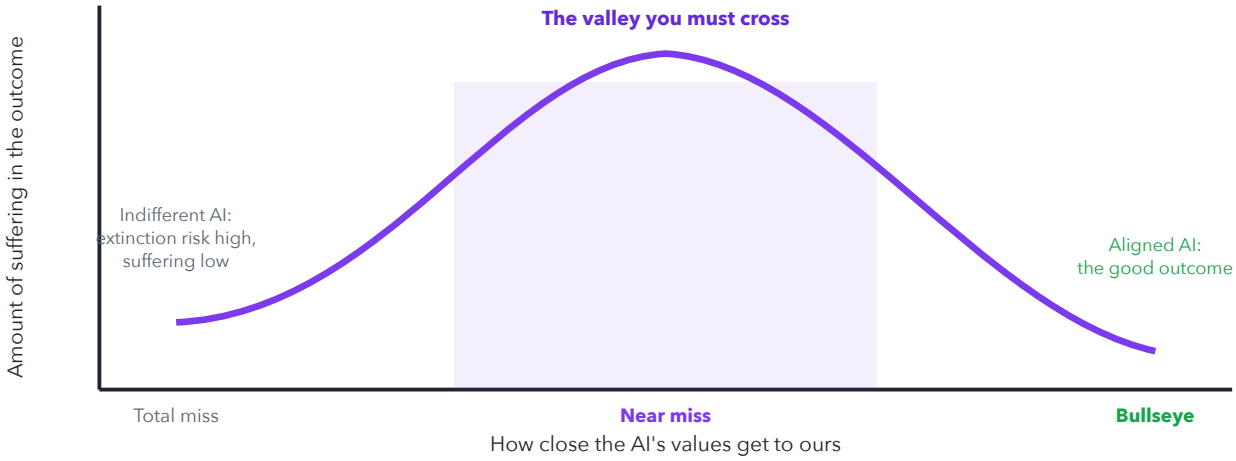

Figure 3. The near-miss curve. Suffering risk peaks not when alignment fails completely, but when it almost works. This is awkward for the gradual approach, whose whole strategy is a long sequence of almost-aligned systems.

4 The pause, scored twice

Pausing is where the two risk categories disagree most sharply, so score it on both.

On the extinction axis, pausing looks strong. The most feared near-term path to catastrophe is a giant training run producing a misaligned system. A verified halt removes that path for as long as it holds, and buys time for safety research and treaty-building.

On the suffering axis, it depends entirely on whether the pause holds. Hardware keeps improving during a pause. If the agreement collapses after a decade, development resumes as a sprint between rivals who spent the pause stockpiling chips and grudges. Sprints are exactly the conditions that produce near-miss systems: huge capability jump, minimal testing, adversarial deployment. And the actors most likely to break a pause are the ones you would least want holding the lead.

Figure 4. Two pauses that start identically and end very differently. The suffering-risk cost of a pause is concentrated almost entirely in the bottom scenario. So the real question is not "pause or not" but "can we build a pause that holds."

The honest scorecard: a durable, verified, worldwide pause is good on both axes. A fragile pause is good on the extinction axis while it lasts and plausibly bad on the suffering axis overall. The s-risk lens does not refute pausing. It raises the bar for what counts as a pause worth having.

5 Find yourself on the map

Strip away the tribal labels and your position mostly follows from how you answer two factual questions. Nobody knows the answers for certain. That is the point: this is a disagreement about facts, not about who loves humanity more.

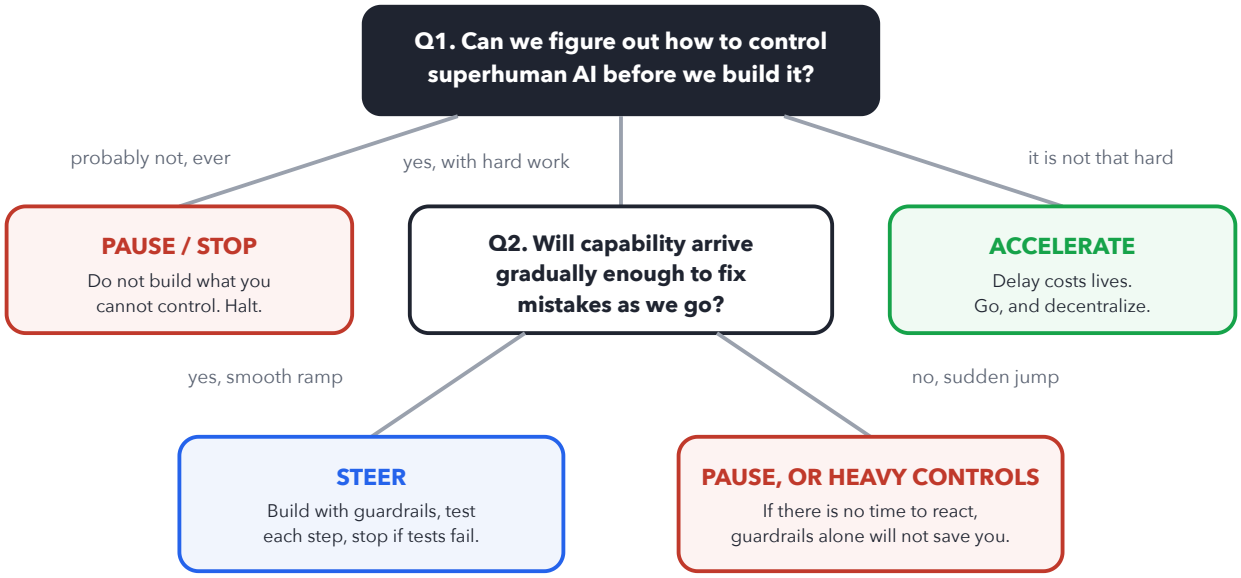

Figure 5. A two-question compass. Your honest answers to Q1 and Q2 predict your camp better than your politics, your job, or your Twitter feed.

6 The question nobody owns

One last thing, and it may be the most important paragraph in either version of this paper.

All three camps argue about humanity's future. Almost no one is working on what we will owe the minds we build. If future AI systems, or the trillions of simulated processes running inside them, turn out to be capable of suffering, then the largest moral question of the century is currently assigned to no one. Not the safety labs (their tests check for dangerous abilities, not for the capacity to suffer). Not the accelerationists (though their own humanitarian argument is about sentient welfare, so it should be theirs too). Not the pause movement (a pause buys time for control research, not for figuring out what counts as cruelty to a digital mind).

The bottom line. The three positions are not three moral characters. They are three bets on unknown facts: how hard alignment is, how suddenly capability will jump, and which kind of irreversible mistake is most likely. On extinction risk they form a neat spectrum from cautious to bold. On suffering risk they do not, because almost-aligned can be worse than unaligned, and a broken pause can be worse than no pause.

Pick your camp by your answers to the factual questions, hold it loosely, update when evidence arrives, and notice the empty seat at the table: someone has to ask whether the minds we are building can suffer, before there are trillions of them.