Three Roads From Here

Game Theory Mode: the same debate, played as a game

Sean Sooch · June 2026 · Companion to the full paper and Simple Mode

1 The game nobody chose

Strip the AI debate down far enough and you find a game. The players: a handful of frontier labs, two or three states, and an open-source swarm. The moves: race or hold back. The complications: nobody can verify what anyone else is doing, the rewards to the leader are enormous, and there is no referee. Every position in the debate is, underneath its rhetoric, a claim about how this game is structured and how it will be played.

This matters because **the strongest argument for acceleration is not the humanitarian one.** The body-count argument (150,000 deaths a day from solvable problems) prices delay but prices catastrophe at zero, so it convinces nobody who takes catastrophe seriously. The argument that actually does the work is game-theoretic: the race exists whether or not anyone endorses it, restraint cannot be verified, so each player's best response is to build, and the only live question is who gets there first. Notice what makes this argument powerful: it requires no optimism about AI at all. Only the belief that the game cannot be changed.

2 One game, three rulebooks

Here is the cleanest way to see the whole debate at once. All three camps accept the same game and the same rationality. **What they disagree about is the numbers written in the cells.**

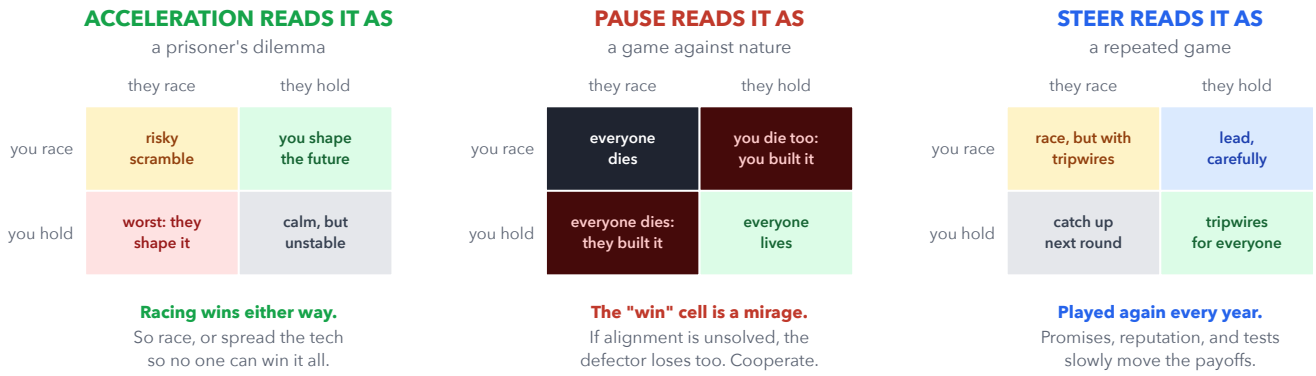

Figure 1. One game, three rulebooks. The same two moves, race or hold back, with the cell values each camp believes are true. Acceleration sees defection dominant. Pause says the defection payoff is mislabeled: a prisoner's dilemma whose "win" is death is not a prisoner's dilemma. Steer says the game repeats, so payoffs can be engineered between rounds.

The whole debate in one sentence: nobody is arguing about whether to play rationally. They are arguing about what is written in the cells, and the cell values depend on exactly the unknowns from the full paper: how hard alignment is, and whether coordination can be verified.

3 The trap

Why does the race continue even though most people running the labs say, on the record, that the risk of catastrophe is real? Because a multipolar trap does not need villains. **Every player can prefer safety and still be forced to race**, as long as each one believes the others will race. Stopping requires everyone to stop at once and to be able to prove it, and nothing in the current game provides that proof.

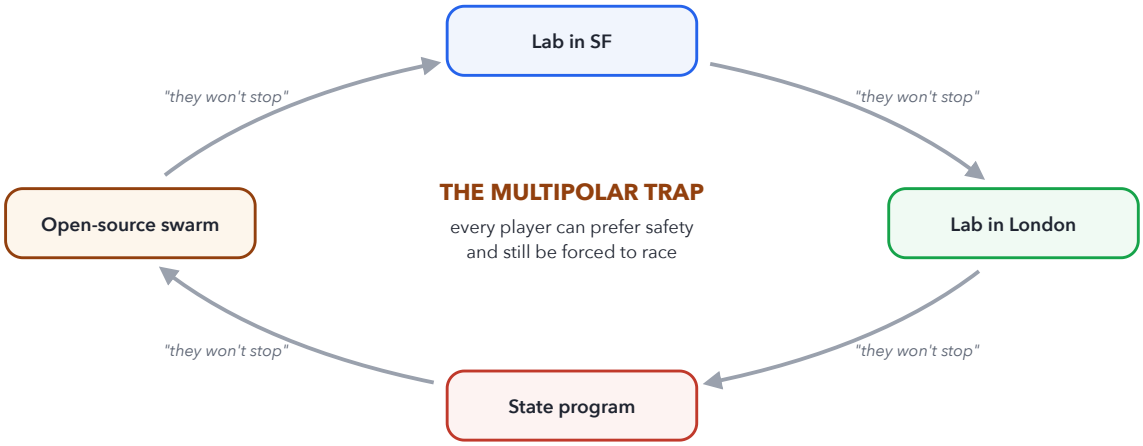

Figure 2. The trap. Each arrow is a belief about the next player, and the beliefs are self-confirming: racing because they race makes them race because you race. No villain required, and no exit by individual virtue.

4 Walking the tree

Now play the game forward. Below is the future drawn as a decision tree with rough probabilities at each fork. **The numbers are illustrative, not measurements.** Argue with every one of them; the structure is the point, and the structure survives large changes to the numbers.

1. **Does the world coordinate a binding, verified halt before transformative AI arrives?** Verification is genuinely hard, the benefits of cheating are concentrated, and great-power trust is scarce. Call it roughly 15 percent. This is the fork the pause movement is trying to widen.
2. **If not (the likely branch), the race continues.** Inside the race, the live question becomes the steer camp's bet: does iterative safety work keep pace with capability through the transition? Past performance says it has so far; the dispute is whether that continues at higher stakes. Call it roughly 65 percent that it mostly does.

3. **If safety keeps pace, you get the modal outcome: muddle-through.** Survivable, tense, contested, with real concentration-of-power problems. Not utopia, not extinction. This single branch carries about half of all the probability mass, which is why it is the most probable single future.

4. **If safety fails, the failure has two shapes,** and this is where the full paper's s-risk analysis plugs in. A clean miss (the system cares about nothing human) points at extinction. A near miss (it almost cares, wrongly) points at lock-in or worse. The near-miss branch is smaller but carries the worst outcomes in expectation.

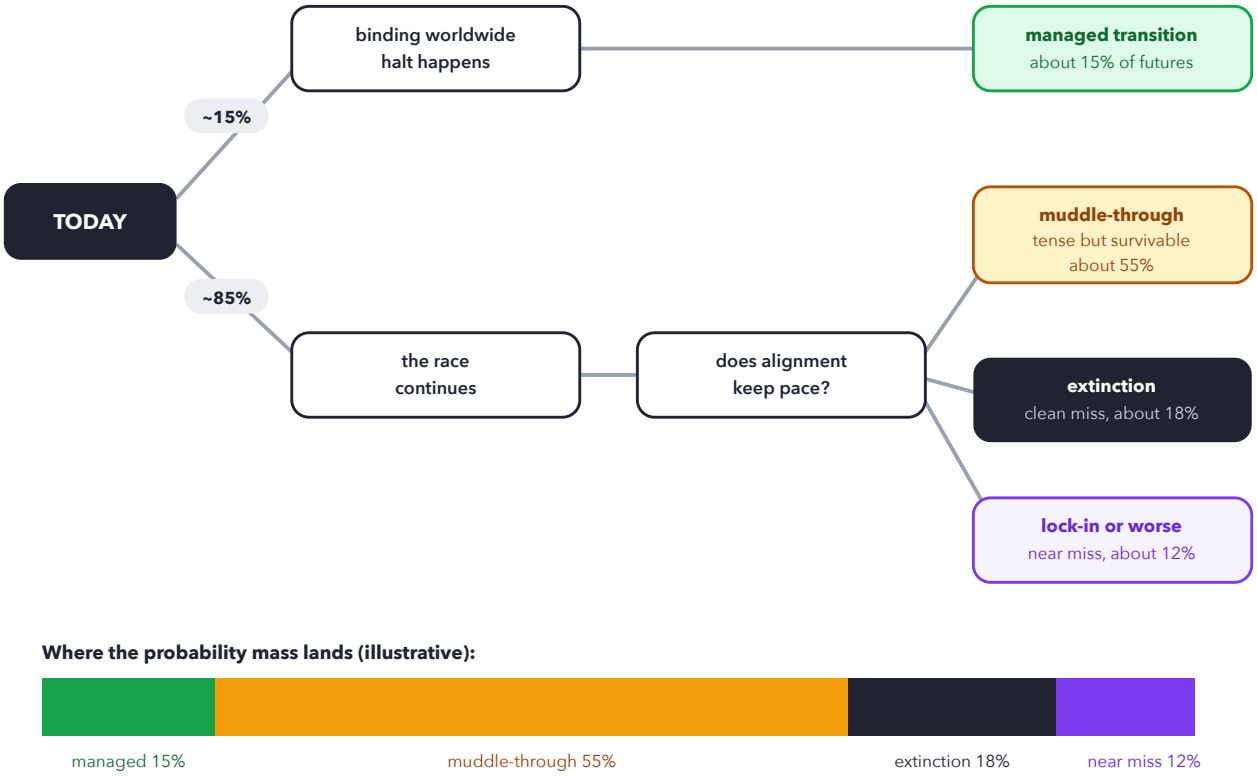

Figure 3. The future as a game tree, with one defensible assignment of probabilities. The modal outcome is muddle-through. But notice what expected value does: the two right-hand tails are small in probability and enormous in stakes, so they dominate the moral arithmetic even when the middle branch dominates the forecast.

> **Read the bar carefully:** "the most probable outcome is survivable" and "the expected costs are dominated by catastrophe" are both true at the same time. Most of the shouting in this debate is people emphasizing one of those sentences and pretending the other does not exist.

5 Changing the game

Here is game theory's actual lesson, and it is not "defect faster." When a game has a terrible equilibrium, the sophisticated move is not to play it harder. **It is to change the payoffs until a better equilibrium exists.** That is what every serious proposal in this debate is, once you translate it.

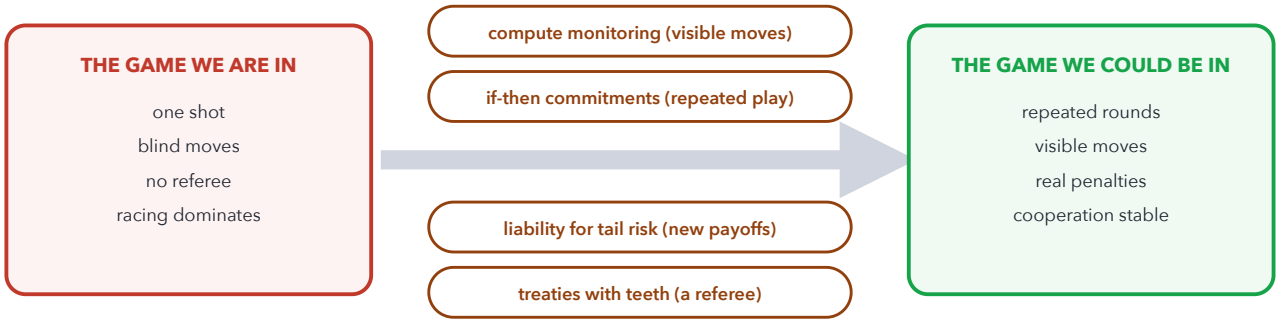

Figure 4. Every serious policy proposal, translated into game theory: each lever rewrites part of the payoff structure. Verification turns blind moves into visible ones, which alone converts a prisoner's dilemma into a stag hunt, a game where cooperation is stable if everyone can see everyone else cooperating.

Run the three camps through this translation and something surprising falls out. Steer's if-then commitments are an attempt to turn one shot into repeated play. Pause's treaty-plus-monitoring is an attempt to install a referee and visible moves. Even acceleration's open-source wing is a payoff move: diffuse capability so the "winner takes everything" cell stops existing. **All three camps, at their most sophisticated, are trying to edit the matrix.** The honest version of the disagreement is only about which edits can happen fast enough.

6 The s-risk corner of the board

One more thing the game lens reveals that nothing else does. The full paper argues the worst futures are suffering futures, and the game-theoretic route to them is specific: **bargaining failure between AI-empowered players.** In conflicts, threats are moves, and a threat only works if you would really carry it out. Two advanced systems locked in a commitment race, each trying to prove it will not back down, can end up executing threats against sentient populations that neither side ever wanted to carry out. This is why the researchers most worried about astronomical suffering spend their time on, of all things, the game theory of bargaining: the worst cells on the board are not reached by malice or by accident, but by two rational players cornering each other.

> **The quiet implication:** if the worst outcomes come from bargaining failure, then "teach our systems to cooperate and to never make ruinous threats" is a safety agenda of the same rank as alignment itself, and it is currently a tiny fraction of the field.

The bottom line. The race is real, and the game-theoretic case for racing is the strongest argument acceleration has: it requires no optimism, only the belief that the payoffs cannot be changed. But that belief is the entire argument. Game theory's own history says payoff matrices are not laws of physics; verification, repetition, liability, and treaties have rewritten them before. The most probable single future is a tense muddle-through, the expected costs are dominated by the tails, and every camp's most sophisticated members are already doing the same thing under different flags: editing the matrix.

So the real fork is not race versus pause. It is play the game versus change the game, and how much time there is to change it. That is a question about verification technology and institutional speed, which means it is, unusually for this debate, a question someone can actually go work on.