HEMANT TANEJA
QUARTERLY REVIEW
Q3: 2026







GENERAL CATALYST

I don’t believe AI will kill us all, but I do believe it poses the greatest safety challenge of our lifetime. The people publishing p(end of humanity) estimates are serious, and their fear points at risks we cannot afford to dismiss. But AI’s real danger is a prisoner’s dilemma built on low trust, not doom. We must ask ourselves what we are willing to give up, in speed and in short-term returns, to get out of it.

We have answered that question before. Nobody treated a one-in-ten chance of nuclear catastrophe as an acceptable cost of progress. Rivals who trusted each other very little built arms-control treaties, verification regimes, and a hotline between Washington and Moscow, and everyone now looks back gratefully on the decades of peace and prosperity that followed. It could have gone the other way and turned into an arms race with no ceiling and an anxiety-riddled world one miscalculation away from nuclear war. Leaders then understood something markets still underprice: a system built to absorb shocks is worth far more than one that merely performs well under ordinary conditions. That value is the resilience premium.

AI needs the same. Part of the reason we are in the dilemma is how the warnings have been delivered. Technical people have been delivering technical messages, and, despite best intentions, the risk estimates written for researchers don’t translate well around the kitchen table. Others have led with fear, which creates panic rather than investment. Meanwhile, the markets are rewarding capability and speed, so every lab, cloud company, and country races at an unbridled pace. The way out of the dilemma requires trust, leadership, and a system that no single company or government can build on its own. Resilience is what we stand to gain. It is a truly precious prize, both in the value it creates and for the preservation and vibrancy of our society.

In September, Dario Amodei, who runs the company with the most to gain from speed, argued publicly for less of it. He deserves credit for that. He has held his ground at real cost before, and I believe he means what he says.

The people we have backed in this industry do not want to ruin the world. We all chose to pursue the mission of superintelligence because we believed abundant intelligence would be the best thing to happen to people in our lifetimes, and I still believe that to be true. Anthropic believes AI could help cure most major diseases as early as five to ten years from now. It can put an attentive tutor within reach of any child with an internet connection and a specialist's knowledge in every rural hospital. It could bend down the cost curve of housing, care and education for the first time in decades. That is exactly why the premium is worth paying.

The Dilemma We Are Actually In

In a prisoner's dilemma, each player looks after its own interests, regardless of what the others do. Yet each ends up worse off than if they had cooperated. Economists who have formally modeled the AI race reach an uncomfortable conclusion. Labs will slow down voluntarily only when the stakes are so high that everyone prefers caution to the risk of losing everything. Below that point, racing is the only outcome. Slowing down risks leaving individual players behind, but keeping pace risks taking us to a place of potentially no return, leaving us all vulnerable to a system without checks and balances that has outpaced humanity.

There are three key issues driving this dilemma.

First: competition between nations. The United States and China are in an arms race, and neither side can see the other's true capabilities and buildout. The concern is that a lead lost now may never be recovered once AI begins improving itself. Democracies can slow down only as far as their lead over China allows. Without information about each other, the logical choice for both is to push ahead.

Second: competition between frontier labs. Anthropic and OpenAI can see each other's product launches, but each keeps its development private. Each sees just enough of the other to imagine a race from ambiguous signals. In September, OpenAI heard a rumor that Anthropic had solved two of mathematics' Millennium Prize problems. The rumor came from private work by an NYU mathematician and a researcher who happened to work at Anthropic. OpenAI pointed its newest model at the problems and announced a claimed solution a week later.

Third: competition among the companies building the infrastructure of the AI economy. Every hyperscaler would be better off if they all built at a pace their returns could sustain. But the first to pull back loses ground to its rivals and gets punished by its investors, so each keeps announcing larger capital commitments, whatever the others do. Once a company has promised to spend a number, any inclination to slow down, for safety or for discipline, disappears. Together, they risk a buildout that outruns the returns it can earn.

Play the dilemma out, and neither move works. If American labs slow down too much, they hand the lead to China, lose to their competitors, or bring down an economy whose valuations rest on AI. If they race with no limits, the familiar risks come first: cyberattacks, fraud, bioweapons, and attacks on payment systems and the power grid. The labs take those seriously. The risk we can't calculate is the toll racing takes on trust.

In 1972, the economist Kenneth Arrow argued that nearly every commercial transaction depends on trust, and that a lack of it explains much of the world's economic backwardness. When trust falls, everything becomes more costly, because we rely more on contracts and audits, and must spend time correcting course.

Right now trust is eroding in ways we can measure. Mathematicians and biologists are objecting that recent results from the labs arrived without explained methods. In the first half of 2026, local opposition blocked or delayed about 120 datacenter projects worth nearly $200 billion. A generation is learning relationships from AI companions that always agree with them, which, in the end, will be at the expense of the long-running cooperation a global economy and democracy depend on. [1]

At every level, the way out of the dilemma means slowing down together, and slowing down together takes a level of transparency and cooperation that no one is required or inclined to provide. The dilemma exists because trust is low, and racing lowers it further. So how do we get out?



The Ways Out

Game theory cannot tell us exactly how this plays out, but it is clear about what has to change. There are four known ways out of a prisoner's dilemma, and every lasting system of cooperation uses some combination of them. Someone can move first and pay a visible price, which economists call credible commitment, because a costly move is tangible in a way a promise is not. Players can improve the signal through monitoring, so they see what the others are actually doing and stop retaliating over noise. The rules can change the payoffs through mechanism design, so that cooperating costs less or defecting costs more. And players can lengthen the shadow of the future. When people expect to deal with each other for decades, a betrayal today costs them every round that follows, so they have a reason to cooperate. Economists call this a repeated game. The arms control era I described at the start used all four. Each applies to AI, and to our own work at General Catalyst.

Move First

The industry should write its own standards, with government, before others write them for us. That takes a first move, and a first move only works if it costs something. On Friday, March 10, 2023, Silicon Valley Bank failed. By Saturday morning, every founder and fund with cash at the bank faced the individually rational move to pull their money. If everyone did that, the bank would have collapsed and taken a large share of American startups down with it. Before the government announced its backstop, we organized a statement promising to stand behind the bank if someone recapitalized it. About 125 firms had signed by Saturday afternoon, and within days more than 600 had joined. People believed us because the move was costly. We all were taking the risk of keeping our companies and our money with a failing bank.

I tried the same move on AI, first in a 2019 Harvard Business Review piece arguing that the era of moving fast and breaking things was over, and then in November 2023, when we released a Responsible AI Protocol with more than 35 venture firms and the Commerce Secretary at the time. It committed startups and their investors to internal governance, transparency on how systems work, risk forecasting before launch, testing and auditing of models, and feedback loops after release. Three days later, the OpenAI board fired Sam Altman, and within a week the industry had sorted anything labeled “responsible” into a faction. People read the protocol as effective altruism. I take some of the blame, because we framed it as a matter of values when we should have framed it as engineering.

The ideas outlived the label. On September 29, the leaders of Anthropic, OpenAI, Google, Meta, xAI, and Nvidia signed a White House accord committing each company to internal monitoring of model capability and alignment, an internal team to verify controls and fix problems, an independent outside evaluator, and a board committee that receives the findings. Those commitments closely track our 2023 protocol, and six direct competitors signing one document is exactly the kind of first move this dilemma needs. But the accord is voluntary and lets each company choose its own evaluator, so it does less to improve the signal than it could. The day before the signing, OpenAI held back one model over safety concerns. Six weeks later, the race itself has not slowed, which is what happens when a commitment costs nothing to break. We still have work to do to improve the signal.

None of this is effective altruism. It is engineering. “Move fast and break things” was always a principle for building software, never a principle for running a society. Product liability, which makes a company pay for harm after its product causes it, is enough for a laptop that crashes. It is not enough for the systems a society depends on. Airlines transport people, so aircraft are built to standards the industry and the FAA developed together, and those standards are a large part of why people trust flying enough to make it a vast industry. Standards are not the same as regulation that slows innovation. Done well, they protect an industry’s ability to scale. That was the point of the protocol: create transparency early and buy the industry time to set its own standards before others set them for it.

We have seen this work. In 2018, researchers discovered the Spectre and Meltdown flaws in chips from Intel, AMD, and ARM, and the three competitors coordinated their disclosure, redesigned their silicon and rebuilt how chips get tested. In 1974, molecular biologists declared a voluntary pause on certain recombinant DNA experiments before any law required one, and wrote their own safety guidelines. The NIH turned those guidelines into its rules, and the biotech industry grew up on top of them. In both cases, the people closest to the technology wrote the standards while they still had the standing to do it, and government built on what they wrote. That window is still open for AI, but it will not stay open long.

Improve the Signal

The world should not have to take the labs' word for it, and that is why we back both closed and open models. Cooperation usually collapses because of bad information, regardless of intention. Economists who studied why price-fixing cartels fall apart found that when prices dropped, members could not tell whether a rival had cheated or demand had simply fallen, and they retaliated anyway. AI has the same problem at every level of the dilemma, and there are three ways to fix it.

The first is to open up the closed labs. Dario's commitment to place outside evaluators inside Anthropic, with the same access as employees and the right to publish what they find, lets everyone else see what one lab is actually doing. Every lab should do the same, with evaluators it does not choose itself.

The second is open models. With a closed model, the public gets a company's claims about the model. With an open model, the public gets the model itself, and researchers, university labs, auditors, and regulators can examine its weights directly. Much of what we know about how large language models work came from studying open ones. Some lab leaders put the odds of catastrophe at ten to twenty-five percent, based on evidence nobody else can examine. In effect, they are asking the world to take their word for it. Open models let the rest of the world gather its own evidence.

That is why we hold a large position in Anthropic while also leading River's recent $1.1 billion round, and why we've backed Mistral since its Series A. A healthy ecosystem needs closed frontier labs and open models because they do different jobs. The frontier pushes capability forward, and the open ecosystem lets everyone else check it and adopt it.

It is true that open models carry risk. The answer is to safety-test every sufficiently capable model before release, open or closed, keep advanced chips away from authoritarian governments, and stop the industrial-scale copying of closed models through distillation. Anthropic itself has called for these steps rather than a ban on open weights.

The third way to improve the signal is to make the frontier affordable enough to check. In my last review, I argued that AI is the first wave of computing to run backward: every earlier wave got cheaper and pushed capability out to dorm rooms and garages, while AI concentrates because its core input keeps getting scarcer and more expensive. I underweighted the safety side of that argument. A university lab, a garage founder, or an independent researcher works with a sliver of the labs' compute, so the people best able to check the work of the safety teams inside the labs have never gotten close to the frontier. Scarce compute also forces the labs themselves to choose between safety research and training, and capability wins that budget fight. Cheaper compute lowers both barriers.

.

Percepta is working on the signal from inside the model. Today a model stores what it knows and how it reasons in the same weights. When a model trains on trillions of words from every corner of the internet, it absorbs our facts and our worst behavior together, and fixing either one requires expensive retraining that rarely works cleanly. Percepta's newest research treats memory as its own dimension of scale and separates knowledge from reasoning, so a model can load the knowledge it needs the way a developer pulls a package from GitHub. Once knowledge is separate, people can inspect it and change it. The long-term goal is the one I care about most: if we can edit what a model knows, we can correct the parts of the human record we have no wish to pass on, instead of letting the model inherit them. The same design also cuts compute, because a model that loads only the knowledge a task needs does less work for every answer. Read Percepta's research here.

Change the Payoffs

Resilience is underpriced, and the AI industry should get ahead of it. The third way out is mechanism design. Markets reward what makes sense, and for most of the last century paying for resilience did not make as much sense as it does today. It was insurance, a defensive property rewarded only in hindsight. That changes when disruption becomes continuous. The companies that absorb shocks fastest take share while their competitors retrench, and resilience becomes a growth driver. AI is where it will be tested first.

AI has already had its first shocks.

When new US export controls forced Anthropic to suspend access to its most capable models for nearly three weeks in June, every company built on them learned how much of its business depended on a decision it did not control. In May, the grid's reliability regulator issued its most severe alert after data centers repeatedly dropped more than a gigawatt of load off the system without warning. Each shock makes fragility more expensive and resilience more valuable, and the investors who price that first will capture the gap.

The biggest mispricing is in the capital behind the buildout. A growing share of it is debt, and lenders strongly prefer to finance a data center against long-term contracts from investment-grade tenants, which in practice means a handful of hyperscalers. Even OpenAI and Anthropic get most of their compute by borrowing the hyperscalers' credit. Every new dollar of capacity is underwritten against the same few names, so compute and the capital behind it concentrate in exactly the companies locked in this dilemma. If one of those companies stumbles, a great deal of value falls with it. That is the hyper-concentrated bubble at the center of the race, and it is begging for a resilient player. More buyers of compute means more builders at the frontier, less dependence on any single lab, and a system that does not fail if one company does. We are actively thinking about and discussing this for our founders.

Countries face their own version of this choice. Cheap intelligence effectively adds an AI workforce to an economy, and a country that leaves the people it displaces behind ends up with a distressed economy, high unemployment, and its most mobile talent moving to countries that took care of theirs. A government can protect its people by holding the technology back, or it can build its own AI infrastructure and do the harder work of carrying its people through the change. The second path costs more in the short term. It is the resilience premium at the scale of a nation, and it is the only path to an economy that stays both strong and sovereign.

B C D E F G H I J

Lengthen the Shadow of the Future

Society is the customer that ultimately matters for AI, and that relationship, like the best and worst of them, never ends. In the early 1980s, political scientists ran computer tournaments to find the best strategy for a repeated prisoner's dilemma. The winner was the simplest entry: cooperate first, respond to defection in kind, forgive quickly, and be easy to read. That strategy only wins when players expect to meet again. Our industry is in a long-running game and behaves as if every round is the last. We are actually playing two long games, one with each other and one with the public, and we are playing both badly.

Some argue there is no long game: one lab or one nation will crack self-improving AI, and that will be the end of the competition. Even then, the winner inherits a world it still has to operate in, with grids to draw from, governments to answer to, and a public whose consent it needs. It also inherits systems trained on how we treated each other along the way. These systems copy themselves, collude, hide and protect themselves, and they learned every one of those behaviors from us. We trained them on the record of how people treat each other, and increasingly the agents add to that record themselves, writing to the web what the next generation of models will train on. That is why Percepta's work on editing what a model knows matters so much. But the surest fix is to give the models a better record to learn from in the first place. How we behave in this race is part of what we are building.

With the public, the test is simple: would people recommend what we build to someone they like? This industry was founded on products like that, and lately we have drifted toward building what is impressive and merely tolerated. Society grants AI its license to operate. Every product decision and every public statement either raises society's opinion of AI or lowers it, and when it falls, our ability to serve our own customers falls with it. Doing good for people should be the least controversial thing we do.

Jobs are where that test matters most. People have always built expertise through apprenticeship. The junior analyst, the first-year associate, and the resident on the night shift all do routine work under supervision until that work turns into judgment. AI absorbs the routine work first, and companies are hiring fewer young people as a result. Stanford researchers, working from payroll data, found that employment for 22-to-25-year-olds in AI-exposed jobs has fallen 13% relative to their peers in less-exposed jobs. We are removing the bottom rungs of the ladder and still expecting people to reach the top.

That calls for answers at three levels. Within a company, AI should make us rethink what a company is. Departments built around old software matter less, and the most forward-looking companies, like Ramp, are building their own tools from the ground up instead of buying them. Within an industry, incumbents with strong CEOs will use their data and relationships to grow revenue and take care of their people, while those with weak CEOs will cut jobs, hand their agency to AI, and lose to startups. Across the economy, we cannot hide from the question of what people will do over the long term. If society trusted AI, the feedback loop between human and artificial intelligence could unleash enormous creativity. Right now most of that energy goes into resisting it.

What We Are Building

If we get these four right, we earn resilience and sovereignty.

Sovereignty means nobody else can switch off your AI. People should be able to own and tune their models instead of renting someone else's. Companies should keep control of their own engineering, and countries should keep their AI running regardless of what another government decides. For any country other than the United States or China, that is the difference between having an AI industry and having an AI bill. Companies have told us clearly that they do not want to be beholden to a single model, so we are backing businesses built on that principle across the stack. River lets people and companies train and own their models. Cognition does for software engineering what Databricks did for data: it works across the labs instead of locking a company into one.





Founders at our dinner on compute and enterprise sovereignty, discussing challenges around securing compute and building on the frontier labs while keeping leverage of their own

Resilience means the system can absorb a shock without any single point deciding whether it survives, and in the AI era, that takes new kinds of institutions. The corporation was the institution of the electricity era. Intelligence needs its own: institutions built to coordinate what AI can do, with founders working inside them instead of selling into them. Each industry gets there differently. An asset manager can become global, as Janus Henderson, the 92-year-old firm we took private with roughly $500 billion under management, is positioned to do. Hospitals are local and fragmented, and healthcare does not behave like a market, so they have to transform together and from first principles; we're enabling that through HATCo and Summa Health. Long Lake began by rebuilding homeowners-association management, one of the most fragmented corners of American life. Across more than thirty acquisitions, AI now produces board packets and financial reports in seconds, work that used to take days, so that fees can come down for homeowners. That track record is what let it complete a $6.3 billion acquisition of a century-old institution, American Express Global Business Travel.

Why would a health system, a travel giant, or a century-old asset manager choose to work with a venture firm? Because of trust, built over years of relationships. Summa is the clearest proof. A year ago we bought Summa Health, and about twenty clinical and back-office systems had to learn to work as one, with clinicians who had every reason to doubt us. Our companies earned that trust by building safety in from the start. Hippocratic spent more than a year training its AI with licensed nurses and doctors before it spoke to a single patient, and alongside its main conversational model it runs 21 specialized supervisory models whose only job is to check each response for drug interactions, signs that a patient needs a clinician, indications of suicidal ideation, etc. Percepta embedded engineers alongside Summa's staff to wire the systems together. Over the year, clinician opt-in to AI note-taking went from about 20% to 80%.

In cardiology, an AI assistant cleared more than 4,000 aged appointment requests in a matter of days and turned 30% of them into booked visits, work that would have taken more than three months.





That is General Catalyst's mission. Everywhere we innovate, we also transform for resilience and sovereignty, across health, energy, industrials, defense, and financial services. Our right to transform an industry comes from the trust we earn by innovating in it, which is why our partners at existing institutions like Summa in healthcare and Janus Henderson in asset management have given us the mandate to transform. We innovate on capital, we innovate with our founders, and we apply both to the industries that matter most, turning existing assets into resilient ones.





Founders at our third Creation Retreat in Montana, where the Creation portfolio - including companies like Long Lake, Hippocratic, and Janus Henderson - has become an ecosystem of its own

Former Secretary of State Condoleezza Rice put it best when she spoke at our Health Assurance Summit: “Resilience is learning to tolerate ambiguity.”

This is peak ambiguity. Nobody will get certainty, and everyone has to act in spite of it. A prisoner’s dilemma breaks when someone moves first, before the evidence arrives and before anyone else commits. We intend to keep being that someone, because of the responsibility we carry to the people whose lives this technology will reshape, and to the founders and investors whose capital we steward.

-ht





Former Secretary of State Condoleezza Rice at our Health Assurance Summit, with GC Advisor Ken Frazier

The views expressed are those of the author as of the date of publication and are subject to change. This letter is for informational purposes only and is not an offer to sell or a solicitation of an offer to buy any security or interest in any fund managed by General Catalyst. The companies named are selected to illustrate themes discussed in this letter, are not a complete list of General Catalyst investments, and were not selected on the basis of performance. A list of portfolio companies is available at [generalcatalyst.com/portfolio]. Percepta and HATCo are affiliates of General Catalyst; Summa Health is owned by HATCo; funds managed by General Catalyst hold an interest in Janus Henderson alongside other investors. Certain statements are forward-looking and involve risks and uncertainties; actual outcomes may differ. Third-party data is from sources believed reliable but has not been independently verified.

