ESSAY · COMMENTARY · AUGUST 30, 2026

Deployable Trust

A Practitioner Reads Dario Amodei's "Machines of Loving Grace"

On the October 2024 essay, read in August 2026

ANANTHA PADMANABHAN · FOUNDER & PRINCIPAL RESEARCHER, CAPITAL MARKETS AI

> In the regulated world, the binding constraint on beneficial AI is not capability. It is *deployable trust*.

Opening

I read this essay in August 2026, two years after it was published. That gives me hindsight, and I have tried not to abuse it: what I am reviewing is the argument about what powerful AI does to the world once it arrives, not the 2024 predictions. I have read it as a reviewer should: position first, contest where warranted, ideas not the author.

The essay is strongest where it is most grounded. "Marginal returns to intelligence" is the right question — when intelligence becomes cheap, what else becomes limiting? — and the five limiting factors he names (the speed of the outside world, the need for data, intrinsic complexity, constraints from humans, physical laws) are an honest answer to it. Where the essay soars, it is a vision, and he says so: this is what the world looks like if everything goes right. I take the genre as given.

The interesting question is what the vision leaves out.

Three things, which I will carry into each of the five sections. They are not a sixth limiting factor. They are pressure on the one he already named and then underweighted: constraints from humans.

First, the benchmark. "Smarter than a Nobel Prize winner across most relevant fields" is his measure of a single model, and he scales it, correctly, to "a country of geniuses in a datacenter" — millions of copies, acting in the world, directing teams of humans. But a million geniuses is more intelligence, not an institution. Most of the value in the modern economy is produced through institutions — firms, agencies, exchanges — not by the smartest person in the room, and institutions are how intelligence acquires accountability and pace; they move for reasons that have little to do with the intelligence available to them. The most famous fund with Nobel laureates as partners was Long-Term Capital Management; it failed because the institution around the models could not absorb a tail the models had assumed away. Halberstam wrote the book-length version of the point: the most credentialed cabinet in American history, wrong about Vietnam for a decade, because brilliance without domain feedback compounds error instead of correcting it.

Second, slack. The essay describes many societal structures as "inefficient or even actively harmful," and some are. Inefficient from whose perspective, and on what time horizon? Structures that evolved bottom-up carry redundancy because they survived. What looks like waste to whoever wants to move fast is, from inside the system, often the reason it is still standing. Putting risk technology into banks taught me to price that slack before trying to remove it.

Third, verification. The essay wants the verification step shorter. On clinical trials it notes that mRNA vaccines for COVID "were approved in 9 months," argues they "arguably should have been approved in ~2 months," and treats "~1 year end-to-end for a drug" as compatible with its 5–10 year compression. That is a coherent speed argument. It is also where the argument is most exposed. When generation becomes cheap, verification becomes the product. A desk that can produce a thousand models a day does not abolish model risk; it industrializes it. Second- and third-order effects are where the bill arrives, and they come, almost by definition, from the things we were sure we understood — the unknown knowns, the one quadrant the famous taxonomy managed to leave out. The question for each section is not how fast the country of geniuses can invent, but how fast the world can validate what it is asked to live with.

Running through all three is a single claim I will test section by section: in the regulated world, the binding constraint on beneficial AI is not capability. It is deployable trust.

1. Biology and health

This is the section with the highest ambitions — compressing 50–100 years of biological progress into 5–10 — and the essay is candid that the physical world resists: cells grow at their own speed, and experiments take the time they take. Fair enough. My reservations are about two layers the section does not model.

The first is ownership. He is explicit that he does not mean AI as a data-analysis tool. He means AI that will "perform, direct, and improve upon nearly everything biologists do," including giving directions to human lab workers — the AI as principal investigator. The regulated world's response to that is not alarm; it is a question: who is the owner of record? Every model running in a bank has a named human who signs for it and answers for it when it is wrong. That is not friction. It is how accountability is constructed, and the essay has no owner-of-record layer at all.

The second is irreversibility. Biology is a fat-tailed domain in Mandelbrot's and Taleb's sense: raising the rate of interventions raises the exposure to tail outcomes unless verification capacity rises with it. And some interventions cannot be unwound. He himself floats gene drives to wipe out the mosquitoes that carry malaria. Some proposed drives are designed to be confinable; the suppression case he is using is not a trade that can be closed out. And second-order effects in this domain have a track record: antibiotics bred resistant bacteria — Fleming warned of it in his Nobel lecture — insecticides bred resistant mosquitoes, and a drive that selects hard enough can be expected to select for whatever escapes it. What role the target species play in the systems that evolved around them is something we partly know — and "partly" is doing a great deal of work in a decision that cannot be taken back. The verification bar should scale with irreversibility, which is the opposite of treating trial requirements as overhead to minimize.

History supports the slack argument here, told carefully. Thalidomide, marketed in Europe from 1957 and withdrawn in late 1961, devastated families in the countries that approved it quickly. The United States never licensed it, largely because one FDA reviewer, Frances Kelsey, withheld approval under heavy commercial pressure while the safety file was thin. Americans were not untouched — thousands received the drug in company "trials" — but the country was spared a marketed catastrophe. The regulator's slowness was the entire value, and the 1962 Kefauver-Harris amendments built the modern verification regime on the back of it. Two years on, the picture is consistent with that discipline: the first AI-designed candidates are in late-stage trials — Insilico's rentosertib among them — which is a real landmark, and none has yet reached full approval. The pipeline is real. The clock is institutional.

Where the essay is right and deserves the credit: eradication is his example, and it is a good one. Smallpox was ended by logistics and political will sustained over decades, not by a molecule alone. That is an institutional achievement wearing a medical coat, which is rather my point.

2. Neuroscience and mind

Much of the biology argument transfers, so I will note what is different. The essay is better here than a fast read suggests: it does not reduce human improvement to pharmacology. There is a serious paragraph on behavioral interventions and an "AI coach who always helps you to be the best version of yourself" — baseline experience improving beside drugs, not only through them. I agree, and a critical reader should concede it.

The observation worth keeping is about who decides. Every time this essay reaches a "who chooses?" question — cognitive enhancement, biological freedom, self-actualization — its answer is the individual: "everyone should be empowered to choose what they want to become." As a value, I share it. As a model of how change reaches people, it is incomplete, and the gap runs through the whole essay: between the individual choosing and the state deciding sits the institution — the family deciding for a child, the employer, the insurer, the hospital system, the school — and that middle layer is where most of these technologies will actually be procured, filtered, or refused. The essay does not have that layer in any operational sense.

Both of these sections describe the happy path — by design, and fairly. But any architect will tell you a system is defined by what happens when it fails, and brains and bodies do not ship with rollbacks. The essay has no error-handling path, and in these two domains that is the part I would want designed first.

3. Economic development and poverty

I am a technologist; my domain is the application of technology in capital markets. So I will not argue development economics with this section — and by his own admission it is where the author is least confident: "I am not as confident that AI can address inequality and economic growth as I am that it can invent fundamental technologies." I respect the honesty, and I agree with the concession behind it: invention and distribution are different problems, and the second is institutional.

What I can speak to is adoption, and here the essay runs on two modes. Technologies spread by individual choice (cell phones permeating Africa "via market mechanisms"; anti-technology movements "more bark than bite") or they are blocked by collective political decision (nuclear power). Both are real. But there is a third mode, and it runs most of the developed economy: institutional adoption inside regulated enterprises — neither a consumer choosing nor a legislature banning, but a risk committee, a validation team, an audit function, working through a process. Companies appear in the essay. This process does not.

I have watched that third mode from inside for two decades, and it has a signature. The functions that adopt first are the ones with the most to gain from speed and the clearest way to measure it: the trading floor before the back office of the same bank, typically by years and often by a decade. The sorting variable is not the sector. It is how much validation a function requires before something is allowed to run. Reversible, P&L-measured functions move fast. The regulated core does not, and it is not supposed to.

Which is why "AI finance ministers and central bankers" is the sentence I underlined twice. Having sold risk platforms to banks and to a monetary authority, I can report the first questions such an institution asks of any system: not how intelligent it is, but who controls it, who validated it, what happens when it is wrong, and whether it is sovereign to us. Those questions are not obstruction. They are the adoption process itself, and any AI advisor to a government will live or die by them, not by its intelligence. The essay's own skepticism about central planning — Hayek on dispersed knowledge, the "socialist calculation problem," the economy as a chaotic system that "has to be managed in a mostly decentralized manner" — points the same direction; it just stops before drawing the institutional conclusion.

On distribution the essay and I agree more than we differ: growth without distribution is his stated worry, his proposed remedy is state capacity — an institutional answer — and development, in the end, is something countries do through their own institutions, not something done to them, however benign the intent. A country that adopts an AI advisor it cannot run, audit, or replace has not adopted a technology; it has acquired a dependency — and dependencies, unlike technologies, come with terms.

4. Peace and governance

I will keep this short, because most of what I believe about geopolitics does not belong in a technology review. Two notes.

The essay is honest that AI does not structurally favor democracy: "if we want AI to favor democracy and individual rights, we are going to have to fight for that outcome." Agreed — Franklin's reported answer at Philadelphia, "a republic, if you can keep it," was the same warning two centuries earlier: survival is maintenance, not architecture. He treats propaganda and surveillance as tools in the autocrat's kit, and he notes the fight inside each country. What he still underweights is the temptation inside democracies themselves. Surveillance does not require an autocrat. It requires a state that can, and a public that gets used to it.

And the "who decides" gap appears once more: the remedies here are state capacity and individual empowerment, with the same missing middle — the courts, the benefits administrations, the agencies through which government would actually absorb AI, at the speed their own assurance processes allow.

5. Work and meaning

This is the section that stayed with me, and the reason is uncomfortable for a reviewer: it is the section where the essay claims least. On the post-labor economy, the author writes that "no one today has done a good job of envisioning" it, offers candidates — a universal basic income among them — without endorsing any, and says it is not possible to know whether they will make sense "without lots of iteration and experimentation." I trust the essay most where it claims least. On the substance I agree with its two anchors: meaning comes mostly from human relationships and connection, not from economic labor; and comparative advantage is a fair account of the near term.

What the section leaves open is what sets the pace. Displacement does not arrive economy-wide; it arrives when employers adopt, and employers are institutions. Everything in section 3 therefore applies to the labor transition directly: it will move at the speed institutions can validate what they deploy — function by function, with a long distance between "the technology exists" and "the institution runs it."

I have one measurement, not a theory of history. At Calypso, a Tier-1 institution averaged fifteen months from first contact to contract signature — working demonstrations on the client's own data, onsite walkthroughs of their test cases, legal redlining — with go-live after that. Almost none of that time was spent on whether the software could do the job; the first demo settled that. It was spent on whether the institution could rely on it — on their data, under their controls, with their names on the sign-off. That was deterministic software; for probabilistic systems the bar rises. One vendor's clock is not the whole labor transition. It is the part of the transition the essay does not describe.

That clock is human. For the people inside a transition, pace is the difference between a workforce that gets time to adapt and one that gets a shock. So the question this section needed to ask is not whether meaning survives the destination — I think the essay gets that right — but who governs the speed of travel, and whether we are strengthening the institutions that set it.

This is more hopeful than it sounds. The oil minister Yamani said the Stone Age did not end for lack of stone. Add the other half: better options ended it. Work changes when better options arrive, and inside institutions, "better" includes "validatable." Building AI that institutions can validate is the unglamorous work of the transition.

Concluding notes

Predictions about powerful technologies age badly, and the people with the most standing to make them have often known it. Oppenheimer, who could have prophesied about the atom with more authority than anyone alive, put his public energy instead into how it would be governed — the Acheson-Lilienthal plan, the international-control debates of the late 1940s. This essay, to its credit,

hedges like a man who knows the same thing ("everything I'm saying could very easily be wrong"). Where it soars anyway, take it as intended: a direction of hope, not a schedule.

My one addition, made from four directions in these notes, is the same each time. Intelligence is measured in models; value in the regulated world is produced through institutions. Institutional slack is redundancy, not waste. When generation becomes cheap, verification becomes the product. And decisions in the essay belong to the individual or the state, when most of them will actually be made in the middle layer between the two. The binding constraint on beneficial AI in that world is not capability — it is deployable trust, and the reliable way to shape a powerful technology to productive purposes is to strengthen the governance structures, private and public, where that trust is manufactured.

That is not a counsel of delay. It is where the work is.

Postscript (September 2026). The companion essay, The Adolescence of Technology, is reviewed separately in Buying Time — capmarkets-ai.com/essays/buying-time.

A note on method. I wrote this review from my own reading notes, with Claude as research and drafting partner — a tool made by the company whose co-founder's essay is under review, which the reader is entitled to weigh. The draft then went through one round of adversarial review by Gemini and Grok under a findings-only brief. I adjudicated every finding, accepted some and rejected others, and checked every quotation against the essay myself. The judgments, and any errors, are mine.

References

- Dario Amodei, *Machines of Loving Grace* (October 2024) — darioamodei.com/essay/machines-of-loving-grace
- J. Robert Oppenheimer, "International Control of Atomic Energy," *Foreign Affairs* (January 1948); *Atom and Void: Essays on Science and Community*
- David Halberstam, *The Best and the Brightest*
- Benoit Mandelbrot and Richard Hudson, *The (Mis)behavior of Markets*
- Nassim Nicholas Taleb, *Antifragile: Things That Gain from Disorder* and *Statistical Consequences of Fat Tails*

ABOUT THIS ESSAY

Anantha Padmanabhan is Founder and Principal Researcher at Capital Markets AI. He spent two decades putting risk technology into banks, exchanges, and a monetary authority, and now builds agentic AI platforms and publishes research at the intersection of institutional risk and frontier AI.

Published August 30, 2026; postscript added September 8, 2026. The canonical version of this essay lives at `capmarkets-ai.com/essays/deployable-trust/`

© 2026 Anantha Padmanabhan. All rights reserved. Quotation with attribution is welcome; for republication in full, write to anantha@capmarkets-ai.com.