TEC | THE EPISTEMIC CENTURY

VOLUME I

The Epistemic Century

How AI could change the institutions that generate, test, fund, and act on claims.

ARYA SOMU

OCTOBER 2026

CONTENTS

01	The bottleneck was never intelligence	3
02	What actually changes, mechanically	4
03	Institutions as epistemic machines	5
04	Replication as a continuous background process	7
05	Forecasting as institutional infrastructure, not special-occasion theater	8
06	Resource allocation without the queue	9
07	Adjudicative institutions: courts, regulators, and the standard of scrutiny	10
08	A market for adjudication	11
09	What builders should actually do differently	12
10	The compressed decade	13
11	A concrete morning	14
12	What could go wrong, said plainly	16
13	Why write this down now	18
	Notes	19

The Epistemic Century

ARYA SOMU

I want to try something Dario Amodei did well in "Machines of Loving Grace"[1] and that most AI writing does badly: pick one domain, go deep enough to be concrete, and let the optimism earn itself through mechanism. The domain I know best isn't biology or neuroscience; it's the machinery underneath science and institutions themselves, the process by which claims get generated, tested, funded, believed, and acted on. I think this machinery is about to change more than it has in a century, and almost nobody is writing about it like it matters, because it's boring in the way that plumbing is boring, right up until you understand that plumbing is the reason cities don't have cholera.

So this is an essay about plumbing, specifically the plumbing of belief.

01

The bottleneck was never intelligence

Here is a claim I want to defend carefully, because it's the load-bearing one: for most of the history of organized inquiry, the limiting factor has not been how smart anyone was, but how much epistemic *work*, generating hypotheses, running tests, checking each other's tests, updating the results, a fixed number of human minds could do per unit time.

A researcher today has access to more accumulated knowledge, better statistics, better instruments, and a faster literature search than anyone in

1950, and yet the replication crisis is worse, not better, than it was in 1950, because the crisis was never about individual competence. It was about throughput. More claims are being produced than there are careful eyes to check them, more grant proposals than qualified reviewers can read closely, and more papers published per year than any human being can track even within a single subfield. Peer review, in practice, is two or three tired experts spending a few unpaid hours on a document that took a team six months to produce, in a system where saying "I didn't have time to check this carefully" is never an option on the form.

This isn't a story about bad people; it's a story about an activity, checking, replicating, cross-examining, synthesizing, whose supply has been essentially fixed by the number of trained experts alive at any given time, while the demand for that activity has grown with every dataset, every subfield, and every institution that needs to decide uncertainty. Science, forecasting, policy analysis, legal reasoning, and monetary policy are all activities that consume the same scarce resource: careful attention applied to a specific, checkable claim, and for the entire history of these institutions, that resource has scaled at the rate of population and training pipelines, which is to say slowly, linearly, and with enormous friction.

What if it didn't have to?

02

What actually changes, mechanically

I want to be specific about the mechanism, because "AI will help science" is a sentence that means almost nothing until you say what activity gets faster and why.

The activity that gets faster is *adjudication*, the process of taking a claim, breaking it into checkable sub-claims, and determining which ones hold up. This is different from discovery. Discovery, the flash of insight that produces a new hypothesis, may or may not be something AI systems

are good at, and that's a separate and more contested question. Adjudication is more mechanical, more legible, and more clearly a place where a system that can read a hundred papers in the time it takes a human to read one, hold the entire relevant literature in working memory simultaneously, and apply a consistent standard of scrutiny across every single claim it checks produces a straightforward multiplier on throughput.

Consider what a single competent human reviewer actually does when carefully reviewing a paper. They check whether the stated methodology matches what's in the supplementary materials, whether the statistics are computed correctly, whether the cited prior work is characterized accurately, whether the claimed effect size is consistent with the sample size, and whether an alternative explanation was considered and ruled out. Each of these is a bounded, checkable task, and none of them requires genius, but all of them require time, patience, and the discipline not to skim past something because you're on your fourth review of the week and the conference deadline is tomorrow.

A system that never skims, never gets tired, and never has a fourth review that week, applied not to one paper but to the cross product of a claim against the entire literature it should be consistent with, doesn't just do this faster; it does it differently in kind, because it removes the sampling problem. Right now, we catch the errors that happen to fall in front of a reviewer who happens to have the specific expertise to notice them, on the specific day they have the bandwidth to look closely, and almost everything else goes through uncaught, not because it's fine, but because nobody had the hours.

03

Institutions as epistemic machines

Step back further. An institution, whether it's a peer review system, a court, a central bank's forecasting committee, or a grant panel, is best un-

derstood as a machine for converting distributed, noisy, partially-motivated human judgment into a collective decision that's more reliable than any individual's. That's the whole design goal. Committees, juries, review boards, and forecasting tournaments all exist because we learned, empirically, that aggregating many perspectives with the right structure beats trusting one smart person's gut, but that aggregation is expensive, slow, and vulnerable to whoever has the most time, the loudest voice, or the most institutional power in the room.

What I think is coming is not "AI replaces the committee." It's AI as a new kind of *component* inside these machines, one that can do the parts of the aggregation problem that are about coverage and consistency, freeing the human components to do the parts that are about judgment and values. A concrete way to say this is that an institution's job is to combine three things: generating candidate claims or decisions, checking them against the largest possible amount of relevant prior evidence, and exercising judgment about what to do given irreducible uncertainty and competing values. AI is extremely good at the second of these; it's not obviously a good substitute for the third, and it's a mixed bag at the first. The institutions that get rebuilt well over the next decade will be the ones that figure out how to route the checking work, the exhausting, uncelebrated, and currently under-resourced work of checking things against everything, to systems built for exactly that, while keeping the judgment work squarely with people who are accountable for it.

This is the actual shape of the optimistic case, and I want to walk through what it looks like in four places I think about a lot: replication and peer review, forecasting, resource allocation, and adjudicative institutions like courts and regulators.

04

Replication as a continuous background process

Right now, replication is an afterthought, something a handful of dedicated researchers do as a side project, usually years after a finding has already shaped a field and sometimes after it's shaped policy. The reason it's an afterthought is that replication is expensive relative to its status: it produces no novel result, gets no prestige, and consumes the same scarce resource, trained expert time, as original research.

Imagine instead a norm, enforced by tooling rather than willpower, where every empirical claim submitted to a journal is automatically checked against the full text and data of every paper it cites in support of its claims, every subsequent paper that has tried to build on or replicate it, and a battery of statistical sanity checks appropriate to its methodology, things like power analysis, robustness to alternative specifications, and whether the effect size is plausible given the measurement noise. None of this requires resolving open scientific questions; it requires reading carefully and consistently at a volume no group of humans can sustain. A claim that survives this gauntlet isn't proven, but it's been checked in a way that almost nothing published today actually has been, and a claim that doesn't survive it gets flagged before it becomes the premise of someone's dissertation five years later.

The failure mode to be honest about is that this could easily become a bureaucratic tax that slows science down instead of speeding it up, if it's implemented as a gate that has to be satisfied rather than a service that helps. The version worth wanting is one where the check happens automatically, cheaply, and continuously in the background, surfacing concerns for a human to weigh rather than issuing verdicts. The version worth fearing is one where "the model flagged it" becomes an unaccountable veto that nobody can argue with, because the whole point was supposed to be restoring the ability to argue with claims on the merits.

05

Forecasting as institutional infrastructure, not special-occasion theater

I spend a lot of my own time thinking about forecasting, about markets and models that try to put a number on uncertain future events, and the thing that strikes me most is how rarely institutions actually use calibrated forecasting as an ongoing input to decisions, versus how often they use it as a one-off exercise for a specific occasion. The Good Judgment Project[2], prediction markets, and superforecasting tournaments have all demonstrated, repeatedly, that structured aggregation of many independent judgments, scored against actual outcomes over time, produces better-calibrated probability estimates than expert committees issuing point predictions with no feedback loop, and yet almost no institution, not central banks, not intelligence agencies, and not most companies, has actually restructured its decision process around this finding. Forecasting tournaments remain a research curiosity that gets cited approvingly and then ignored operationally.

Part of why is cost. Running a good forecasting tournament, recruiting and vetting participants, scoring them over time, and building the infrastructure to actually feed those forecasts into decisions is expensive and slow relative to just asking the three most senior people in the room what they think. What changes when AI systems can generate, track, and continuously update calibrated forecasts on a huge number of questions simultaneously and cheaply is that the cost of running this kind of process drops close to zero. An institution, whether it's a company deciding whether to enter a market, a government agency weighing a policy change, or a fund evaluating a macro thesis, could have a live, continuously-scored set of probability estimates on the outcomes that matter to it, checked against its own track record over time, instead of a single point estimate from whoever spoke last in the meeting.

This is not a claim that AI forecasts will be *right*. Calibration is not the same as accuracy, and a system can be well-calibrated and still tell you that an outcome is 60% likely when the honest answer, given genuinely irreducible uncertainty, is that nobody can know. The value isn't oracular; it's that the process becomes auditable and improvable in a way that "ask the smartest person in the room" never was, because you can actually check, over hundreds of resolved questions, whether the forecasting process is well-calibrated, and fix it if it isn't. That's a fundamentally different epistemic object than institutional intuition, even when institutional intuition happens to be right.

06

Resource allocation without the queue

Grant panels, hiring committees, and editorial boards are, structurally, rationing devices, mechanisms for allocating a scarce resource, funding, positions, or publication slots, using a scarcer one, expert reviewer time. The queue length is not a feature of how important the decisions are; it's a symptom of a fixed supply of qualified evaluators applied to a growing supply of proposals. Grant success rates at major funding agencies have fallen for decades, not primarily because the average proposal got worse, but because the number of proposals grew faster than the number of qualified reviewers, so panels compensate by reviewing more shallowly or by defaulting to safer, more legible bets that are easier to evaluate under time pressure, which is one of the more boring and structural explanations for why funding panels are often accused of being conservative and risk-averse. It's not only that reviewers lack imagination; it's that genuinely novel proposals are more expensive to evaluate carefully, and careful evaluation is the resource in shortest supply.

If the cost of a thorough, well-informed first-pass review, checking a proposal's claims against the literature, checking whether the proposed

methodology is sound, and checking whether the budget is realistic given comparable projects, drops enormously, panels can spend their now-freed human attention on the part of the job that actually requires human judgment, deciding what's worth funding given values and priorities that are legitimately contested rather than verifiable facts about whether the proposal is competently constructed. The scarce resource stops being "can anyone check this competently" and starts being "do we, as an institution with particular goals, want to bet on this," which is exactly the question a funding panel should be spending its human hours on and currently mostly isn't, because it's drowning in the other one.

07

Adjudicative institutions: courts, regulators, and the standard of scrutiny

I'll go somewhere a little more contested. Courts and regulatory bodies are also, at bottom, machines for adjudicating claims against evidence and precedent, and they suffer from the same throughput bottleneck. A regulatory agency evaluating whether a company's safety claims match its internal testing data, or a court trying to establish whether a pattern of conduct is consistent with a legal standard, is doing the same kind of work as a peer reviewer: checking a claim against a large body of relevant material, under time pressure, with a caseload that outstrips the available expert hours.

The optimistic version here is that discovery, cross-referencing, and precedent-matching, the parts of legal and regulatory work that are about coverage of a large evidentiary record, get faster and more thorough, while judgment about culpability, intent, proportionality, and the application of values to hard cases stays exactly where it is now, with judges, juries, and accountable regulators. I want to be honest that this is the place where I'm least confident the boundary holds, because legal reason-

ing is unusually vulnerable to the temptation to let "the system said" substitute for "a person decided," since it comes with the appearance of objectivity that makes it easier for humans to defer to rather than own. The right design principle, one I don't think we've solved, is that these systems should make the evidentiary record more thoroughly examined rather than make the decision, and institutions need to build in friction that keeps that boundary from eroding under the ordinary human pressure to offload responsibility onto whatever looks most authoritative.

08

A market for adjudication

There's a version of this that I think is underrated even among people who buy the general argument: markets are themselves a form of distributed adjudication, and prediction markets in particular are a case where the "checking" function and the institution are the same thing. A prediction market doesn't defer to an expert panel's judgment about whether a claim is likely true; it lets anyone with money and a disagreement bid against the current consensus, and the price becomes the aggregated verdict. This is, in miniature, exactly the epistemic machine I've been describing, a mechanism for converting distributed, partially-informed, partially-motivated judgment into a single number that's more reliable than any one participant's opinion, because it's continuously checked against real stakes and real outcomes.

What's interesting is watching what happens when you feed AI-generated analysis into that kind of market rather than into a committee. A committee has to decide, as a group, how much weight to give the AI's output, which reintroduces exactly the bottleneck: limited expert hours to evaluate a new kind of input that started this whole essay. A market doesn't have that problem, because it doesn't need consensus about how much to trust the AI; it just needs some participants who think the AI is

right and are willing to bet on it, and some who think it's wrong and are willing to bet against it, and the price sorts it out. Markets are, in a sense, the institutional form best suited to absorbing a new, unproven information source without requiring anyone to first agree on how much to trust it, and that's a structural advantage worth taking seriously as more of these institutions get designed, not as a replacement for peer review or grant panels, but as a template for how you let a genuinely new kind of evidence earn trust gradually, through track record, rather than needing it granted by fiat on day one.

I'd also flag the limits of this cheerfully, since it's a case I'm predisposed to like. Markets aggregate incentives, not truth, and a market can be perfectly liquid and still systematically wrong if the underlying question resolves ambiguously, if manipulation is cheap relative to the stakes, or if the pool of participants is small enough that a few large actors can move the price without actually knowing more than anyone else. Prediction markets are a powerful adjudication mechanism for questions with clear, checkable resolution criteria and enough participants to make manipulation expensive, but they are not a general-purpose substitute for expert judgment on questions that are inherently ambiguous, slow to resolve, or contested on values rather than facts, which is most of what courts and grant panels actually do. The lesson isn't "replace committees with markets." It's that the space of institutional designs for adjudication is larger than "expert panel" or nothing, and AI lowers the cost of running more of these designs in parallel to see which one actually earns trust on which kind of question.

09

What builders should actually do differently

If the throughput story is right, it implies a specific and somewhat unglamorous research agenda, and I think it's worth naming it plainly rather

than leaving it implicit. The interesting work isn't making a bigger, smarter model and hoping institutional adoption follows; it's building the specific connective tissue. That means interfaces that let a checking system surface concerns to a human in a form that preserves the human's ability to disagree and overrule, rather than a form that just outputs a verdict. It means scoring infrastructure that tracks a forecasting or checking system's track record over time, in public, so that trust is earned through resolved predictions rather than assumed from capability benchmarks. And it means adversarial red-teaming built into the checking layer itself, so that gaming the checker is treated as a known threat model from day one rather than a scandal discovered five years in. None of this is the part of AI development that gets keynote slides, but it's also, I'd bet, the part that determines whether any of the optimistic scenarios above actually happens, versus becoming another case study in a promising technology captured by whoever adopted it first and used it to entrench an existing advantage.

10

The compressed decade

Dario's essay uses the frame of compressing a century of scientific progress into a decade. I want to borrow the frame but point it somewhere slightly different: not a compressed century of discoveries but a compressed century of *institutional maturation*, the slow process by which fields learn, the hard way, which of their practices are actually reliable.

It took science roughly a century to go from "publish and mostly trust" to the modern apparatus of preregistration, replication requirements, statistical review, and open data, and that apparatus is still incomplete and unevenly adopted across fields. It took forecasting research about seventy years, from operations research in World War II to the Good Judgment Project, to establish rigorously that calibrated aggregation beats ex-

pert intuition, and another decade-plus to barely begin operationalizing that finding anywhere outside of academic exercises. Institutions learn slowly because the feedback loop between "we used this process" and "here's whether the process actually worked" is long, expensive, and usually never closed at all, since most institutional decisions are never scored against their outcomes in any rigorous way.

If AI systems can make the checking and scoring loop cheap enough to run continuously and automatically, with every claim checked, every forecast scored, and every decision process eventually auditable against how it actually performed, then the thing that compresses isn't the rate of discovery; it's the rate at which institutions can learn what's working and fix what isn't. That's a more modest claim than "AI cures diseases in a decade," but I think it might matter just as much, because it's upstream of almost everything else. A civilization that gets meaningfully better and faster at knowing which of its institutions' practices actually work is a civilization that compounds every other kind of progress faster too.

11

A concrete morning

Abstractions like "throughput" and "adjudication" are easy to nod along to and hard to actually picture, so let me try to make one version of this tangible.

A funding panel is reviewing forty proposals for a research grant cycle. Today, each proposal gets one primary reviewer and one secondary reviewer, both volunteering their time on top of comprehensive research and teaching loads, each spending perhaps three or four hours per proposal if they're conscientious and closer to one if the week has been hard. The reviewers catch the errors they happen to have the specific background to catch, so a statistician on the panel notices a power calculation that doesn't add up, but nobody else would have. A methodology that re-

sembles a discredited approach from an adjacent subfield sails through because none of the three reviewers happened to work in that subfield, and the panel meets for a single afternoon, argues about the five proposals everyone has opinions on, and defers to whoever wrote the most confident-sounding summary for the other thirty-five.

Now picture the same forty proposals arriving alongside a report, generated overnight, for each one. Every statistical claim gets checked against the raw numbers in the appendix, every cited prior result gets checked against what that prior paper actually says and whether it's been contested since, every methodology gets checked against known failure modes from adjacent fields, and every budget line gets checked against comparable funded projects for plausibility. It's a checklist of specific, cited concerns and specific, cited strengths, attached to each proposal and available to every reviewer before they've read a single page. The three human reviewers on each proposal now spend their four hours doing something different, not hunting for errors they might not have the background to notice, but weighing the concerns that have already been surfaced, deciding which of them actually matter, and doing the part of the job that was always supposed to be the point: judging whether this specific bet, with its specific risks made visible, is one this institution wants to make. The panel's afternoon meeting spends less time discovering that a budget number looks off and more time arguing about whether a risky, exciting proposal is worth funding given everything now known about it. That's not a faster version of the same meeting; it's a meeting about a different, more consequential question, because the more mechanical one got handled before anyone sat down.

Nothing about this scenario requires the checking system to have good scientific taste, to know what's exciting, or to make the actual funding decision. It requires it to be thorough, consistent, and tireless at a task that's fundamentally about coverage, and to hand its output to humans in a

form that helps them argue rather than a form that tells them what to conclude.

12

What could go wrong, said plainly

I don't want this essay to be the genre where the caveats section is three throat-clearing sentences before returning to unqualified enthusiasm, so let me actually sit with the failure modes, because I think they're serious and specific rather than generic AI-doom gestures.

HOMOGENIZATION OF JUDGMENT. If every institution's checking process routes through a small number of underlying models, an error or a blind spot in those models stops being one reviewer's mistake and becomes every institution's simultaneous, correlated mistake. Human committees are diverse and slow partly because diversity and slowness decorrelate errors, since three people are unlikely to all miss the same thing for the same reason. A world where the checking layer is centralized needs deliberate architectural diversity, meaning different models, different methodologies, and adversarial cross-checking, or it trades a slow but decorrelated error rate for a fast but correlated one, which is a worse failure mode, not a better one.

METRIC CAPTURE. The moment "passed the AI check" becomes a legible signal that confers status or funding; people will optimize for passing the check rather than for being right, exactly as happened with citation counts, h-indices, and every other proxy metric science has ever adopted. A system built to catch bad claims will, given enough time and enough incentive to game it, start generating claims specifically engineered not to trigger it. This isn't a hypothetical; it's the empirical track record of every scientometric proxy in history, and there's no reason to expect this one to

be different unless it's designed from the start assuming adversarial pressure.

RESPONSIBILITY LAUNDERING. This is the single failure mode I'm most worried about, mentioned above in the legal context but true everywhere. A finding, a decision, or a verdict that came from "the model said so" is psychologically easier for a human to defer to than one a colleague argued for, because it feels less like a choice someone is accountable for. Institutions exist partly to hold identifiable people responsible for judgment calls, and any design that lets a human say "the AI flagged it" instead of "I decided, based on what the AI surfaced, that this was a problem" has quietly moved accountability somewhere it can't actually be enforced, onto a system that can't be fired, sued, or voted out. This is a design failure rather than an inevitability, but it's the one I'd bet is hardest to avoid in practice, because it's so much more comfortable than the alternative.

WIDENING RATHER THAN CLOSING GAPS. Every institution described above is already unevenly resourced, and elite universities, well-funded agencies, and large firms will adopt these tools first and best, so the throughput gains could easily widen the gap between institutions that can afford to rebuild their processes and ones that can't, rather than democratizing careful checking the way the optimistic story implies. The technology doesn't have a built-in tendency toward equitable diffusion; that has to be built deliberately, and historically it usually isn't.

None of these are reasons to expect the whole project to fail. They're reasons the good outcome has to be built on purpose, with these specific failure modes in mind, rather than assumed to fall out automatically from making the underlying models more capable.

13

Why write this down now

I think the reason nobody writes the "Machines of Loving Grace for institutions" essay is that it's genuinely less thrilling than curing cancer or ending poverty, since checking claims against literature more thoroughly is not a sentence that produces goosebumps. But I keep coming back to the fact that almost every other kind of progress people are excited about new drugs, better policy, smarter markets, more just legal outcomes, is bottlenecked at some point in its pipeline by exactly the kind of throughput-limited, expert-attention-constrained checking process I've described here. The institutions are the load-bearing wall. If you rebuild the wall well, you get to keep building on top of it faster than before, but if you rebuild it badly, with correlated errors, gamed metrics, laundered accountability, and gains captured only by whoever already had resources, you've built a taller structure on a wall that was already the weak point, and it will show.

I don't think this happens automatically, and I don't think it happens without people who actually understand both the mechanics of these institutions and the mechanics of the systems being built to augment them sitting in the room when the design decisions get made. That's most of why I'm writing this down, not as a prediction but as an argument for where the effort should go, and a bet that the century ahead of us will be defined less by any single discovery than by whether we got the plumbing right.

There's a version of the AI story that treats institutions as the thing being disrupted, legacy structures that a sufficiently capable model will simply route around, the same way a search engine routed around the reference librarian. I don't think that's right, and I think believing it is a mistake with real consequences. Courts, funding agencies, regulatory bodies, and peer review exist because someone, at some point, needs to be ac-

countable for a judgment call, and no model, however capable, changes the fact that accountability has to attach to an entity that can be questioned, overruled, and held responsible in ways that make sense to the humans who live with the consequences. The institutions aren't the obstacle to a faster epistemic future; they're the thing that has to survive it, still recognizable and still trusted, on the other side, and whether they do depends entirely on the specific, unglamorous design choices made now rather than on how good the underlying models eventually get.

That's the actual bet underneath this essay. It isn't that intelligence is about to get radically cheaper, since that part is close to consensus at this point, whatever else people disagree about. The bet is that the institutions built to convert cheap intelligence into trustworthy collective belief are the part nobody is racing to build, precisely because they're less exciting than the models themselves, and that this is exactly backwards. A civilization with brilliant models and unreformed institutions for checking, trusting, and acting on what those models produce doesn't get the compressed century; it gets a much faster rate of confident, unaudited error, which is a worse outcome than the slow, careful, bottlenecked status quo we're trying to improve on. The plumbing is the whole point, and I'd rather spend the next decade on it than assume it fixes itself.

NOTES

1. Dario Amodei, "Machines of Loving Grace" (2024). darioamodei.com/essay/machines-of-loving-grace
2. Barbara Mellers et al., "Psychological Strategies for Winning a Geopolitical Forecasting Tournament," *Psychological Science* (2014). doi.org/10.1177/0956797614524255



The Future Belongs to Those Who Understand It

The Epistemic Century is a private investment and research firm founded by Arya Somu in November 2026.

Originally published October 11, 2026, at aryasomu.com/writing/the-epistemic-century.

© 2026 Arya Somu. Set in Crimson Pro and Inter.