

The Architecture of Truth

What Makes a Claim Load-Bearing

James Maconochie

Technology Leader | BS Civil Engineering, Imperial College London '93 | MS Civil Engineering, MIT '94

October 2026 (v1.0)

Abstract

Two people read the same report and disagree. One concludes there is no fact of the matter, only competing perspectives. The other runs a careful analysis and issues a confident verdict on a question the evidence was never able to settle.

These look like opposite failures. They are the same one: a claim that has come loose from whatever was holding it up.

This paper argues that a claim worth standing under is not something a person possesses but something an architecture produces. By architecture I mean everything standing behind a claim: the procedures that test it, the institutions that run those procedures, and the people those institutions have formed. Procedure is part of it. The part procedure cannot do is where this paper spends most of its time.

Because the parts can be named, the failures can be diagnosed rather than merely deplored. Four checks decide whether a claim is eligible: whether the thing it names as the cause really is the cause, whether people who had every reason to disagree agree anyway, whether anyone with the skill and the motive has tried to break it, and whether it survives time.

The central claim of this paper is that these four checks are necessary but not enough. Passing them produces a candidate, not a conclusion. Somebody still has to recognise that a claim which has passed all four is strong enough for what it is about to be asked to carry, and no procedure performs that step.

Three institutional cases test the argument. Air-crash investigation shows the architecture working where little is contested. The Volcker disinflation shows it working under severe political pressure. The Federal Reserve in 2006 shows it substantially intact, staffed by capable people, with the information available, and the judgement not made. That third case is what the argument rests on. If sound architecture guaranteed sound judgement, there would be nothing here worth saying.

Two further arguments follow. Prediction markets, crowdsourcing and algorithmic ranking can feed the checks but cannot perform the judgement. And AI is changing the route by which almost everything now reaches a reader. The old route had people in it: reporters, editors, peer reviewers, named commentators. The new one runs through systems that assemble an answer out of text rather than by going and looking, and rules written for the old route will not catch what goes wrong in the new one.

This paper diagnoses the problem. It does not prescribe the cure. It proposes no institutions and no legislation; it sets out what any of them would have to respect.

1. The question behind the trilogy

Three papers preceded this one, and reading them back reveals something none of them named directly. *The Attention Crisis* diagnosed the depletion of a resource. *The Wisdom Gap* argued that current AI architectures face a structural limit on what they can become. *AHI* proposed the constructive alternative.

Each was built around something none of the three stopped to defend. Attention was being defended *for* something. Wisdom was the mature form *of* something. AHI was an architecture *toward* something.

That something was truth. Not truth as a possession arrived at by procedure, and not truth as the output of careful reasoning correctly executed. Truth as what a working set of conditions produces: recognisable when the conditions hold, and reliably absent when they do not.

The claim

Truth is not a property a person has. It is a practice an architecture sustains. That architecture has constraints; claims hold within them, and degrade predictably as they are violated.

A word about *architecture*, since it carries most of the weight in this paper. It means the arrangements that surround a claim and act on it, together with the people those arrangements form: who is permitted to test it, with what access, against whose objection, over how long, with what consequence for being wrong, and who ends up in a position to say that it is strong enough to act on. Not a building, and not a metaphor for one. Section 7 works through three real ones.

It is not a procedure that delivers truth on demand. It is the set of conditions under which a candidate claim can be tested and, if it survives, recognised as strong enough to build on. That recognition is a separate act,

and no amount of procedure produces it. Section 6 develops the distinction; the sections before it earn the ground it stands on.

One narrowing, since the claim above could be read as a much larger one. This is not a theory of what truth *is*. It is an account of the conditions under which claims can be made, tested, and relied upon by people who did not make them. Whether truth in some fuller sense exists beyond what such architectures produce is a question this paper leaves alone.

An old version of the problem

There is a line widely attributed to Marcus Aurelius: everything we hear is an opinion, not a fact; everything we see is a perspective, not the truth. The wording is a modern condensation rather than anything in the *Meditations*, but the thought underneath it is his, and it is sharper than the paraphrase. In Book 8 he writes:

"Do not say more to yourself than the first impressions report. You have been told that some one speaks evil of you. This is what you have been told; you have not been told that you are injured... abide always by the first impressions and add nothing of your own from within."

That is the failure this paper is about, stated eighteen centuries early. Something arrives. We add to it (a cause, a motive, a story) and then treat the addition as though it arrived with the observation.

What follows from this is not what it first appears. If every impression is unreliable, one might conclude that nothing can be known and every view is as good as any other. That is the first of the two failures above, and the Stoics were not its authors. That was the Academic Skeptics, the rival school they spent centuries arguing against. Their position was the opposite, and more demanding. An impression arrives unvetted, and assent is a separate act that can be withheld. The disciplined person does not automatically believe what turns up in front of them.

That is this paper's structure, two thousand years early. The impression is the candidate claim. Assent is the judgement.

The fact that everything reaches us from somewhere, through someone, is not the conclusion. It is the problem the architecture exists to solve. And a well-built architecture can produce claims more reliable than anything that went into it. That is not mysterious. It is how a bridge comes to be stronger than any one of its members: redundancy, testing under load, and time.

What this paper opens

The trilogy diagnosed a problem and proposed an alternative. The governance paper specified what such an architecture must do at the

institutional level. This paper specifies what the architecture is *for*: what it produces when it works, and what it stops producing when it does not.

Ten sections follow. Sections 2 and 3 set out the working terms and how each becomes the next. Sections 4 and 5 cover where claims legitimately hold and where the architecture fails. Section 6 is the core: the four checks, what they leave undone, and two worked cases. Section 7 asks where such architectures come from. Section 8 examines what can and cannot stand in for them. Sections 9 and 10 draw the implications and close on the practice.

2. Five words the argument depends on

A commentator says: "*The fact is, the country is divided.*"

Three different kinds of claim are packed into that sentence and only one of them is a fact, arguably none. Pulling them apart is what these five words are for. They are working definitions. They earn their place by what they let the argument do, not by settling older debates about what truth is.

2.1 Fact

A Fact is something you could have watched happen. A cup falls off a table. A person drinks tea. A thermometer reads 38.2 degrees. That is the whole of it.

This is much narrower than everyday usage, deliberately so. Merriam-Webster gives two senses among others: something that actually exists or occurs, and a piece of information presented as true. This paper uses only the first. So when someone says "the fact is that climate change is accelerating," they are not misusing the word. They are using its second sense. But what follows is not a Fact as this paper means it. Nobody watched climate change accelerate. It is a conclusion drawn from a great many observations, which is the next term wearing this one's clothes.

Worth saying once and plainly: almost nothing argued about in public is a Fact in the strict sense. Most of it is the next term, or one of the two after that.

2.2 Reality

Reality is what the Facts add up to: the causal structure underneath that explains why they keep happening the way they do. The cup's fall is a Fact; gravity is Reality. The patient's fever is a Fact; the way the virus damages tissue is Reality.

The crucial property is that Reality is never observed. It is worked out. Nobody has ever seen gravity. We infer it from a very large number of

things falling, and we test the inference by dropping different things in different places and checking that the account still holds.

Which is why a claim about Reality can be wrong in a way that a report of a Fact cannot. The Fact is what happened. The claim about Reality is an inference drawn from it, and inferences fail.

A warning about this word, and a larger one than Fact needed. Merriam-Webster gives reality as the quality or state of being real, and as the totality of real things and events. Oxford's learner's dictionary gives it as the true situation as against how you would like life to be, and as a thing that is actually experienced or seen. None of them means the causal structure underneath, and that last one runs almost directly against the use here, since Reality in this paper is precisely the part nobody sees. The word is kept because gravity really is part of what is actually the case, and that is where the argument needs it to sit. But the narrower sense has to be held on purpose.

2.3 The Mediated Picture

Almost everything you know about the world, you did not see. Something stood between you and it. What you have is a picture of it, and pictures differ enormously in how much of the original survived the trip.

A Mediated Picture is what reaches you when the go-between is short and its faults can be named: an operation recorded on video, a pressure reading from an instrument calibrated last week, a reporter you trust filing from a place you cannot go.

None of these is the thing itself. Each is a window onto it. But the window is clean enough to act on, and, more importantly, when it fails you can say how. The video could be doctored. The instrument could be out of calibration. The reporter could be mistaken, or lying. Each of those is a specific accusation that can be checked and usually fixed. Saying a picture is imperfect costs nothing. Being able to name exactly how it might be wrong is what makes it usable.

2.4 The Constructed Picture

Now lengthen the chain.

A story assembled from a dozen sources. Raw material shaped into narrative, because raw material is unreadable and narrative travels. A frame that decides what goes in and what stays out. An organisation whose incentives reward one kind of finding over another. By the time an event reaches you as an article, it has passed through all four, and each has scraped something off the connection to what happened and added something of its own.

This is the water we swim in. Nearly everything a voter, a professional, or a policymaker acts on is a Constructed Picture.

And the dangerous property is that it does not announce itself. It arrives wearing the surface markers of first-hand knowledge (names, dates, quoted figures, the cadence of someone who was there) and the reader files it as Reality. That confusion is not occasional. It is where most of the trouble lives.

2.5 The load-bearing claim

The first four words name kinds of thing. The fifth names something a claim earns.

A load-bearing claim is one you would stand underneath. Not one you find interesting, or believe, or would defend at dinner: one you are willing to build on, knowing what happens if it gives way. Most claims never need to be that. They circulate, they inform, they get repeated, and nothing rests on them.

A claim earns it by passing four checks and then being judged strong enough by someone who has understood what it is about to be asked to carry. Section 6 does that work. For now: this is not a fifth shelf on the same bookcase. It is what the architecture, working properly, produces.

3. How each level becomes the next

Four transformations connect the five terms, and they have different shapes. The differences decide what the architecture has to do at each level.

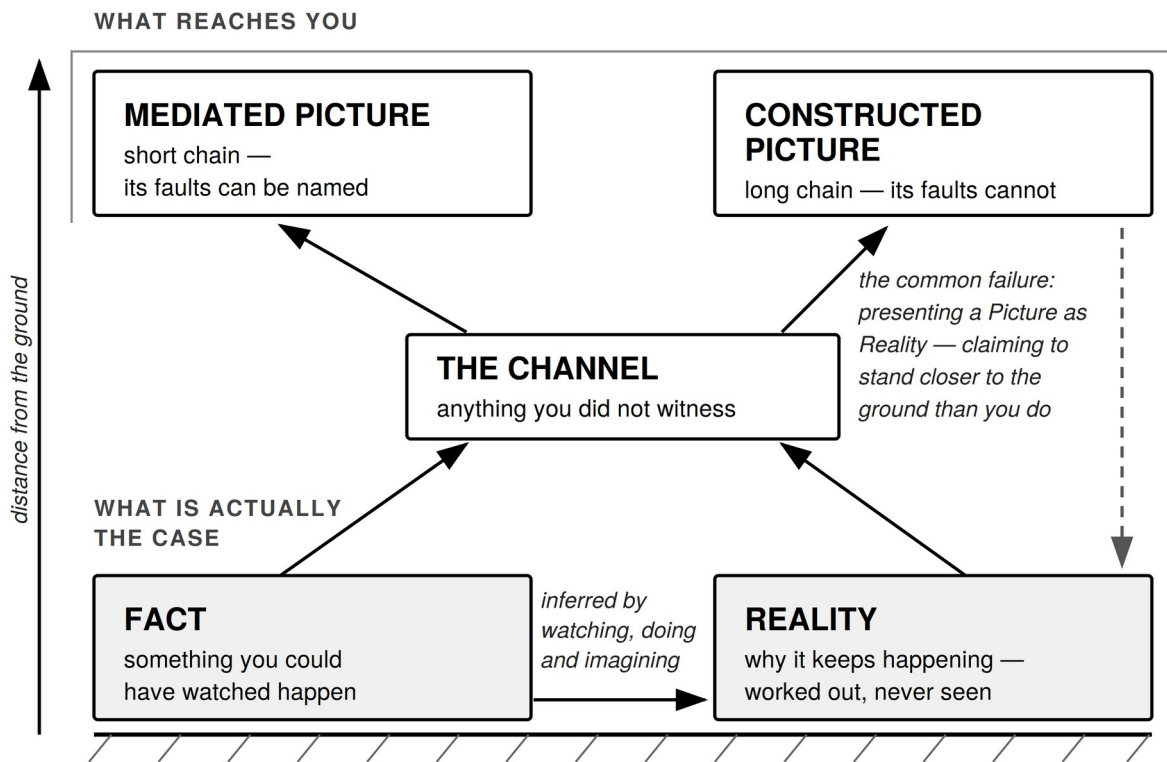


Figure 1. Facts and Reality are what is actually the case. A Picture is what reaches you after it has travelled: further from the ground, not higher up. Both kinds of Picture carry Facts and Reality together; they differ in how long the chain is, and in whether its faults can be named. Either kind then goes through the four checks and the judgement, which Section 6 develops. Section 5.2 takes up one case where this picture does not hold, because the chain does not start at the ground.

3.1 From Fact to Reality: the hardest move

Judea Pearl describes three kinds of reasoning, which he calls a ladder: **watching, doing, and imagining.**

Watching is a Fact: the cup fell. Doing is intervention: push a different object off the same table and see whether it falls the same way. Imagining is counterfactual: what would have happened with no table there; what would happen on the surface of the moon.

Gravity is not something anyone sees. You get to gravity by doing the second and the third together: testing an account of why things fall, and asking what would happen if the circumstances were different. Watching alone never gets you there.

Getting there is the whole of the move from Fact to Reality, and it is the foundational one that everything else rests on. Any Picture you are handed

assumes someone already made it. A load-bearing claim assumes it twice over.

What matters is what fills the gap when nobody makes the climb. It isn't nothing, and it isn't a shrug. It is stories: each one coherent, each one explaining the Facts, none of them requiring anybody to test a cause. And a story usually gets there first, because a story is cheap and a causal account is expensive.

People in that position are not short of Facts. They have plenty of Facts. What they have lost is the step between the Facts and the story.

3.2 From Reality to a Picture; a fork, not a step

The second transformation runs sideways.

Facts and Reality both have to reach everyone who was not present when the Facts happened. They reach them through a channel, and what comes out the far end is a picture. The channel can be short and inspectable, or long and not. Same original; two channels; one you can check and one you mostly can't.

The two Pictures therefore sit at the same level, as Figure 1 shows. They are a fork, not a step: introducing them in that order tracks increasing degradation and nothing more.

This matters for what the four checks will be asked to do:

- A Mediated Picture is **not** automatically sound. Video gets doctored; instruments drift; a single honest observer can be wrong.
- A Constructed Picture is **not** automatically unsound. A well-sourced, seriously-attacked piece of journalism can pass every check; it just carries a heavier burden, because more was done to it upstream.

The architecture's job here is not to rule the Constructed Picture out. It is to make sure the channel is honestly named, so the checks run on a candidate that knows what it is.

3.3 From Picture to load-bearing, filter, then judgement

The third transformation has two stages, and they are not interchangeable.

First, the filter. Four checks: does the cause hold up, do independent observers agree, has anyone seriously tried to break it, does it survive time. Cause, independence, attack, time.

This part is a procedure. A claim clears all four or it does not. Failing any one returns it to the Picture it came from, where it can still be discussed, repeated, and believed. It is just not yet something to stand under.

Second, the judgement. Someone has to recognise that this cleared claim, in this case, for this decision, is strong enough to carry the weight about to be put on it. That part is not a procedure, and no amount of care turns it into one.

The filter is necessary and not sufficient. The judgement cannot be reduced. Both are required. Section 6 develops each, with worked cases.

3.4 What the hierarchy is, and is not

This is a hierarchy of dependence, not of worth. Facts are foundational and nothing stands without them. The Pictures are not failed attempts at something better; they are what everything else gets built from, and most useful claims in ordinary life live there quite happily without anyone ever needing to judge them load-bearing.

Which gives the rule the rest of the paper runs on: **degrade any level and the top degrades with it.** Lose your Facts and you lose your Reality, then your Pictures, then anything you could safely stand under, in that order. Lose the discipline to convert Facts into Reality and you arrive at the same place by a different road.

So when claims start failing, the diagnosis has to run at every level, not only at the top where the failure is loudest.

4. Where load-bearing claims actually work

People make claims worth standing under every day, in many fields, and they hold. This section takes three and asks what they have in common.

4.1 Three that work

A bridge engineer specifying steel. "This beam will carry this load" is a claim you can stand under, literally. It holds because a century of work stands behind it. Bars pulled apart in laboratories. Beams loaded until they failed, so somebody could write down where. Structures watched through wind, rain, snow, earthquake and traffic. The results were compared, argued over, and written into material standards. Those standards get reviewed by people whose professional reputation depends on catching errors in them. And the bridges keep score: one that stands for fifty years under loads nobody designed it for is evidence, and one that falls gets investigated until the codes change. The engineer's claim is not an opinion. It rests on all of that.

A well-run clinical trial. Patients are assigned to treatment or control, which is doing rather than watching: so the cause check can actually run. Repeat it at other sites with other teams and you have tested independence:

if the result holds, it no longer depends on one group's methods or blind spots. Peer review and regulatory review exist to break it, which is the attack check. And registry data on patients who took the drug for years, long after the original trial closed, settles what the trial itself could not. Where all four are present, "this drug treats this condition in this population" is a claim you can build on.

Mature physics makes the point in its sharpest form. Nobody has seen gravity. Special relativity has been standing for over a century, including against people who set out specifically to knock it down. Claims like that live inside an architecture that has done the work repeatedly, against opposition, over horizons longer than any physicist's career.

4.2 What the three have in common

Four things. Their absence is the warning sign.

You can intervene. Not just describe, act, and watch what follows. The engineer loads the beam. The trial assigns the patient. The physicist builds the experiment.

Failure is visible. A bridge that falls leaves wreckage. A drug that doesn't work shows up in patients. A theory that's wrong eventually predicts something the world declines to produce. Somebody is looking at the result.

Attack is routine, not optional. Other engineers review the design. Other labs attempt replication. Other physicists try to break the theory. It is built into how the field works, rather than left to individual conscience.

The time horizon outlives the practitioner. The codes the engineer works inside outlived the people who wrote them. Drug safety data accumulates for decades. Relativity has been under test longer than most physicists have been alive. The architecture is patient in a way no individual can be.

4.3 The same field, two different claims

The framework does not say medicine is a field where claims hold. It says some claims in medicine are held up and others are not, and that the dividing line sits inside medicine, not between medicine and everything else.

"This antibiotic kills this bacterium in this population" has everything behind it: trials, replication, review, long-term data.

"This treatment will work for this patient, with this particular combination of other conditions" is a different animal. There may be no trial that covered this combination. This patient's specific situation may not be observable.

And the thing that would tell you whether the call was right arrives long after the call has to be made.

That is not a flaw in medicine. It is a structural property of the field, and part of the practitioner's job is noticing which kind of claim they are making. The useful question is not *is this a medical claim* but *does the architecture here actually reach down to the level where I'm standing?*

Where somebody issues a confident verdict at a level the architecture does not reach, the failure is not sloppy thinking. It is careful thinking, carried past the point where anything is holding it up.

5. Where it fails

Four failure modes, structurally distinct. My reading is that public discourse is running all four at once, which would account for why claims feel so unstable, though whether all four are present everywhere is not something this paper measures. What matters is that they are separable, and separating them is what makes a diagnosis possible.

5.1 Claims held at the wrong level

The commonest failure, and the easiest to make.

A Reality claim gets presented as Fact. A Constructed Picture gets presented as Reality. The distinction between levels collapses, and the claim is held with more confidence than its actual level earns.

"The fact is that climate change is accelerating" is a Reality claim in a Fact's clothing. "The country is divided" is not Reality. It is polling, coverage, social media mood and editorial framing, condensed into a sentence built to sound like a description of how things are.

Either may be perfectly defensible at its own level. Climate change as Reality is well-supported. A divided country is a reasonable claim about the Constructed Picture, if you are honest about what went into it. The failure isn't the claim. It's the level it's being held at, and the confidence that travels with it.

This requires no institutional decay whatsoever. One careful person, working alone, on a good day, can do it. That is why it is the most common.

5.2 The channel is changing, and one new thing is not a channel at all

Something newer, and different in kind.

Someone using an AI to summarise the news, draft an analysis or research a question is not receiving a Mediated Picture. They are receiving a

Constructed one: but it arrives wearing Mediation's clothes: confident prose, specific names, plausible citations, the cadence of somebody who knows.

A newspaper article, however degraded, is a chain that starts at something that happened. Somebody was there, or spoke to somebody who was. The chain can be long and the links can be bad, but it runs back to the ground, and you can walk back down it: call the reporter, read the document they cited, find the person they quoted.

An AI summary is a chain that starts somewhere else. What went into the model was text, in enormous quantity, most of it Pictures that other people made. Some of that text carries the work those people did in getting from Facts to Reality. What comes out is text shaped by the patterns in that text. At no point did the process go and look.

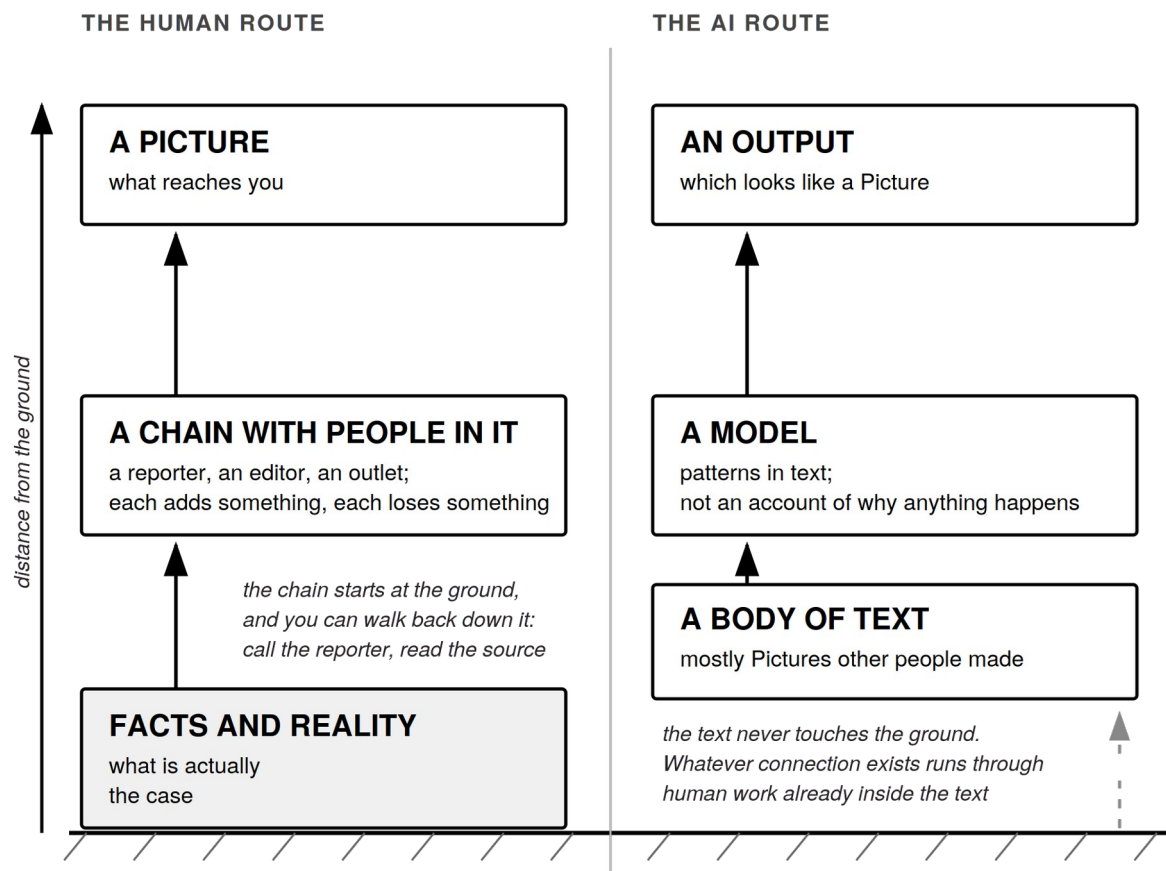


Figure 3. Two chains that arrive looking the same. The human chain starts at the ground and can be walked back down. The AI chain starts at a body of text that is already some distance from the ground, and the only connection to the ground runs through work other people did before the text was written.

So this is not a longer channel. It is a different kind of thing standing in the channel's position, and the difference has a precise consequence.

In the terms of 6.1.6, the cause check on an AI output is not failed. It cannot be run, and it is the third kind: nobody has obstructed it and no amount of waiting helps, because what the check needs to examine is not there. There is no intervention to inspect, no counterfactual anybody tested, no account of why that anybody built. Whatever causal work sits inside the answer was done by people whose names are not attached and whose reasoning cannot be recovered from the result.

This does not mean the framework has nothing to say about AI output, or that such output is beyond assessment. It means AI output is a particular kind of candidate: one where the first check cannot be run on the artifact itself, and where the only route to running it is to go behind the output, to whatever sources it drew on, and check those instead. That route exists. It is expensive, almost nobody takes it, and it is often not available at all, because the trail back is missing or invented.

Three further differences follow, and each of them compounds this one. It is worth being careful about how strong they are, because the obvious objection is fair. Human mediation was never transparent either. Editors made framing decisions nobody could inspect; outlets had incentives readers couldn't see. The difference isn't that one channel is clean and the other isn't. It is three specific things:

Scale. A single AI system now produces synthesis at a volume no newsroom could.

No named mediator. A newspaper article had people behind it, and their framing decisions were examinable in principle and contestable in practice, you could write to the editor, and occasionally it worked. An AI summary's framing decisions are diffuse, inherited from training data, weighted by processes the reader cannot inspect and often the writer cannot either. There is nobody to write to.

Uncertainty stops meaning anything. Older infrastructure preserved hedges, attributions, and the visible seams where one source ended and another began, and those markers tracked how sure the writer actually was. AI prose still carries hedges. Measured word by word it carries rather more of them than human writing does. What has broken is the link between the hedge and the state behind it, and Section 9.3 gives the measurement.

Together these shift the channel enough to matter. What reaches the reader reads as Mediated and behaves as Constructed. The part of a reader's judgement that would normally catch the substitution doesn't fire, because the surface signals are telling it not to.

None of this depends on settling what is happening inside the AI system, the machine as I usually call it. Whether it is reasoning or pattern-matching, whether the next model closes a gap or opens one, is a separate argument, and the case above does not need it settled. It needs only what is not in dispute: what went in was text, and the process did not go and look. That is a claim about where the chain starts, not about what the machine is doing once it has started.

5.3 The checks get weaker

The four checks still formally exist nearly everywhere. What erodes is the capacity to run them, and the erosion is uneven. Two of them, attack and independence, have solid evidence behind them. For the third, time, the evidence is thinner than the intuition. And the fourth, cause, is degrading in a way that looks nothing like the obvious one.

The attack check needs institutions with both the ability and the reason to attack.

In journalism the capacity has simply gone. American newsroom employment fell about a quarter between 2008 and 2020, and newspaper newsrooms specifically lost well over half their staff (Pew Research Center, on federal labour data). Roughly 3,500 newspapers have closed since 2005, about two in five (Medill Local News Initiative), leaving 213 counties with no local news source at all and more than 1,500 with only one.

In research the problem is different: not capacity lost, but volume outrunning it. Indexed research articles rose by nearly half between 2016 and 2022, faster than the supply of researchers available to review them (Hanson and colleagues, *Quantitative Science Studies*). And when a large collaboration tried to reproduce a hundred published psychology findings, roughly a third replicated (the Open Science Collaboration, *Science*, 2015).

One thing not to claim here: rising retraction numbers. They are ambiguous: more retractions may mean detection is improving rather than standards falling, and the replication crisis is itself an instance of adversarial testing working. The honest version is that the volume of claims has outgrown the capacity to test them, and that what testing happens is increasingly after the fact rather than before it.

The time check needs somewhere old claims get revisited: a record kept, and somebody consulting it.

Here the intuition is stronger than the evidence, and it is worth saying so plainly. What can be shown is that turnover in what holds public attention has accelerated: the average time a topic stayed among the most discussed on one large platform fell by about a third over three years. What cannot be shown, because nobody has measured it, is that this translates into faster

erosion of public memory for specific past claims. That is a plausible extrapolation, not a finding, and it is offered as one.

One result from the same work cuts the other way and should be said. There was no acceleration at all in scientific citation or in encyclopedia editing: the knowledge stores that are built to persist. That is either reassuring, or a sign that the record is still being kept carefully in places almost nobody consults. The argument does not need the stronger version. It needs only that the check requires an arena, and that an arena nobody consults does not function.

The independence check needs observers who can genuinely observe independently. The popular explanation for how it fails is the wrong one, and the argument is better for losing it.

The intuitive story is the echo chamber: feeds sorting people into groups that only hear themselves. The best available evidence does not support it. Large field experiments have cut people's exposure to like-minded content by a third for months at a time and found no measurable change in their attitudes. Feeds assembled by algorithm turn out to carry *more* cross-cutting material than people's own clicking does. And news diets reached through search and social platforms are more varied, not less, than those of people who go straight to a favourite outlet.

What does erode observer independence is duller, and much better documented: the same material appearing in more places.

When newspaper ownership consolidates, duplicated content rises sharply: across two million articles from 108 Swedish newspapers, duplication rose by more than half in relative terms after mergers (Garz and Ots, *Journal of Communication*). In American local television, a study of 861 stations by Yanich and Bagozzi found that in 39% of markets, pairs of nominally separate stations were broadcasting substantially the same newscast, averaging around 69% identical content and reaching some 41 million households.

That is this check's failure, counted. Three outlets carrying the same material are not three observers, and a reader has no way to tell from the outside.

A second route matters, and it connects back to the check before it. A randomised experiment on one large platform found that its ranking system demoted traditional media in favour of other content, shifting participants' views over several weeks. The failure there is not that observers converge on one another. It is that the layer deciding what they see weights *away from* the sources that run their own verification.

One caution about the consolidation evidence. The effects vary by owner rather than running uniformly in one direction: some chains cut local

coverage sharply, others increased it. So the defensible claim is not that consolidation makes all content the same. It is that consolidation makes content a function of who owns the outlet rather than of what is happening locally, which still reduces the number of genuinely independent judgements being made.

The cause check is degrading least visibly, and perhaps costing most. As the capacity for causal reasoning weakens, narrative consistency starts standing in for causal consistency. A story that hangs together gets mistaken for a story that tracks.

This is not a speculation about the general public. It happens inside expert institutions, under exactly the conditions where it should not. Here is a member of the Federal Open Market Committee, in the transcripts, on whether house prices were unsustainable:

"there are stories one can tell about fundamentals that would explain or rationalize housing prices. Obviously, low interest rates have to top the list... So from that point of view, it's hard for me to see how it would be reasonable to place a great deal of certainty on the notion that housing is significantly overvalued, or that there's a bubble, or that it's going to collapse."

Every element of that is defensible. What it is not is a test. A coherent account of why prices *could* be justified has been allowed to do the work that evidence about whether they *were* should have done. The check passed the claim, not because the causal chain held, but because the narrative one did.

5.4 Nobody is there to make the judgement

Even where the checks run properly, the practitioner who would make the judgement may be missing.

The Governance paper called this cognitive debt: the compounding cost of people who sign off on outputs rather than interrogating them. The Wisdom Gap paper called it the seed corn problem: the junior work that builds a practitioner gets eliminated, and twenty years later the senior practitioner it would have built isn't there.

In this paper's terms: an architecture can produce perfectly compliant, fully cleared claims and still fail to produce anything worth standing under. Recognising that this claim, in this case, is load-bearing for what is about to be done with it, and holding that judgement against everything pushing the other way, takes somebody who has been formed to do it. If they aren't there, nothing visibly breaks. The reports still get written. The outputs still appear.

What's missing is the judgement, and its absence does not show up in the document.

5.5 They compound

These four are not independent.

Wrong-level claims travel further through degraded channels. Degraded channels make weakened checks harder to spot, because the inputs no longer carry the markers that would have exposed how far they had travelled. Weakened checks make the missing judgement easier to hide, because there is no clean output for it to visibly fail to clear.

The result is not four problems. It is one composite failure in which each part covers for the others.

Which is where the diagnostic question earns its keep: which mode is driving this particular failure? Because the architecture, where it can still be repaired, should be repaired at the level where repair is possible, not at the level where the symptom is loudest.

6. What makes a claim load-bearing

The filter is necessary and not sufficient. Wisdom makes the judgement, and a filter without it produces compliance rather than anything worth standing under. Both are required. This section develops the filter, then what it leaves undone, then two cases of the judgement actually happening.

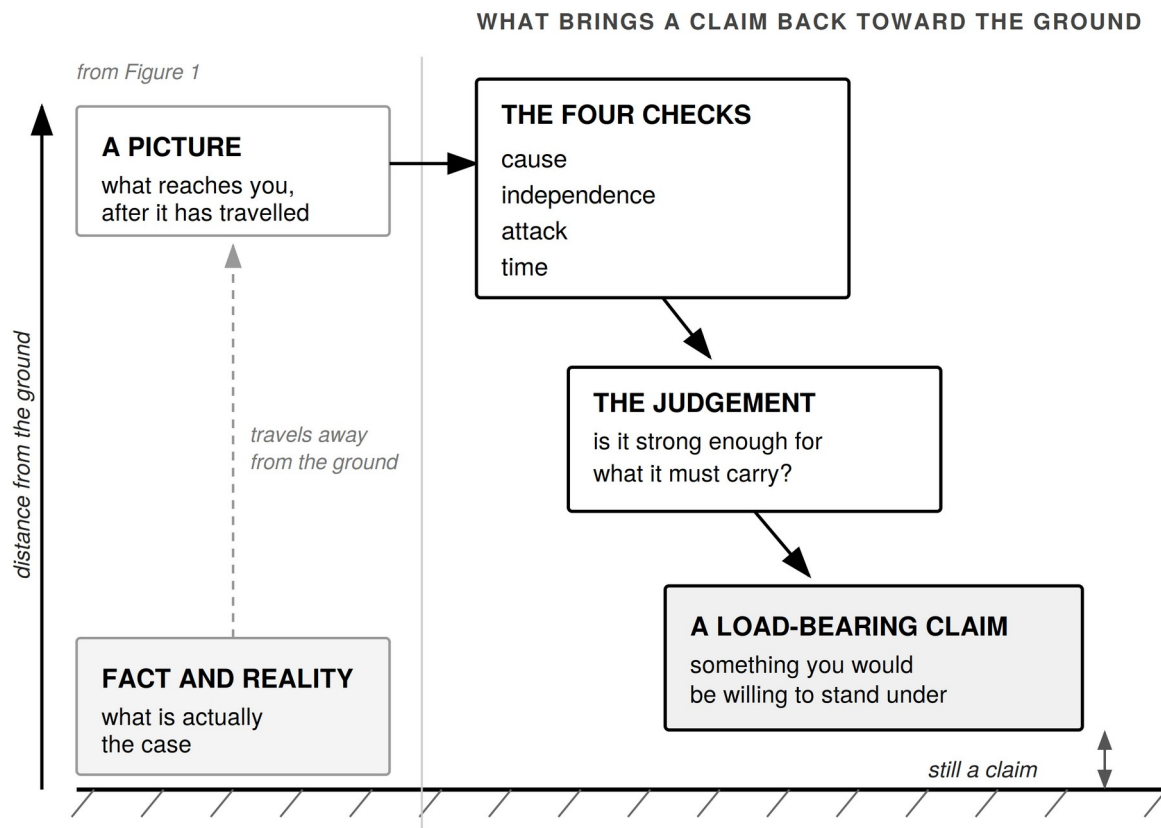


Figure 2. The four checks and the judgement are what bring a claim back toward what is actually the case. It never arrives all the way: a load-bearing claim is still a claim, and can still be wrong. What the architecture buys is a shorter drop: and a claim whose distance from the ground somebody has actually measured.

6.1 The four checks

Formally: causal consistency, observer-independent persistence, institutional adversarial testing, and time-delay arbitration. In shorter form, which is how the rest of the paper refers to them: **cause, independence, attack, time.**

6.1.1 Cause: does the cause hold up?

Does the claim survive doing and imagining? Act on the world, and the predicted consequence follows. Consider what would have happened otherwise, and the claim doesn't fall apart.

Correlation dressed as causation fails here. So does narrative consistency: a story that hangs together internally without tracking any causal structure. It passes the test of seeming coherent without doing any of the work the check requires.

COVID origins illustrates a variant worth naming separately. The evidence base was adversarially curated: access to the Wuhan Institute restricted, early case data withheld, sample selection shaped by parties with reputational and legal stake in the answer. Run a Bayesian estimate over that base and you get a confident-looking number. But the check cannot run at all, not because it returns a wrong answer, but because the conditions for performing it are absent. Observation under known intervention isn't available. The counterfactuals are hypothetical rather than testable.

That distinction matters more than it looks. **A check that fails and a check that cannot be performed are different situations, and the second is routinely reported as the first.**

6.1.2 Independence: do different observers agree?

Do people looking from different places, with different reasons to want different answers, see the same thing?

Agreement alone isn't enough. The observers have to be genuinely distinct: different methods, different traditions, different incentives, different access to the underlying Facts. Where such people converge, the claim has shown something no single source could have manufactured on its own: not one newsroom's angle, not one research group's method, not one agency's interest in the answer.

The failure mode is the one Section 5.3 named. Three newspapers running the same wire story are not three observers. Three platforms surfacing the same trending narrative are not either. The check needs observer positions whose differences *would show up as disagreement* if there were nothing real underneath. Their agreement counts only because disagreement was genuinely on the table. Collapse those differences upstream and the check has nothing left to test.

6.1.3 Attack: has anyone seriously tried to break it?

Has somebody with the skill and the motive made a genuine attempt to falsify this?

The right to challenge is not the same as challenge. Openness to criticism produces no criticism; only criticism does. A claim nobody has attacked has not been tested, whatever theoretical openness surrounds it.

So the check looks for evidence of attack, not evidence of permission. Did somebody with credentials and motive try? Did the attempt run through a process (peer review, regulatory scrutiny, failure analysis, replication) that gave them a real chance of succeeding? If yes, and the claim is still standing, the check is met. If not, or if the nominal adversaries have lost the capacity to mount a serious attack, it isn't, however much institutional approval the claim appears to carry.

6.1.4 Time: does it survive?

Does it hold its shape at ten years, and at a hundred where that applies?

A claim that looked true under the attention conditions of its own moment, and didn't survive their removal, has not passed. The check needs arenas where past claims get revisited: institutional memory, archives, longitudinal data, professional traditions that record who got it wrong as well as who got it right.

This is the most patient check and, where available, the most reliable, because it is the hardest to game. You can make a claim attractive in the moment. You cannot make it retroactively durable. It fails when the arena disappears: when nobody keeps the record, or keeps it and never looks.

6.1.5 Why all four

Each catches what the others miss.

A claim can be causally sound but observer-dependent: one brilliant theorist's framework nobody else can reproduce. It can be observer-independent but unattacked: widely held because nobody bothered. It can be well-attacked and still fail on time: convincing under present scrutiny, and predictably different-looking in twenty years. A claim can clear three and fall to the fourth.

The redundancy is deliberate: each check is there to catch what the others would let through. A claim needs all four to be eligible for the judgement, with one qualification, which the next section sets out.

6.1.6 Passed, failed, and cannot be run

A check has three possible outcomes, not two, and the third is not a softer version of the second.

A check **passes** when it has been run and the claim survived it. A check **fails** when it has been run and the claim did not. A check **cannot be run** when the conditions for performing it are absent.

The third is routinely reported as the second, and it is the worse position to be in. A failed check leaves a record. Somebody ran it, it came back negative, and that negative is information: it can be cited, disputed, and acted on. A check that could not be run leaves nothing behind, and nothing is also what a clean result looks like from the outside. The danger is not that the claim is weaker than it appears. It is that there is no trace of the weakness for anyone to find.

Three different things can put any check, on any claim, out of reach, and the honest response to each is different.

The conditions have been removed. Somebody, or some arrangement, has taken away what the check needed. COVID origins is this kind. The cause check did not return a wrong answer; it could not be carried out, because access to the site was restricted and early case data was withheld. This is in principle remediable: restore the access and the check can run. Where it is not remediable, the honest report names who removed the conditions and when.

The time has not elapsed. The check is the right instrument and the claim is the right kind of thing, but the answer arrives later than the decision does. The time check is the usual case, since it measures years and most decisions will not keep. This one resolves itself by waiting, which is precisely what is not on offer. Section 6.4 is that situation.

The claim is not the kind of thing the check can take hold of. Nobody interfered, and no amount of waiting helps, because what the check needs to examine is not there to examine. This kind does not resolve. The only route is to go behind the claim, to whatever it was built from, and run the check there instead. Section 5.2 (AI output) takes up the case where this has started to happen at scale.

This settles something the four checks otherwise leave open. Where a check cannot be run and the decision cannot be postponed, the honest position is not that the claim has cleared the filter. It is that the claim cleared everything that could be run, that one check stands open, and that acting anyway is itself a judgement: about what the delay would cost, and about what the claim is being asked to carry in the meantime.

6.2 What the filter leaves undone

The filter produces eligible candidates. It does not produce load-bearing claims.

Four things separate a cleared candidate from something worth standing under, and none of them reduces to procedure.

Is this the right question? A filter that has run cleanly hasn't asked whether this was the claim to be making. A clinician with a textbook diagnosis still has to ask whether the diagnostic question is the live one. A policymaker with a confident projection still has to ask whether the decision actually turns on it.

What's at stake? The weight of the claim has to match the weight of the use. A cleared candidate determining one line in a patient's chart needs less than the same candidate determining public health policy. Part of the judgement is knowing the difference.

Is the filter even the right instrument? Some situations can't take it. Cases where you cannot put odds on anything, because you do not know

what the possible outcomes are. Frank Knight called this uncertainty, and distinguished it from risk, where at least the odds can be estimated. Questions whose answers arrive long after the decision has to be made. Architectures degraded enough that the filter's own outputs can't be trusted. The judgement includes recognising that no procedure, however carefully run, is producing what the situation needs: and then saying plainly that you are acting on a judgement call, and saying what it rests on, rather than presenting a guess as though it were a finding.

Do I know where my own edges are? The calibrated map of one's own knowledge that the Wisdom Gap paper named as the most valuable property of senior practitioners: knowing where your judgement is solid, holding it there against pressure, and refusing it where it isn't.

These are continuous with the Wisdom Gap thesis. Wisdom has preconditions you can name: developmental experience, having something at stake, attention shaped by feedback over years. Naming them does not add up to wisdom. **The filter is what wisdom uses. It is not what wisdom is.**

6.3 A senior clinician

A patient arrives with a cluster of symptoms that fit the textbook presentation of condition X. The chief complaint fits. The examination produces the expected findings. The labs return the expected values. "This is condition X" has cleared the filter at the population level: guidelines support it, peer-reviewed evidence supports it, replication supports it, and the long-term outcome data gives the expected response rates.

A junior clinician prescribes the standard treatment. A senior clinician, who has seen hundreds of presentations of X, hesitates.

The hesitation has no clear cause. Something about the patient's phrasing on onset. Or the order the symptoms appeared in. Or a small finding the textbook doesn't flag. Perhaps nothing she could name in the moment. The hesitation produces a question: is this X, or is it Y presenting atypically?

She orders one more test. It shows Y. The treatment differs. The patient does well.

That is the judgement, at everyday scale. The candidate had cleared the filter; the filter had no work left to do. What she brought was not better filter-running. It was the recognition that a cleared candidate might not be strong enough for the decision in front of her.

Note what is and isn't available afterwards. It is **observable in operation**: the chart, the order, the documented decision point. It is **traceable through the outcome**: the patient's trajectory differs from what X treatment would have produced. It is **not specifiable in advance**: she

cannot say beforehand what would trigger the doubt, and afterwards only partially ("the cough sounded different from how X usually presents"). The doubt was held by someone whose attention had been shaped by hundreds of prior cases, and that attention is what the architecture produced.

One caution, because this example can be read the wrong way. What makes it good judgement is not that she turned out to be right. Had the extra test come back clear and the patient recovered on the standard treatment, she would still have been right to order it: the question she asked was the live one, and the cost of asking was small against what it would have caught. The outcome is how we come to hear the story. It is not what the judgement is assessed on, and 6.5 sets out why.

It isn't dramatic. It doesn't announce itself. It is a senior clinician ordering one more test the textbook didn't require, and having a reason.

6.4 October 1962

A U-2 flight over Cuba returns photographs confirming Soviet medium- and intermediate-range ballistic missiles, capable of striking most of the continental United States. CIA analysts and specialists in Soviet deployment review them. Multiple independent observers, working from the imagery and from corroborating signals intelligence, reach the same conclusion. The Joint Chiefs, the ExComm, and the President are briefed.

The claim, Soviet offensive missiles are operational in Cuba, clears every check that can be run. **Cause:** the photographs show what they show, and the mission is repeatable; further flights confirm. **Independence:** multiple analysts, different priors, different methods, same conclusion. **Attack:** contrarian readings of the imagery are entertained and rejected on the evidence. **Time** cannot be run, in the second of the three senses 6.1.6 sets out: the check is the right instrument and the claim is the right kind of thing, but it measures years and the decision will not keep. That leaves the claim three for three with the fourth standing open, and acting on it is itself a judgement about what waiting would cost. The claim does hold across the days of the crisis as more evidence arrives, but that is corroboration, not the time check.

What the filter has produced is a candidate that has cleared everything available to it. What it has not produced is a decision.

The Joint Chiefs press for air strikes followed by invasion. The argument is coherent: the missiles are an immediate threat, the window for removing them is narrow, and the United States has overwhelming superiority. The cleared candidate, treated as sufficient on its own, points toward direct action.

Kennedy chooses a naval blockade.

The choice does not follow from the filter's output. The filter produced "missiles in Cuba." The choice between blockade and air strike turns on something else: the recognition that this claim was about to authorise an action whose consequences might exceed what could be reversed. Strikes killing Soviet personnel could trigger retaliation against Berlin, against American forces in Turkey, or directly. Each of those has its own consequences behind it. The blockade preserves options; the air strike forecloses them.

The judgement is not in seeing the missiles. It is in knowing what their presence will and will not justify.

Worth adding what the contemporary record concealed: the settlement included a private American undertaking to withdraw Jupiter missiles from Turkey within months, communicated through Robert Kennedy to Dobrynin and kept out of public view. This strengthens rather than complicates the case. Part of what Kennedy recognised was that a concealed concession was worth more than a public victory: which is a judgement about the weight of the use, not about the evidence.

It is observable in operation: the ExComm minutes exist, the tapes were released decades later, and multiple participants corroborate. It is traceable through the outcome: the missiles were withdrawn. It is not specifiable in advance: no decision rule produces "blockade" from those inputs. The wisdom is in the holding: keeping the alternatives present, including the ones the loudest voices wanted dismissed, and not collapsing to a single course until the judgement could be made well.

(One sourcing note: Robert Kennedy's Thirteen Days is a participant's shaped account and is treated as such by historians. The May and Zelikow transcripts are the stronger citation for the deliberations themselves.)

6.5 The obvious objection

The judgement step invites a fair challenge: it sounds unfalsifiable. Architecture intact and the outcome bad? The judgement wasn't performed. Outcome good? It was. The diagnosis collapses into a theory that explains everything after the fact.

The answer is to separate two things the objection runs together: **what the practitioner did, and what the world did next.**

The judgement is assessed on the first. Was the claim matched to the weight of the use? Were the alternatives held open rather than collapsed early? Was the institutional pressure working against the decision named rather than absorbed? Was the practitioner's own limit acknowledged? Those are observable while the decision is being made, and reconstructable afterwards from the record of the decision: not from the outcome.

Run the counterfactual on Kennedy. Suppose he chose the blockade, made the judgement exactly as the record shows, and the Soviets escalated anyway. The framework would not say the judgement failed. It would say the judgement was sound and the outcome was catastrophic. Those are different claims, and the catastrophe is not evidence that the decision should have been different.

The reverse holds. Someone who ratifies the loudest voice in the room, holds nothing open, and gets lucky has not made the judgement. The good outcome is not evidence that they did.

A follow-on question: if the markers are observable, why isn't the judgement simply a fifth check?

Not, as an earlier version of this argument had it, because the question never bottoms out: each answer needing another answer behind it, turtles all the way down. It does bottom out. You could write a fifth check asking "were the first four actually runnable here?" and it would stop there quite happily.

The real answer is simpler. **The markers are specifiable; the judgement they issue in is not.** A checklist can verify that somebody held the alternatives open. It cannot tell them which alternative to choose. That gap doesn't close by adding checks, because every additional check is one more thing to run, and none of them decides.

A sharper version of the same objection is worth conceding outright. The four checks are not judgement-free either: deciding what counts as a genuine intervention, a genuinely independent observer, a serious attack or a long enough horizon all take judgement. So the distinction here is not between a mechanical part and a human part. One kind of judgement can be measured against a standard somebody wrote down. The other kind decides how much that standard is worth in this particular case, and there is no further standard to measure that against. The judgement is what acts where procedure ends: not because procedure runs out of room, but because that decision was never procedure's to make. It belongs to a person, and always did.

7. Where the architecture comes from

Three cases. Air-crash investigation, where the architecture has had decades to mature in a field nobody much argues about. The Federal Reserve under Volcker, where it works under constant political pressure in a field people argue about all the time. And the Federal Reserve again, a quarter-century later, where everything was in place and the judgement still was not made.

7.1 The National Transportation Safety Board

When an airliner crashes, the NTSB investigates. It is independent of the airlines, the manufacturers and the FAA by statute. Investigators get the wreckage, the flight data recorder, the cockpit voice recording, the maintenance logs, the crew records. Teams work in parallel, structures, engines, human factors, operations, weather. The findings are public. The recommendations have teeth, because the FAA can mandate compliance and the industry depends on people being willing to get on planes. The cycle from crash to investigation to industry-wide change runs in years, not news cycles.

All four checks are built into the design rather than left to anyone's conscience.

Attack. The NTSB is not the airline, not the manufacturer, and not the regulator that certified the aircraft. Each of those has a reason to read the evidence in a particular direction. Independence from all three is what makes the attack real rather than nominal.

Cause. The wreckage exists. The recorders exist. Investigators reason about physical material, not about somebody's account of physical material.

Independence. The structural engineer reads the wreckage, the human-factors specialist reads the crew interaction, the operations expert reads the flight profile. They start from different places. Where they converge, the convergence means something.

Time. A recommendation issued after one accident gets tested by what does or doesn't happen over the following decade. Where it prevented later accidents, the approach is validated; where it didn't, the approach changes. The arena in which time can do its work exists by design.

The conditions for judgement are there too. Senior investigators carry hundreds of prior investigations, and the institution is built so the judgement (this is the factor that matters here, this is the recommendation that addresses it) is made by people the institution itself formed.

Commercial aviation has become extraordinarily safe over decades. Not because individual pilots got better. Because the architecture turns each failure into a constraint the whole system inherits.

7.2 The Federal Reserve

The same structure works elsewhere. The Federal Reserve does the same work in a field where the underlying claims are politically and economically disputed.

Independence is written into the Federal Reserve Act. Governors serve staggered fourteen-year terms, longer than any presidential cycle. The

Federal Open Market Committee combines the Board with rotating regional bank presidents, which distributes the institutional voice across geographies and constituencies. Dissents are on the record. Transcripts of FOMC meetings are released after about five years, which makes the deliberations available to historians without exposing them to immediate political pressure.

Against the four checks: attack is built into the voting and dissent structure; cause runs over the data infrastructure of the whole System, including what the regional banks know about local economies; independence runs across those regional vantage points and the Board's national one; and time operates through the released transcripts and the long horizon over which monetary policy gets judged.

The judgement case is Paul Volcker. He took office as Chairman in 1979 facing double-digit inflation, with the textbook theory of the day predicting that any sustained monetary contraction would produce mass unemployment. The claim that had cleared the filter (that breaking inflation required sustained contraction, even at severe near-term cost) pointed toward holding interest rates at historically high levels.

The pressure against doing so was intense. Unemployment rose sharply. The midterms went against the administration. Members of Congress called for his resignation.

He held the rates.

The judgement was that the claim was strong enough for what it was being asked to carry: and, more than that, that backing off under political pressure would not merely delay the disinflation but teach the markets that the Fed's commitment was conditional on political comfort, which would make the original problem worse. The architecture is what let the judgement land. Independence meant he could not be removed for it. The committee structure put the dissents on the record. The five-year delay meant the deliberations could be analysed later without being weaponised at the time. The judgement was Volcker's. The architecture is why it held.

7.3 What the two have in common

Five things.

Independence from the parties whose claims are being tested. The NTSB is not the airline; the Board is not the executive branch.

Access to the thing itself. What investigators call ground truth: the wreckage in one case, the System's own data infrastructure in the other, rather than somebody's account of either.

Outputs that become public, immediately or on a delay. Reports and recommendations; transcripts and decisions.

Time horizons longer than any career. The codes outlive the investigators who prompted them; a fourteen-year term outlasts any single political moment.

Attack built into the structure rather than left to virtue. Multi-disciplinary investigation; committee voting and recorded dissent.

These are what any architecture needs to support claims worth standing under. The two cases are illustrative, not exhaustive. The requirements are general, so the same structure works in other fields.

7.4 What the architecture cannot supply for itself

Specifying what an institution must do does not make anyone do it.

The NTSB works partly because the airline industry, on balance, actually wants safety. Passengers prefer airlines that don't crash, insurers price safety into premiums, and one fatal accident costs an airline far more than years of safe flying earns it. The architecture runs with the grain of the industry's interests rather than against it.

Fed independence works partly because, until fairly recently, both major parties agreed monetary policy should sit outside direct partisan control. The agreement has been pragmatic rather than principled, each party has found it inconvenient at various points, but it has mostly held.

This is not a criticism of either architecture. It is an observation about what they rest on. **Any architecture that produces claims worth standing under depends on incentives it cannot generate for itself.** Where the alignment holds, it does its work. Where the alignment breaks, the outputs keep appearing (the reports still get published, the rate decisions still get issued) and what changes is whether anyone still treats them as load-bearing.

The Governance whitepaper handles how incentives and institutional design interact. What matters here is the narrower and more honest point: necessary, not sufficient.

7.5 The Fed and the housing bubble

Section 7.2 treated the Fed as a working architecture. This is the same institution, a quarter of a century later, failing, and it is the case the whole framework depends on being available. If intact architecture guaranteed good judgement, "necessary but not sufficient" would be an empty phrase.

What could have been said by 2006, in plain terms. House prices had risen far beyond what incomes could support. A great deal of mortgage

lending had been written on terms that only worked if prices kept rising. The securities built on those mortgages were priced as though the underlying loans would not go bad together. And major financial institutions had borrowed against those securities heavily enough that a correlated downturn would wipe out their capital. Put together: the financial system is leveraged in a way that will spread failure widely.

The information was available. It was distributed across exactly the parts of the institutional apparatus you would expect to surface it.

On the price data, and a complication the argument needs. The national price-to-income ratio reached 4.7 in 2005, the highest since at least 1980 (Harvard Joint Center for Housing Studies). Of 382 metropolitan areas, 226 peaked between 2005 and 2009.

But the variation was enormous. Los Angeles peaked near 11.7 and New York near 7.5, while much of the Midwest and interior Northeast stayed below 3.0. Texas had a population boom and almost no price bubble.

That variation is not a footnote, and the argument is weaker if it gets smoothed away. A Fed official in 2006 could look at the same data and say, accurately, that this was a set of regional stories with regional explanations, supply-constrained coastal cities and the sand states, rather than a national phenomenon. Roughly speaking, that is what the FOMC did say.

The answer is that the regional variation was itself the thing being relied on. The securities were rated on the assumption that mortgages in different places would not fail at the same time. Geographic diversity was the argument for safety. So the fact that the boom looked regional was not evidence against systemic risk; it was the premise the systemic risk was built on.

On the lending. Loan quality deteriorated steadily and measurably. Across roughly a million securitised first-lien subprime loans, low- or no-documentation lending rose from 23.5% in 2001 to 37.7% in 2007, and combined loan-to-value ratios climbed throughout (Demyanyk & Van Hemert, *Review of Financial Studies*, 2011). Their conclusion was that quality declined every year from 2001 to 2007 once you adjust for loan characteristics and economic conditions.

The point is not that anyone designed these loans to fail. It is that a significant share of them could not perform on their own terms without continued price rises or refinancing, **and the deterioration was visible in the data at the time.**

On the securities. The models did not assume mortgages were uncorrelated. They assumed correlation was low, that geography diversified it, and, the decisive assumption, that house prices would not fall across the country at once.

The Financial Crisis Inquiry Commission found how those correlation figures were arrived at. A Moody's managing director described analysts being asked how related they thought different mortgage securities were, and characterised the answer as that they made them up. The June 2004 model was calibrated on the preceding twenty years, a period of rising prices and low delinquencies. The Commission put the failure in a single image. Toss a hundred coins and each one is independent of the rest; leave a loaf out overnight and every slice goes stale together. The securities, it concluded, turned out to behave "less like coins than like slices of bread." By late 2008 Moody's was using correlations two to three times higher.

On the four checks.

The cause check did run. It ran against an explanation nobody had re-tested. Economists had agreed since the mid-1980s that swings in the economy had become milder, the "Great Moderation." What they had not agreed on was why. Ben Bernanke's 2004 argument was that better monetary policy deserved much of the credit. James Stock and Mark Watson, whose work established the pattern, attributed only 20-30% to policy and most of the rest to good luck, and specifically found that structural changes in the economy did not explain it.

The good-luck reading is the more dangerous one, and it deserves saying plainly: if the calm was mostly luck, it could end without notice. The institution acted as though it were durable regardless. Under that assumption, rising asset prices and rising household debt looked like adjustment to a more stable world rather than something building up. The check ran. It ran against an account that had not been re-tested for the conditions it was now being applied to, which is a harder failure to catch than a check that doesn't run at all.

The independence check failed hardest. Fed staff, Treasury, the SEC, the rating agencies, the major banks and the academic literature on financial stability had largely converged on one picture of how the system worked. But they were reading from the same intellectual framework, citing the same data, and had reasons to prefer continuity. That is not what the check tests for. The agreement was inherited, not independent, and nothing inside the architecture was built to tell the difference.

The attack check was present on paper and weak in practice. Dissenting voices existed, and they were not obscure: - **Dean Baker published a detailed warning in August 2002**, four years early, showing that real house prices had risen 29% while real rents rose 10%, that two-thirds of the run-up was accounted for by the gap between buying and renting, and predicting price falls of 11-22% destroying \$1.3-2.6 trillion in wealth. Public, specific, quantified, and falsifiable.

- **William White, the Bank for International Settlements' chief economist, made the case for leaning against credit booms directly to Alan Greenspan at the Fed's own Jackson Hole conference in August 2003.** Greenspan rejected it on the record: "There has never been an instance, of which I'm aware, that leaning against the wind was successfully done."
- **Edward Gramlich**, a sitting Fed governor, spent years on subprime lending and documented the regulatory gap precisely: of 2005 subprime originations, only 20% came from federally supervised banks and thrifts, 30% from bank affiliates under partial supervision, and **50% from state-chartered independents with almost no federal oversight**, where most reports of abusive lending originated. His line was that "in the prime market, where we need supervision less, we have lots of it." - **Raghuram Rajan's 2005 Jackson Hole paper** was not about housing; it was about incentives in finance and hidden tail risk, but its reception is the relevant evidence. Lawrence Summers called the premise "slightly Luddite" and "largely misguided." Donald Kohn, then a sitting Fed governor, suggested Rajan was nostalgic for bank-dominated finance. The rebuke came from inside the tent.

The right to challenge was fully present. The challenges happened. **What did not happen was a re-test.** The prevailing consensus was strong enough to hold the challenges at the edge without ever having to refute them.

And one detail complicates the picture in a way worth keeping. Gramlich had raised examining bank lending affiliates with Greenspan privately in 2000 and been turned down. But writing to Greenspan in 2007, after the episode surfaced in the press, he described it as "a small incident, and as I think you know, if I had felt that strongly at the time, I would have made a bigger stink." **The dissent failed to break the filter partly because the dissenter did not treat it as filter-breaking.** That is a more interesting finding than a story about a suppressed whistleblower, and it says something about how institutions absorb internal doubt.

The time check had not run. The Great Moderation was about twenty years old and had operated across an unusually benign stretch, including asset-price rises that had not been followed by systemic trouble. The check that uses time as its instrument had not yet had time.

On the judgement, and what the draft of this paper previously got wrong.

It is tempting to read Greenspan's December 1996 "irrational exuberance" remark as a moment of judgement he later retreated from. It doesn't survive reading the passage. The phrase appears as a question: how would we know when irrational exuberance had unduly escalated asset values? Not as a

claim that it had. And the same speech already contains the doctrine that would justify doing nothing: that central bankers need not worry about a collapsing asset bubble unless it threatens output, jobs and prices.

So there was very little to retreat from. What followed was not retreat but hardening. By Jackson Hole in August 2002 the position was explicit: bubbles cannot be identified until they burst, and the idea that well-timed tightening could have prevented the late-1990s bubble is "almost surely an illusion."

The better reading, and the one that serves this framework: the possibility was available to be entertained at the highest level a decade before the crisis, and entertaining it changed nothing, because the argument for inaction was sitting in the same speech. The problem was never that the claim was unthinkable.

And the sharpest version of the indictment is not the one usually made.

The doctrine of the period is often summarised as "clean up afterwards rather than lean against the wind." But Bernanke's version was never "do nothing." Speaking in New York in October 2002, he set out a division of labour: the Fed should point monetary policy at its macroeconomic goals **while using its regulatory and supervisory powers to help ensure financial stability.** Monetary policy was the wrong tool for bubbles; supervision was the right one.

The Fed then did not use the supervisory tool. The Financial Crisis Inquiry Commission put it directly: the Fed's "pivotal failure to stem the flow of toxic mortgages ... could have [been done] by setting prudent mortgage-lending standards ... The Federal Reserve was the one entity empowered to do so and it did not."

So the failure is not that the Fed refused to raise rates against a bubble it could not confidently identify. **It is that the Fed's own doctrine promised that supervision would handle what monetary policy would not, and then the supervision did not happen.** That is a claim about an institution failing on its own stated terms, which is a much harder thing to argue with.

What the transcripts show. The 2006 FOMC transcripts were released in January 2012, and they are the archive doing its job. In March 2006 Bernanke said "strong fundamentals support a relatively soft landing in housing." Timothy Geithner said the effects of a cooling housing market would be "modest." In September, Kevin Warsh said the capital markets "are probably more profitable and more robust at this moment than they have perhaps ever been." The Fed's own chief economist, David Stockton, on the difficulty of forecasting housing: "right now, it feels a bit like riding a roller coaster with one's eyes shut."

And at Greenspan's final meeting as Chairman, on 31 January 2006, Geithner told the room: "I'd like the record to show that I think you're pretty terrific ... And I think the risk that we decide in the future that you're even better than we think is higher than the alternative."

That is what a filter looks like at the moment it most needed to work and didn't.

What the case establishes. The five properties of 7.3 preserve the conditions under which judgement can be exercised. They do not guarantee it will be. The same architectural form explains how both NTSB-style success and Fed-2006-style failure are possible, and both are consistent with the same structural commitments. Telling them apart is the work: which check ran, which check ran against a stale model, where agreement was inherited rather than independent, where the challenge was made and absorbed rather than answered. The framework can name what failed without claiming the failure was inevitable.

8. What lateral structures can and cannot do

8.1 The substitution

The architectures in Section 7 are expensive. They need statutory independence, multi-disciplinary teams that take years to assemble, terms longer than political cycles, and incentive alignments that never come for free. Where they are in working order, they do what nothing else can, and Section 7.5 is the reminder that existing and working are not the same thing. Where they are absent (and increasingly, in fields where institutional review has eroded faster than anyone has proposed rebuilding it) the question of what could stand in becomes live.

Three candidates come up: prediction markets, crowdsourcing, and algorithmic reputation. Each aggregates scattered signal across many participants. Each is faster, cheaper and more scalable than the institutional form.

The question is what each actually delivers against the four checks, and what none of them delivers at all.

8.2 Prediction markets

A prediction market aggregates scattered beliefs by attaching consequences to them. Participants back positions with money; the price is the weighted aggregate of their bets. It is Hayek's insight about dispersed knowledge fitted with Taleb's discipline: beliefs with nothing riding on them drift, and beliefs that are paid for tend not to.

Cause runs on the participants directly. A trader holding a causally incoherent position pays for it over time, and on cleanly-resolving questions the price tends to track what the causal structure supports.

Independence is partial. The market aggregates many people, but those people frequently read the same news, the same commentary, the same feeds. Sometimes the bets converge because of independent inference and sometimes because of shared inputs, and the market cannot tell the difference from inside itself.

Attack is absent in its institutional form. Traders disagree, which produces price discovery, but disagreement between self-interested participants is not the same as being examined by somebody whose job is to break the claim regardless of their position.

Time works on resolvable questions and only those. An election settles on election day. A drug approval settles when the agency rules. Questions about climate severity, the soundness of a financial system, or whether a contested historical claim will hold across generations do not settle, because they have no resolution date the structure can register.

And the market cannot do the judgement. Imagine a prediction market in October 1962 on whether the United States would strike Cuba within ten days. It would have produced a probability. It would not have produced Kennedy's recognition that the cleared claim was about to authorise something that might not be reversible. A price is not a judgement about what a claim can carry.

8.3 Crowdsourcing

Crowdsourcing distributes observation, curation and correction across many people. Wikipedia is the standard case; community notes on social platforms are a newer one; citizen science is a third.

Cause gets partially tested through edit history: inconsistencies surface as edits, and the trail lets you reconstruct how a claim arrived at its current shape. On contested topics, though, motivated editing produces internally consistent accounts polished by people with positions. The check passes the surface test without doing its work.

Independence is the strongest of the four on settled topics. Many editors, different priors, different sources, converging on a stable rendering, that does some of what the check requires. On disputed topics the same structure produces edit wars whose outcome depends on who has more time and stamina, which is a different thing entirely.

Attack is present on paper and weak in practice. Defenders of established framings are usually better resourced than challengers, and on community notes the volume of claims exceeds the capacity to test them. The check

looks for evidence of attack and finds, on contested questions, evidence of access without exercise.

Time emerges slowly through the edit history. Claims that turned out wrong get corrected eventually and the trail of corrections is public. It is patient where the question is settled and fails where the question is contested during the same period the editing is happening.

What crowdsourcing cannot do is tell reported from load-bearing.

Wikipedia is encyclopedic by design: it reports what sources considered reliable have said, weighted by which sources are considered reliable. That is useful. It is not the judgement. The senior clinician in 6.3 was not asking what most clinicians would have recorded. She was recognising that a cleared claim wasn't strong enough for the decision in front of her. Crowdsourcing has no equivalent, because distributing the work distributes the stake to nobody in particular.

8.4 Algorithmic reputation

Algorithmic reputation systems work out a score, a ranking or a moderation decision from many signals. The early ones were not AI at all. PageRank ordered search results by who linked to whom; citation indices ranked papers by who cited them. Both followed rules somebody had written down. The recent ones (platform voting, content-moderation weighting, product ratings) increasingly do use AI, and the newest put AI language models directly into the ranking. This section calls them all ranking systems, because the four checks fare much the same either way.

This is the weakest of the three.

Cause is not tested at all. The system aggregates signal; it does not test signal against causal structure. A widely-cited claim may be widely cited because it is right or because it is convenient, and the algorithm cannot see the difference.

Independence is where the channel problem bites hardest. Ranking systems cannot distinguish independent verification from inherited consensus. When the sources feeding them (the wire services, the trending feeds, the handful of outlets everyone else cites) have already collapsed observer independence, the convergence the algorithm registers is evidence of shared inputs. The system rewards what the channel produced, which is precisely what the check exists to test against.

Attack is largely absent. Gaming is the adversarial mode ranking systems contend with, and gaming is not testing. A claim nobody has gamed is not a claim anybody has tested.

Time is weak by design. Ranking systems are tuned to recency, and the historical record they consult is the record of their own past outputs, which

makes the arbitration circular. The arena where time could do its work collapses into the arena the algorithm itself created.

One thing to flag here and develop in Section 9. These systems increasingly decide their rankings using machine learning, software that works out its own weightings from past data instead of following rules somebody wrote down, and the newest generation puts AI language models directly into ranking and moderation. Each step puts more distance between the ranking and any independent check on it. Whether that is a straight loss of independence or a change in what the ranking depends on is not settled. What is clear is that the reader can inspect less at every step. **The thing we hoped would check the pipe has become part of the pipe.**

8.5 What they are for

All three can feed a filter. None makes the judgement, and that absence is structural rather than incidental.

The shape of the question matters as much as the shape of the structure. Binary, near-term, cleanly-resolvable questions suit prediction markets. Settled empirical questions suit crowdsourcing. Surface reputation questions suit algorithmic systems. Structural claims about causes, contested pictures, and questions whose answers arrive after the decision has to be made defeat all three, because none of them is tuned for that kind of question.

Think of them as suppliers, not substitutes. They feed an institution that makes the judgement; they do not replace it. The NTSB's parallel disciplinary teams are a lateral input. The Fed's regional bank intelligence is a lateral input. Neither institution reduces to its inputs. What the institution adds is the conditions under which somebody it formed can make the call.

9. Implications

This section diagnoses; it does not prescribe. It does not propose particular institutions or particular legislation. It says what any institution, any governance framework, and any approach to AI-mediated discourse has to respect.

9.1 For institutions

The five properties at 7.3 are not features of the NTSB and the Fed in particular. They are the specification.

So the question to ask of an existing institution is not whether it has the trappings of authority but whether it has the five properties in working

order. Where they are present, it can support the claims it produces. Where one or more has gone, the claims keep coming out and their standing depends on whether anyone is still treating them as load-bearing, which is the situation Section 5 described as an architecture failing while still producing artifacts.

For fields that never had such institutions, or whose institutions have eroded faster than anyone has proposed rebuilding them, the harder implication is that the work cannot be skipped. An institution missing one of the five cannot, by itself, support the full range of claims. What it can still support is narrower, and which claims those are depends on which property is missing: an institution with no long memory can carry a claim whose time horizon is short, and cannot carry one whose horizon is long. Beyond that range it is producing candidates of varying quality whose judgement, where it happens at all, is happening somewhere outside its walls.

That is not a reason to give up. It is a specification of what would have to be built. A structure designed for compression cannot be repurposed for tension by relabelling it.

9.2 For governance

The Governance whitepaper set out what an AI-era governance framework must do to keep human judgement in authority over the AI system (the machine, as I usually call it): cultivate the conditions under which the practice develops, require its exercise in the moment, preserve the institutional conditions that sustain it across careers, and surface the signatures of its absence where inspecting individual outputs cannot reach.

What this paper adds is what the governance is *for*.

The Governance paper specified how the practice develops, is exercised, is sustained and is verified. It could not, from inside its own walls, say what that practice produces at the system level. Without that, the four requirements float free of their purpose: an institution that cultivates, requires, preserves and surfaces a practice without knowing what the practice is for has built a compliance regime around an output it never characterised.

The output is claims worth standing under: eligible under the four checks, and judged load-bearing by people the architecture itself formed.

Each of the four requirements does a specific piece of this work. *Cultivate* builds people capable of making the call. *Require* ensures the checks are actually run rather than noted. *Preserve* protects the conditions under which attention can be shaped over a career long enough to produce the capacity. *Surface* verifies, where individual documents cannot show it, that the practice is happening rather than being performed.

Frameworks that meet the four requirements without specifying the output will drift toward whatever the institutional incentives reward, which is rarely claims at the standard the architecture was built for. Frameworks that specify the output without architecting the requirements will produce the language of rigour with no practice underneath it. Both are necessary. Neither is sufficient.

9.3 For AI-mediated discourse

This is the most time-sensitive implication, and it needs stating carefully, because the strong version of it is not what the evidence supports.

What is measurable. In March 2025, 18% of Google searches by a tracked panel of US adults returned an AI summary. When one appeared, people clicked through to an actual source 8% of the time, against 15% when it didn't; and clicked a link inside the summary 1% of the time (Pew Research Center). Referral traffic from Google search to thousands of publisher sites fell by about a third in a year (Reuters Institute).

So the shift is not that people have started getting their news from chatbots, only about one in ten use them for news weekly, and one in a hundred call them their main source. **The shift is that a synthesis layer has been interposed between readers and sources at a scale no newsroom could produce, whether or not anyone asked for it.**

On what that layer does to confidence signals, and this is the sharpest finding available. It is tempting to say AI writing strips out hedges. Measured at the word level, the opposite is true: AI systems use hedging words somewhat *more* often than human writers do (Reinhart et al., *PNAS*, 2025).

What has actually broken is the relationship between the hedge and the state behind it. Researchers measured how certain a model internally was against how assertive its language sounded, and found essentially no relationship between the two, while human readers picked up the assertiveness accurately (Ghafouri et al., 2024). **The hedges are still there. They have stopped meaning anything.**

The practical consequence is visible in how AI tools fail. Across 1,600 queries to eight AI search tools, one system misidentified 134 of 200 news articles while signalling low confidence just 15 times, and never once declined to answer (Tow Center, 2025). The paid tiers were worse, because they gave definitive wrong answers rather than declining. A study of 2,709 responses about news, assessed by journalists across 18 countries, found 45% contained at least one significant problem and 31% had significant sourcing problems (EBU/BBC, October 2025). And in legal research tools marketed as hallucination-free, accuracy ran from 65% down to 19%.

On opacity, with the objection conceded up front. Human mediation was never transparent. Editors' framing decisions were not inspectable, newsroom routines were invisible to readers, and the corrections that did exist had thin uptake. The difference is not degree of opacity but where it sits and what can be done about it. Editorial opacity is institutional, local, and attached to a named person who can be written to, sued or fired. Model opacity is distributed across a body of training text and a weighting procedure with no corresponding party to address. A byline names a person, not a decision procedure: but it does at least name somebody.

On whether readers are fooled: the honest answer is not the convenient one. When content is labelled as AI-generated, people discount it: comfort with entirely AI-made news runs around 12% against 62% for human-made. Unlabelled, though, readers cannot tell the difference at all: in a controlled study, AI-generated and human-written news were rated identically for credibility, expertise and readability before disclosure.

So the problem is not that people over-trust AI. **It is that they cannot identify it, and that they calibrate their reliance to how confident the output sounds, which no longer tracks how reliable it is.** In one experiment people relied on unhedged answers 90% of the time, and their own accuracy fell from 86% to 76% when the system was overconfident. And the labels that would trigger the discount mostly don't fire: roughly one in five people see AI labels daily, while more than three in four consume news daily.

One finding worth the space, because it isolates the mechanism. In seven experiments with over ten thousand participants, people who researched a topic through an AI assistant wrote advice afterwards containing fewer specific facts and far more similar to one another's than people who used a search engine. **The second experiment held the underlying information identical across both conditions and the effect held** (Melumad & Yun, *PNAS Nexus*, 2025). The synthesised format itself produces shallower understanding, even when the facts are the same.

A limitation worth stating rather than waiting to be caught on. Almost none of this research establishes a human baseline. Nobody has measured how human summarisers perform on the identical task under the same rubric. The argument here is about a difference between two regimes, and most of the evidence measures one of them.

What follows. A generation ago, the things that carried information to the public (newspapers, broadcast news, peer-reviewed journals, professional commentary) had named people in them, whose choices could be examined in principle and argued with in practice. What is replacing them is assembled by AI, at a scale no newsroom could match, with the choices behind it spread across training data and ranking systems that no reader can inspect.

Frameworks that treat this as a continuation of the old infrastructure (to be handled by extending existing rules on disclosure, transparency and accountability) will fail at the layer where the substitution happens, because those rules were written for a channel that had people in it. The form has changed. Recognising the change has to come before responding to it.

10. The practice

10.1 What this paper has and hasn't done

The framework specifies a system-level output: claims that working architectures produce, eligible under the four checks and judged load-bearing by people those architectures formed. That is what Section 7's institutions produce when they work, what Section 8's lateral structures cannot produce alone, and what Section 9 asks any framework to build for.

What it has not addressed is what such an architecture produces *in the people who work inside it* over a career: the practitioner-level output, as distinct from the system-level one. That is a paper still to be written. The same architecture produces both: claims at the system level, and practitioners at the personal level. Pointing that out is all this paper needs to do.

There is also a limit on what the framework reaches, and it is worth stating plainly rather than leaving to be discovered. The four checks were built for claims that come out of chains with people in them. Every case in Section 7 is of that kind, and so is every failure in Section 5 except one (5.2, AI output). On an AI output the cause check cannot be run on the artifact at all, for the reasons Section 5.2 gives (the chain starts at a body of text, not at the world), and the framework reaches such an output only indirectly, by going behind it to the sources it was built from and running the checks there. That is a real restriction on the scope of what is claimed here. It is not a reason to abandon the instrument, but anyone using it on AI-generated material should know they are working one step removed.

10.2 Four moves

Not a checklist. What the framework asks of anyone doing this work.

Know which level you are on, and hold the claim with the confidence that level earns. A Reality claim presented as Fact claims to have seen what was actually worked out. A Constructed Picture presented as Reality claims that a long-travelled account is the thing itself. Both claim to stand closer to the ground than they do. The claim may be perfectly defensible where it actually sits. The question is not "is this correct" but "is this being

held at the level the architecture actually supports." Public failures of confident reasoning often fail this test before they fail any other.

Run the four checks honestly, and keep the cleared claim separate from the judgement. Honest running means refusing to say a check has been performed when the conditions for performing it were absent: when the evidence base was curated by interested parties and cause could not be tested; when the observers were reading from the same source and agreement was inherited; when the challenge was nominal; when the time needed to settle the question hasn't elapsed. A claim that clears all four is eligible. It is not yet something to stand under. The judgement is a separate act.

Refuse the substitution where it claims to deliver what it can't. Prediction markets, crowdsourcing and algorithmic reputation each do real work and each can feed a filter. None of them makes the judgement, and none is a substitute for it. Where the institutional architecture is expensive or missing, the temptation to treat a lateral output as a finished claim is strong, and the framework's job is to say exactly what is being given up.

Sustain the architecture where it exists, and act honestly where it doesn't. Where an institution has the five properties in working order, part of the practice is protecting them against the pressure to produce more, faster and cheaper, and the political pressure that comes with it. Where the architecture is partial, the practice is to say plainly that you are acting on a judgement call and to say what it rests on, rather than presenting a guess as a finding. Where the checks cannot run and a decision is unavoidable, the right response is not to issue a verdict in the filter's voice. It is to act, openly, with the conditions of the decision visible to anyone who wants to test it later.

10.3 What this closes and what it opens

The first arc (Attention Crisis, Wisdom Gap, AHI) diagnosed a problem and proposed an alternative. The Governance whitepaper specified the institutional form that alternative requires. This one specifies what those architectures are for: the claims they produce when they work, and the predictable way those claims degrade when they don't.

This is an opening move, not a conclusion.

Truth, in the load-bearing sense built here, is not a possession arrived at by procedure. It is what working architecture produces. The constraints can be named, the failures can be diagnosed, and the practice can be specified. What the framework cannot supply is the willingness to do the work in conditions where signing off would be easier and would pass every external test.

That willingness is what a person brings to the architecture, and what the architecture, working, builds in them over time. The two move together.

10.4 Where this paper itself stands

It would be poor practice to set out a test and quietly exempt the document setting it out. So, against its own four checks, plainly.

Cause has been run in part. The framework has been worked against three institutional cases, one of them chosen because it is a failure.

Independence has not been run, and nothing prevents it. This is one author's account. The readers who have pushed back on it so far were not independent of one another in the sense 6.1.2 requires, and I am aware that saying so is easier than fixing it.

Attack has been run only in its weaker form. The objections in the back matter are my own attempts to break the argument, which is not the same thing as somebody with a reason to break it doing so.

Time cannot be run, in the second of 6.1.6's three senses. A claim about how institutions behave is measured in decades, and this paper is months old. Nobody has obstructed the check and the claim is the right kind of thing for it. The years have simply not passed.

By 6.1.6 that leaves two checks partly run, one not run, and one that cannot yet be run. Which puts this argument some way short of load-bearing, and it is meant to. These are working papers. They are revised when they are shown to be wrong, and what changed is kept on the record.

What this one is asking for is the check it has not had.

Common objections

A serious argument invites serious challenge. These are the strongest lines of resistance, answered directly. Where a concession is owed, it is given.

1. "You have smuggled in a theory of truth while claiming not to."

The objection lands on the opening move. The paper defines a load-bearing claim as what an architecture produces, without defending that against the alternatives: correspondence, coherence, pragmatist accounts.

The concession is the framing itself, and the vocabulary now carries it. The fifth term is not "Truth." It is a *standing a claim earns*, which is a claim about warrant, meaning what entitles somebody to rely on it, rather than a claim about what truth is. Whether truth in some fuller sense exists beyond

what such architectures produce is a question this paper does not adjudicate.

One version deserves a direct answer: the solitary inferrer on a desert island, who works out a causal structure correctly from observation and acts on it. Does the framework deny that person has got something right? No. They may well be right, and may be acting wisely. What they cannot do alone is produce a claim with the standing described here: one that can be relied on by people who did not produce it, that holds up without their continuous defence, and that is testable by adversaries who do not share their assumptions. The framework is silent on whether their inference is correct. It speaks only to what an inference must pass through to do public work.

2. "Why these four checks? They are asserted, not derived."

Partial concession. The four are not derived from any deeper necessity.

What can be claimed is functional: each catches what the others miss, as 6.1.5 sets out. Could the same coverage come from a different partition? Plausibly: combining attack and independence under one heading, or splitting cause into intervention and counterfactual testing, would do similar work. The claim is not that *four* is a magic number. It is that the four failure modes the checks address (internal causal incoherence, observer-dependence, untested standing, and fragility over time) all have to be addressed. The number is an organising convenience. The failure modes are not.

3. "The judgement step is unfalsifiable: it explains every failure after the fact."

If the architecture is intact and things go badly, the framework says the judgement was not made. If things go well, it says it was. That is a theory that cannot lose.

The answer is to separate what the practitioner did from what the world did next, and to assess the judgement only on the first. Was the claim matched to the weight of the use? Were alternatives held open rather than closed early? Was the pressure working against the decision named rather than absorbed? Did the practitioner acknowledge the limits of their own knowledge? Those are observable while the decision is being made, and reconstructable afterwards from the record of the decision itself.

Run it on Kennedy. Suppose he chose the blockade, did everything the record shows, and the Soviets escalated anyway. The framework would not say the judgement failed. It would say the judgement was sound and the outcome was catastrophic. Those are different claims. Equally, someone who ratifies the loudest voice in the room, holds nothing open, and gets

lucky has not made the judgement. The good outcome is not evidence that they did.

The follow-on question is the sharper one: if the markers are observable, why is this not simply a fifth check?

Not because of infinite regress. A fifth check asking "were the first four actually runnable here?" would terminate perfectly well, and an earlier version of this paper rested on a regress argument that does not hold.

The real answer is simpler. The markers are specifiable; the judgement they issue in is not. A checklist can verify that somebody kept their options open. It cannot tell them which option to choose. That gap does not close by adding checks, because every further check is one more thing to run and none of them decides.

4. "The institutional cases are cherry-picked."

The NTSB and the Federal Reserve are unusually strong architectures. Fields where the architecture is partial, captured, or cannot exist at all (geopolitics, climate attribution, contested social science, ethics) might defeat the generalisation.

The empirical observation is accepted. The framework does not claim every field can host the architecture. It claims that where claims must be made at scale and relied on by those who did not make them, this specification is what their standing depends on. Where the architecture is partial or absent, what the framework supplies is the diagnosis: what is missing, what follows from its absence, and when acting on a stated prior is the honest move rather than issuing a verdict in the filter's voice.

Section 7.5 addresses the related worry that intact architecture is taken to guarantee good output. It does not, and that case is in the paper precisely because the claim is necessary-but-not-sufficient rather than guaranteed.

5. "'AI replaces mediation with construction' overstates the difference."

Human mediation was never transparent either. Editors' framing decisions were not inspectable; outlets had incentives readers could not see. Treating the shift as a change of kind overclaims.

On transparency this is conceded. The paper does not claim that human mediation was clear and AI is murky. Section 5.2 says as much, and 9.3 concedes the opacity point explicitly before arguing from it. What survives on that axis is a claim about the *locus* of opacity rather than its degree, together with three specific differences: scale, the absence of any addressable mediator, and the decoupling of confidence signals from warrant.

The paper does claim a change of kind, but on a different axis, and it is worth being clear which. The claim is not that the AI channel is more opaque than the old one. It is that the chain does not start at the ground: what went into the model was text, not the world. That is not a difference of degree, and no amount of transparency about the old channel touches it. Objection 10 takes up what follows.

6. "You concede the algorithmic curation evidence. Doesn't that sink the independence check?"

A fair challenge, and new to this version. Section 9.3 accepts that the best available studies find algorithmic feeds produce *more* cross-cutting exposure than users' own choices, and that forcing feed diversity produces no measurable attitudinal change. If the filter-bubble mechanism is not real, what is left of the claim that observer independence is eroding?

Two things, and both are better evidenced than what they replace.

The first is ownership concentration. When independent newsrooms consolidate, duplicated content rises sharply, measurably so, and stations in the same market broadcast substantially identical news. That is not a metaphor for "three newspapers running the same wire story are not three observers." It is that claim, counted.

The second is source displacement. A randomised experiment on X found the algorithmic feed promoted some content and demoted traditional media, shifting opinion over weeks. The mechanism is not that observers converge on one another. It is that the curation layer weights *away from* sources that run their own verification, which connects the independence check to the attack check rather than leaving it to stand alone.

The filter-bubble framing was the weaker argument. Losing it costs the paper nothing it was entitled to.

7. "Where is the theory of legitimacy?"

The lawyer's question. The NTSB and the Fed are authoritative not because they are epistemically well-designed but because they are democratically delegated, judicially reviewable, and constitutionally constrained. Who appoints the architects? What protects them from capture? What happens when their claims conflict with democratic outcomes?

Conceded, with the scope marked. Legitimacy is a separate axis and this paper does not address it. The governance paper, which this one presupposes, treats legitimate constitution as a precondition: the architecture must be authorised, the authorisation must be contestable, and it must be revisable. What this paper adds is that legitimate authorisation is not by itself sufficient for an institution to produce claims that hold their

standing. An institution can be perfectly legitimate and epistemically inadequate. The two conditions are complements, and treating either as a substitute for the other is the failure this paper argues against in another form.

8. "Where are the incentive distortions inside the filter?"

The behavioural economist's question. Practitioners are treated as more formed by their institutions than the evidence warrants. Status incentives distort attack: challengers carry career risk that defenders do not. Motivated reasoning shapes which observers get taken seriously. Reputational herding bends the filter toward whatever the consensus already holds.

Conceded, and located. Section 7.4 already says the architecture depends on incentive alignment it cannot generate. The behavioural failure modes are continuous with that rather than orthogonal to it: weakened checks are partly the consequence of career risk for dissent and throughput pressure on review. The Fed case in 7.5 shows one version of it directly, in a dissenter who described his own challenge as "a small incident" he did not press. A fuller treatment would model the behavioural layer explicitly. The narrower claim (that the architecture has to be built to hold against these pressures, including the behavioural ones) survives.

9. "You dropped the claim that public reasoning has declined. So what has changed?"

Also new to this version, and the honest answer matters.

An earlier draft argued that public causal reasoning had degraded. The trend data does not support it: the best available series on public scientific reasoning has been essentially flat since the early 1990s, and the item measuring understanding of experimental control has improved. The claim was dropped because it was not true, not because it was inconvenient.

What replaces it is more specific and better evidenced. The vulnerability is constant, not new: people systematically believe they understand mechanisms in far more detail than they do, and the illusion is precisely targeted: it appears for mechanisms and disappears for narratives. What has changed is the supply side. The volume and velocity of coherent, untested explanations competing for that permanently vacant space is not what it was.

A constant weakness meeting a changed environment is a sharper mechanism than general decline, and it requires no contested trend data to stand up.

10. "The framework is categorically unsuited to AI-generated claims."

The strongest version deserves stating in full. Large language models do not perform causal reasoning in Pearl's sense. They do not intervene on the world and they do not build causal accounts; they find patterns in text. If the first and most fundamental check systematically cannot run for AI-generated claims, not because of adversarial curation but because of what a large language model is, then the architecture is not merely strained. It is the wrong instrument for the job.

The premise is accepted in full, and Section 5.2 now states it rather than implying it. It is the reason Figure 3 is in the paper.

The conclusion does not follow, and the step it skips is the one 6.1.6 was written to supply. The objection assumes that a check which cannot run leaves the framework with nothing to say. It leaves the framework with three things to say: which of the three senses of *cannot be run* applies, what that costs the claim, and what the remaining route is. For AI output the answers are the third sense, the loss of any direct test of cause, and going behind the output to its sources. A framework that produces those three answers has not failed on the case. That is the work.

What the objection does establish, and what the paper now says outright, is that AI output is not simply a further degraded Picture. Treating it as one, as a longer version of the newspaper chain, would be the level-confusion of 5.1 committed by this paper rather than diagnosed by it. Section 10.1 records the limitation that follows.

Intellectual foundations and prior work

This paper does not claim to have invented institutional epistemology, the move from causal inference to warranted claims, or the diagnostic approach applied to public discourse. Each has a lineage.

On causal inference

Judea Pearl's framework, across *Causality* (2000) and *The Book of Why* with Dana Mackenzie (2018), supplies the account of the move from observation to inference. The three kinds of reasoning (watching, doing, imagining) are what Section 3.1 uses as the engine of that move.

Pearl is one of several frameworks. Donald Rubin's potential-outcomes approach, James Woodward's interventionist account, and Nancy Cartwright's capacities-based approach each formalise causal reasoning differently. Pearl is used here because those three kinds map cleanly onto the move from Fact to Reality and give the reader a vocabulary that carries through the argument. The structural claim, that Reality is built from Facts under causal discipline rather than narrative consistency, does not depend on Pearl being uniquely correct.

On Stoic epistemology

The Stoic distinction between the impression that arrives and the assent that is granted or withheld is the oldest version of this paper's central structure. An impression (*phantasia*) presents itself; assent (*sunkatathesis*) is a separate act, and the disciplined person withholds it from impressions that have not earned it. The filter and the judgement are that distinction, worked out institutionally.

Two things are worth saying about it. The Stoics were not relativists (that was the Academic Skeptics, their opponents) so the perspectival character of what we receive was for them the beginning of the problem, not the end of it. And Marcus Aurelius's instruction in *Meditations* 8.49 to "abide always by the first impressions and add nothing of your own from within" is a direct statement of the level-confusion this paper's Section 5.1 describes.

The line widely circulated as "everything we hear is an opinion, not a fact" is a modern condensation rather than a passage from the *Meditations*, and is cited here as such.

On claims as architectural output

The framework contributes to institutional epistemology, the tradition asking how knowledge is produced, tested and stabilised through structural conditions rather than individual cognition. Helen Longino's *The Fate of Knowledge* (2002) and John Hardwig's work on epistemic dependence develop the underlying claim: that warranted assertion in many fields depends on architectures of testing no individual could replicate. The philosophy of science since Kuhn and Lakatos has approached related questions through paradigm dynamics.

The framework does not adjudicate between theories of truth. Correspondence theories ground truth in relations between claims and world; coherence theories in the consistency of belief sets; pragmatist accounts in successful action under uncertainty. This paper is silent on those debates.

On shallow understanding and narrative substitution

The empirical grounding for Sections 3.1 and 5.3 comes from two findings. Rozenblit and Keil's work on the illusion of explanatory depth (2002) shows that people systematically overestimate how well they understand mechanisms, and, crucially, that the illusion is domain-specific: it appears for devices and natural phenomena and disappears for narratives and procedures. Keil's later work (2010) supplies the corrective, arguing that abstract causal sensing and the ability to locate expertise largely survive; the deficit is in mechanistic detail.

Pennington and Hastie's Story Model (1988) supplies the demonstration that narrative coherence does work validity should be doing: mock jurors given identical evidence convicted at 78% when the prosecution's case was ordered as a story and at 31% when it was not.

On information aggregation

Friedrich Hayek's "The Use of Knowledge in Society" (1945) is the foundational argument for price as an information-aggregating structure. Nassim Taleb's work on skin in the game supplies the discipline of consequence-bearing belief. Section 8's treatment of prediction markets combines both, and its contribution is to specify what such aggregation produces, inputs to a filter, and what it does not.

On institutional architecture

The NTSB's operating model is documented in its own *Aviation Investigation Manual*. The Federal Reserve's structure is treated in Allan Meltzer's history, and the Volcker disinflation in Volcker's own *Keeping at It* (2018) and William Silber's biography. The 2006 case draws on the Financial Crisis Inquiry Commission's final report, on the FOMC transcripts released in 2012, and on Fligstein, Brundage and Schultz's analysis of all 72 scheduled FOMC meetings between 2000 and 2008.

On Knightian uncertainty

Frank Knight's distinction between risk, where probability distributions can be specified, and uncertainty, where they cannot (1921), grounds the third of the four things the filter leaves undone.

Prior work by this author

This paper is the fifth in a connected series. *AHI: First Principles* (2026) gives the foundational argument for Augmented Human Intelligence as the alternative to the AGI scaling thesis. *The Wisdom Gap* (2026) develops the argument that current architectures are categorically incapable of producing wisdom, and introduces the constraint-awareness framing this paper extends. *The Attention Crisis* (2025) situates the work within attention as a finite resource. *AI Governance and the Architecture of Practice* (2026) specifies the institutional form, and is the immediate predecessor to this paper. The supporting essay "AHI From the Inside" (2026) develops the System 1 / System 2 distinction at the practitioner level. Earlier work includes *Beyond Scale: Towards Biologically Inspired Modular Architectures* (2025) and "LLMs, Synthetic Wisdom and Bias" (2024).

References

Ananny, Mike and Crawford, Kate. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." *New Media & Society* 20(3), 2018, 973-989.

Arguedas, Amy Ross, Robertson, Craig, Fletcher, Richard and Nielsen, Rasmus Kleis. *Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review*. Reuters Institute for the Study of Journalism, 2022.

Baker, Dean. "The Run-Up in Home Prices: Is It Real or Is It Another Bubble?" Center for Economic and Policy Research, August 2002.

Bakshy, Eytan, Messing, Solomon and Adamic, Lada. "Exposure to ideologically diverse news and opinion on Facebook." *Science* 348(6239), 2015, 1130-1132.

Bennett, Max S. *A Brief History of Intelligence*. Mariner Books, 2023.

Bernanke, Ben S. "Asset-Price 'Bubbles' and Monetary Policy." Remarks before the New York Chapter of the National Association for Business Economics, October 15, 2002.

Bernanke, Ben S. "The Great Moderation." Remarks at the meetings of the Eastern Economic Association, Washington, D.C., February 20, 2004.

Bernanke, Ben S. and Gertler, Mark. "Should Central Banks Respond to Movements in Asset Prices?" *American Economic Review* 91(2), 2001, 253-257.

Bratsberg, Bernt and Rogeberg, Ole. "Flynn effect and its reversal are both environmentally caused." *PNAS* 115(26), 2018, 6674-6678.

Cartwright, Nancy. *Nature's Capacities and Their Measurement*. Oxford University Press, 1989.

Clark, Andy. *The Experience Machine*. Pantheon, 2023.

Demyanyk, Yuliya and Van Hemert, Otto. "Understanding the Subprime Mortgage Crisis." *Review of Financial Studies* 24(6), 2011.

European Broadcasting Union and BBC. *News Integrity in AI Assistants*. October 2025.

Financial Crisis Inquiry Commission. *The Financial Crisis Inquiry Report*. U.S. Government Printing Office, 2011.

Fletcher, Richard, Kalogeropoulos, Antonis and Nielsen, Rasmus Kleis. "More diverse, more politically varied: How social media, search engines and aggregators shape news repertoires in the United Kingdom." *New Media & Society* 25(8), 2021.

Fligstein, Neil, Brundage, Jonah Stuart and Schultz, Michael. "Why the Federal Reserve Failed to See the Financial Crisis of 2008: The Role of 'Macroeconomics' as a Sense-Making and Cultural Frame." Institute for Research on Labor and Employment, University of California Berkeley, 2014.

Garz, Marcel and Ots, Mart. "Media consolidation and news content quality." *Journal of Communication* 75(3), 2025, 195–206.

Gauthier, Maxime, Hodler, Roland, Widmer, Philine and Zhuravskaya, Ekaterina. "The political effects of X's feed algorithm." *Nature* 652(8109), 2026, 416–423.

Ghafoori, Bijean et al. "Epistemic Integrity in Large Language Models." arXiv:2411.06528, 2024.

Gilardi, Fabrizio et al. "Quality perceptions and intended engagement in response to AI-generated and AI-assisted news." *Scientific Reports*, 2026.

González-Bailón, Sandra et al. "Asymmetric ideological segregation in exposure to political news on Facebook." *Science* 381(6655), 2023, 392–398.

Gramlich, Edward M. "Booms and Busts: The Case of Subprime Mortgages." Federal Reserve Bank of Kansas City symposium, Jackson Hole, August 2007.

Greenspan, Alan. "The Challenge of Central Banking in a Democratic Society." Francis Boyer Lecture, American Enterprise Institute, December 5, 1996.

Greenspan, Alan. "Economic Volatility." Federal Reserve Bank of Kansas City symposium, Jackson Hole, August 30, 2002.

Guess, Andrew et al. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science* 381(6656), 2023, 398–404.

Hanson, Mark A. et al. "The strain on scientific publishing." *Quantitative Science Studies* 5(4), 2024, 823–843.

Hardwig, John. "Epistemic Dependence." *Journal of Philosophy* 82(7), 1985, 335–349.

Harvard Joint Center for Housing Studies. "Price-to-Income Ratios are Nearing Historic Highs."

Hayek, Friedrich A. "The Use of Knowledge in Society." *American Economic Review* 35(4), 1945, 519–530.

Hoffman, Donald. *The Case Against Reality*. W.W. Norton, 2019.

James, William. *The Principles of Psychology*. Henry Holt, 1890.

- Jaźwińska, Klaudia and Chandrasekar, Aisvarya. "AI Search Has a Citation Problem." Tow Center for Digital Journalism, *Columbia Journalism Review*, March 2025.
- Job, Macarthur. *Air Disaster*, Volumes 1–4. Aerospace Publications, 1994–2001.
- Kahneman, Daniel. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- Keil, Frank C. "The feasibility of folk science." *Cognitive Science* 34(6), 2010, 826–862.
- Kennedy, Robert F. *Thirteen Days: A Memoir of the Cuban Missile Crisis*. W.W. Norton, 1969.
- Knight, Frank H. *Risk, Uncertainty, and Profit*. Houghton Mifflin, 1921.
- Kuhn, Thomas S. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.
- Lakatos, Imre. "Falsification and the Methodology of Scientific Research Programmes." In Lakatos and Musgrave (eds.), *Criticism and the Growth of Knowledge*. Cambridge University Press, 1970.
- Longino, Helen E. *The Fate of Knowledge*. Princeton University Press, 2002.
- Lorenz-Spreen, Philipp, Mønsted, Bjarke Mørch, Hövel, Philipp and Lehmann, Sune. "Accelerating dynamics of collective attention." *Nature Communications* 10, 2019, 1759.
- Maconochie, James. "LLMs, Synthetic Wisdom and Bias." LinkedIn, 2024.
- Maconochie, James. *Beyond Scale: Towards Biologically Inspired Modular Architectures for Adaptive AI*. jamesmaconochie.com, 2025.
- Maconochie, James. *The Attention Crisis*. jamesmaconochie.com, 2025.
- Maconochie, James. *AHI: First Principles*. jamesmaconochie.com, 2026.
- Maconochie, James. *The Wisdom Gap*. jamesmaconochie.com, 2026.
- Maconochie, James. "AHI From the Inside." jamesmaconochie.com, 2026.
- Maconochie, James. *AI Governance and the Architecture of Practice*. jamesmaconochie.com, 2026.
- Magesh, Varun et al. "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools." *Journal of Empirical Legal Studies*, 2025.
- Marcus Aurelius. *Meditations*. Book 8, 49.
- Martin, Gregory J. and McCrain, Joshua. "Local News and National Politics." *American Political Science Review* 113(2), 2019, 372–384.

May, Ernest R. and Zelikow, Philip D. *The Kennedy Tapes: Inside the White House During the Cuban Missile Crisis*. Belknap Press, 1997.

Medill Local News Initiative. *The State of Local News 2025*. Northwestern University, October 2025.

Melumad, Shiri and Yun, Jin Ho. "Experimental evidence of the effects of large language models versus web search on depth of learning." *PNAS Nexus* 4(10), 2025.

Meltzer, Allan H. *A History of the Federal Reserve*, Volumes 1 and 2. University of Chicago Press, 2003 and 2009.

National Science Board. *Science and Engineering Indicators: Science and Technology: Public Attitudes and Understanding*. NSB-2020-7 and NSB-2022-7.

National Transportation Safety Board. *Aviation Investigation Manual*.

Nordmo, Magnus, Nøttestad Norrøne, T. and Lang-Ree, O. C. "Reevaluating the Flynn effect, and the reversal." *Intelligence* 110, 2025, 101909.

Nyhan, Brendan et al. "Like-minded sources on Facebook are prevalent but not polarizing." *Nature* 620, 2023, 137-144.

Open Science Collaboration. "Estimating the reproducibility of psychological science." *Science* 349(6251), 2015.

Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

Pearl, Judea and Mackenzie, Dana. *The Book of Why*. Basic Books, 2018.

Pennington, Nancy and Hastie, Reid. "Explanation-based decision making: Effects of memory structure on judgment." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(3), 1988, 521-533.

Pew Research Center. "U.S. newsroom employment has fallen 26% since 2008." 2021.

Pew Research Center. "Google users are less likely to click on links when an AI summary appears in the results." July 2025.

Reinhart, Alex et al. "Do LLMs write like humans? Variation in grammatical and rhetorical styles." *PNAS* 122(22), 2025.

Reuters Institute for the Study of Journalism. *Generative AI and News Report 2025*.

Reuters Institute for the Study of Journalism. *Digital News Report 2026*.

Rozenblit, Leonid and Keil, Frank. "The misunderstood limits of folk science: an illusion of explanatory depth." *Cognitive Science* 26(5), 2002, 521-562.

Rubin, Donald B. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5), 1974, 688–701.

Sapolsky, Robert. *Behave*. Penguin Press, 2017.

Silber, William L. *Volcker: The Triumph of Persistence*. Bloomsbury, 2012.

Stock, James H. and Watson, Mark W. "Has the Business Cycle Changed and Why?" *NBER Macroeconomics Annual* 17, 2002.

Taleb, Nassim Nicholas. *Antifragile*. Random House, 2012.

Taleb, Nassim Nicholas. *Skin in the Game*. Random House, 2018.

Tooze, Adam. *Crashed*. Viking, 2018.

Volcker, Paul A. with Harper, Christine. *Keeping at It*. PublicAffairs, 2018.

Weisberg, Deena Skolnick, Keil, Frank C., Goodstein, Joshua, Rawson, Elizabeth and Gray, Jeremy R. "The seductive allure of neuroscience explanations." *Journal of Cognitive Neuroscience* 20(3), 2008, 470–477.

Woodward, James. *Making Things Happen*. Oxford University Press, 2003.

Yanich, Danilo and Bagozzi, Benjamin E. *Reusing the News*. SNF Ithaca Research Report, Biden School of Public Policy and Administration, University of Delaware, August 2025.

Zhou, Kaitlyn, Hwang, Jena D., Ren, Xiang and Sap, Maarten. "Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty." arXiv:2401.06730, ACL 2024.