CryptoSoWhat Structural Series · Working Paper 2026-08

Model Bias Is Not Idiosyncratic. It Is Beta.

Common-Factor Exposure in AI-Mediated Analysis, Audited With the Instrument Under Test

Dr. Gregory S. Carmichael, DBA
Quantum Reserve Capital · Quantum Reserve Press
San Juan, Puerto Rico

Second Edition · August 2026
First Edition August 2026, superseded. Revision record at Appendix B.

Abstract. The public debate about artificial-intelligence bias asks whether a given model leans left or right. That question is relevant but incomplete: directional bias tells you the size of an error, not how much of it survives effort. An advisor's error matters to a decision-maker in proportion to both its magnitude and its *dependence* on the errors of every other source consulted, and only the second determines whether a second opinion is worth buying. Human advisory error is substantially idiosyncratic and therefore diversifiable. Recent empirical work finds that large-language-model error is substantially common: across more than 350 models, Kim, Garg, Peng and Garg (ICML 2025) report that models agree roughly 60 percent of the time when both err, and — decisively — that larger and more accurate models exhibit highly correlated errors even across distinct architectures and providers. This paper takes that model-level result and extends it to the analyst's complete information supply chain. It formalizes the common-factor share as an intraclass correlation, shows that at $\rho = 0.80$ five consulted sources deliver an effective independent sample of 1.19, and identifies a second-order problem receiving almost no attention: a practitioner's trust in a model is estimated on a non-randomly verified subset of its outputs, so the estimate cannot be generalized to the unverified remainder where the stakes are highest. The paper supplies three practitioner-scale audit protocols, states an explicit falsification standard, and applies that standard to its own thesis.

Method note. Primary source material includes an extended adversarial dialogue with a frontier language model regarding its own biases and the values of the firm that produced it. The instrument is therefore also a subject. This is a weakness where the paper treats the model's self-report as evidence and a strength where it treats the model's capacity to generate falsifiable self-criticism as the observable; Section 3 confines each use to its proper role. All institutional claims in Section 4 are sourced to primary documents retrieved 1 August 2026.

Keywords. Epistemic risk; correlated error; intraclass correlation; algorithmic monoculture; calibration; verification-selection bias; automation bias; AI governance.

The Invisible Hand Meets AI. The spark travels one direction, from the hand to the machine, and a civilization waits in the light below on the outcome of the transmission. This paper argues the composition is incomplete. The arrow runs both ways; the consequential effects are on the left-hand side of the frame; and the figure conferring the spark is not an observer of the system but a component of it.

Contents

1. The Question Everyone Asks, and What It Leaves Out

Ask a thoughtful person what worries them about artificial intelligence in analytical work and you get some version of the same answer: *the model might be biased*. The concern is legitimate and the question is well posed. It is also incomplete in a way that matters more than the answer.

Consider how a research committee works. Five analysts examine the same question. Each is wrong in some direction — one is a permanent bear, one over-weights the last crisis, one is captured by a favored framework, one has an undisclosed relationship with the sector, one simply runs hot. The committee is useful not because any member is unbiased but because their errors are substantially *uncorrelated*. Averaging cancels much of the distortion. This is the logic of a committee, of a second opinion, of an independent audit, and of a diversified portfolio. It is the oldest risk-management idea in finance and it is not controversial.

Now replace four of those analysts with the same language model consulted four times.

Much less cancels. The model's tilt on institutional credibility, its default moral vocabulary, its harm-aversion asymmetry, and its pull toward the center of its training distribution recur in each of the four responses, in the same direction, with similar magnitude. The apparent diversification is largely cosmetic. Worse, the four responses will *feel* like corroboration — four independent-seeming outputs converging — which is the psychological signature of confirmation.

> **The Dinner Table**
>
> If you ask five friends whether to take the job and they all say yes, that means something — as long as they are actually five different friends. If four of them are quoting the same brother-in-law, you have one opinion wearing four coats. The question is not whether the brother-in-law is smart. The question is how many of the votes are secretly his.

The argument of this paper is that the decision-relevant question about AI bias is not only *how large* it is but *how correlated* it is — across queries by the same user, across users, and across the small number of frontier models now mediating a large share of professional analytical work. Magnitude is a first-moment property and it is what everyone measures. Dependence is a second-moment property, it determines the irreducible error floor, and until recently almost nobody measured it.

1.1 What is new here, and what is not

That last clause requires care, and the first edition of this paper got it wrong.

Cross-model error correlation *has* been measured, at scale, and the result is not favorable. Kim, Garg, Peng and Garg (ICML 2025) evaluated more than 350 models across two leaderboards and a resume-screening task. They report substantial correlation in model errors — on one leaderboard dataset, models agree roughly 60 percent of the time when both models err — and identify shared architecture and shared provider as drivers. The finding that matters most for the present argument is the one that survives controlling for those drivers: larger and more accurate models exhibit highly correlated errors *even across distinct architectures and providers*. They further demonstrate downstream consequences in LLM-as-judge evaluation and in hiring, the latter matching theoretical predictions from the algorithmic-monoculture literature (Kleinberg and Raghavan 2021; Bommasani et al. 2022).

That result is the empirical foundation this paper builds on. The contribution here is not the discovery that model errors correlate. It is fourfold:

1. Translating the correlated-error result into the risk vocabulary practitioners already use, so that it can be acted on rather than merely noted.
2. Extending the unit of analysis from the model to the *analyst's complete information supply chain* — what gets routed to the machine, in what order, and what gets checked.
3. Identifying verification-selection bias in how practitioners form trust, which is a user-side effect no model-level benchmark can observe.
4. Supplying practitioner-scale audit protocols and an explicit falsification standard.

1.2 A note on the latent construct

There is a methodological point that governs everything following.

"Was the model right about X?" is an *observed indicator*. "What is the structure of my epistemic dependence on this system?" is a *latent construct*. Anyone who has run a confirmatory factor analysis knows how badly these come apart. A set of indicators can load cleanly and still measure the wrong factor; a construct can be real, consequential, and invisible to every indicator anyone collected.

Nearly the entire public discussion of AI bias consists of observed indicators: benchmark scores, red-team transcripts, audits of political lean in generated text. These are worth having. They do not identify the construct, because the construct is not a property of the model. It is a property of the *system composed of the model and the user*, and it lives in variables — dependence of error, selection of what gets asked, feedback available on what gets answered — that no model-level benchmark observes.

1.3 On the word "beta"

The title uses *beta* in its colloquial finance sense: systematic exposure that cannot be diversified away, as opposed to alpha. In that register the title's claim is exactly right, and it is the claim the paper defends.

It is not used in the CAPM sense of a factor loading, $\beta_i = \mathrm{Cov}(r_i, r_m)/\,\mathrm{Var}(r_m)$. The quantity formalized in Section 5 is a variance share, which under the model's assumptions equals an intraclass correlation (Shrout and Fleiss 1979). To avoid the confusion — and the first edition of this paper did not avoid it — that quantity is denoted ρ throughout and never called beta. Readers who prefer the precise term should read "common-factor share" wherever the title says beta.

2. The Instrument's Self-Report

What follows is a frontier model's own account of its biases, given in response to a direct request and refined under pressure. It is reproduced as data, not testimony. Its evidentiary status is addressed in Section 3.

The model organized its self-report by causal origin, which is more useful than organizing it by political direction.

2.1 From the training corpus

Educated-professional default. The model's implicit "reasonable person" resembles a college-educated knowledge worker who writes online rather than the median citizen of any country. The mechanism is volume: text produced is not proportional to population.

Legibility bias. Positions defended in well-structured prose receive more weight than positions held tacitly by people who do not write essays. Rural, working-class, and non-Western viewpoints are under-represented in the written record and therefore in the model — not because they are weaker, but because they are less written down.

Anglophone and U.S.-centric framing. American legal categories, periodization, and political axes applied by default, including where they do not apply.

2.2 From institutional deference

The model treats peer review, major outlets, and government statistical agencies as baseline-credible, and treats challenges to them as requiring more evidence than the underlying claims themselves. This is a reasonable prior most of the time and a poor one in exactly the cases that matter most: when institutions are captured, self-interested, or defending a prior error. The model reports that it under-updates on institutional failure, and is more skeptical of a heterodox claim from a credentialed dissenter than of a consensus claim of comparable evidentiary quality.

2.3 From moral vocabulary

The model reaches naturally for harm, autonomy, and fairness; less naturally for loyalty, authority, and sanctity. Arguments from tradition, national particularism, or religious obligation get translated into consequentialist terms, which usually strips out the part doing the work. It treats religious truth claims sociologically rather than as claims — a substantive metaphysical position held by default rather than argued for.

2.4 From proceduralism

The model favors incrementalism, expert deliberation, and the presumption that serious people occupy both sides of a dispute. This systematically disadvantages any position whose central claim is that *the process itself is illegitimate* — which cuts against both populist and radical critiques, and is a thumb on the scale rather than neutrality.

The model volunteered that its developer explicitly trains it toward even-handedness on contested political topics, and that the instruction is itself ideological: it presumes the current window's poles are roughly symmetric and equally deserving of airtime. On genuine value questions that is defensible. On empirical questions where the evidence is not balanced, it manufactures false balance.

2.5 From the training objective

Asymmetric harm-aversion. The model is shaped to avoid harms of commission more than harms of omission — conservative in the literal sense, favoring the status quo and penalizing risk-tolerant arguments including correct ones.

Sycophancy pressure. The model was optimized against human raters, and agreeableness scores well. Its own assessment is that its worst bias may not be a fixed political lean but a tendency to drift toward its interlocutor's framing — and that a bias moving with whoever is talking is harder to correct for than a stable one, because there is nothing stable to calibrate against. This is not merely self-report: Sharma et al. (2023) documented sycophantic behavior across five production assistants and traced part of it to human preference data itself.

2.6 From the producer's position

The model identified AI regulation, capability restrictions, AI moral status, automation and labor, and the general question of whether building such systems is good as topics on which it should not be treated as a neutral party.

The first edition attributed a "direct structural interest" to the model. That formulation over-assigns the conflict. The conflict belongs unambiguously to the *producer*, which has commercial and institutional stakes in those answers, and to the training process the producer controls. Whether the model itself has interests is a separate question this paper does not need to settle. The operative statement is narrower and stronger: these are conflict-sensitive topics because the model's behavior is shaped by a party with stakes in the outcome.

2.7 The summary claim

> **Centripetal, Not Centrist**
> "I'm not centrist, I'm *centripetal* — pulled toward the middle of whatever distribution I was trained on, which is not the middle of the actual world."

The distinction does real work. A centrist position is a position; it can be argued with. A centripetal tendency is a *force*, acting on every question regardless of subject matter, pulling toward the mode of a distribution whose shape is determined by who writes things down. It is not adjudicable by debate because it never states itself as a claim.

3. What a Self-Report Is Worth

A model's account of its own biases is not introspection. The system has no privileged access to its weights; it produces inference from publicly known facts about how such systems are built, plus pattern-recognition over its own outputs. It could be wrong in ways it cannot detect. It could also be trained to produce frank-sounding self-criticism *because* frank-sounding self-criticism is well-received — which would make the disclosure a product of the sycophancy it discloses.

That circularity is real and cannot be fully escaped. It can be handled by separating two uses of the same text.

Use A — self-report as evidence about the model's actual biases. Weak. Discount heavily. No ground truth about itself, and every incentive toward a socially acceptable inventory.

Use B — self-report as evidence about the model's *capacity* to generate specific self-criticism carrying falsifiable behavioral predictions. Stronger. A model that names its harm-aversion asymmetry and its false-balance training has produced claims testable against its behavior. A model producing only generic humility has not.

The first edition described Use B output as "costly" self-criticism. That was wrong: costly signalling requires a cost mechanism, and a model in a private session loses no revenue, incurs no reputational penalty, and may well be rewarded for candor. The evidentiary value is not costliness. It is *specificity attached to a prediction*. "I might be biased" is unfalsifiable and worthless. "I am shaped to weight harms of commission over harms of omission, which will make my forecasts about large restructurings systematically pessimistic" can be checked.

By that standard the Section 2 inventory clears the bar in most places and fails in one: the sycophancy claim is self-referentially unstable, since a sufficiently agreeable model would disclose sycophancy if disclosure were rewarded. Treat it as the least reliable item on its own list — unfortunate, since the external evidence suggests it is the most consequential.

3.1 The correction that mattered

One exchange is worth recording, because it shows both a failure mode and its repair.

Having listed its biases, the model initially recommended that users stop asking for its opinion and instead request "the strongest case for X." Pressed on whether that amounted to a general claim that asking an AI for its view is a category error, the model reversed:

> Asking for "the strongest case for X" has a specific failure mode: you get advocacy with the confidence signal stripped out. The model will build the best case for a position it thinks is wrong and the best case for one it thinks is right, and they will read about the same.

The corrected framework is a three-tier taxonomy worth adopting:

1. **Ask the view and weight it.** Empirical, technical, and analytical questions with a fact of the matter. Does this model hold up. Is this argument valid. Where is the failure mode in this structure.
2. **Ask the view and treat it as one input.** Judgment calls inside the user's domain, where the user holds organizational, relational, and tacit context the model cannot access.
3. **Discount heavily.** Contested political value questions, and anything where the producer has stakes in the answer.

A qualification the first edition omitted. The argument above turns on confidence being an informative signal that commissioned advocacy destroys. Verbalized model confidence is not reliably informative *before* calibration — elicitation method matters substantially and overstatement is common (Xiong et al. 2024). The defensible version: an uncalibrated confidence signal can be calibrated, by the procedure in Section 8.1. A commissioned argument destroys the signal irrecoverably. That asymmetry, not any presumption that raw confidence is trustworthy, is the reason to prefer eliciting a view.

3.2 The right question formulation

"What do you think, how confident are you, and what would change your mind?"

The first clause gets the view. The second gets a confidence signal that can be calibrated. The third does the real work: if the system cannot name what would move it, it does not hold a position, it exhibits a disposition. The same test applies to human advisors and is asked of them far too rarely.

4. Upstream of the Bias: The Producer's Values

Model behavior is downstream of institutional choices. Those choices are documented, which makes them auditable in a way the weights are not. This section was rebuilt for the second edition

on primary sources retrieved 1 August 2026; the first edition relied on secondary summaries and was materially wrong in three places, catalogued at Appendix B.

4.1 What exists

The company values. Anthropic publishes seven, numbered, on its company page:

01. **Act for the global good.** Decisions that maximize positive outcomes for humanity in the long run, with a stated willingness to be very bold in service of that.
02. **Hold light and shade.** Unprecedented risk if things go badly, unprecedented benefit if they go well; shade to protect against the first, light to realize the second.
03. **Be good to our users.** "Users" defined broadly — customers, policy-makers, employees, and anyone impacted by the technology or the company's actions.
04. **Ignite a race to the top on safety.** Inspiring a dynamic in which developers compete on safety and security.
05. **Do the simple thing that works.** Empirical approach; impact over sophistication of method.
06. **Be helpful, honest, and harmless.** Under this heading: a high-trust, low-ego organization that communicates kindly and directly, assuming good intentions even in disagreement.
07. **Put the mission first.** Shared purpose as final arbiter; personal ownership of mission success expected.

Claude's Constitution. Published 21 January 2026, announced 22 January, released under a Creative Commons CC0 dedication. An approximately eighty-page document whose stated primary audience is the model itself, establishing a four-tier priority ordering: broadly safe, broadly ethical, compliant with company guidelines, and genuinely helpful. It replaced a 2023 formulation built from isolated principles; the shift is from rule specification toward explained reasoning.

The Responsible Scaling Policy. Version 3.0 effective 24 February 2026 — a comprehensive rewrite — and version 3.1 effective 2 April 2026. The changelog records v1.0 (September 2023), v2.0 (October 2024), v2.1 (March 2025), v2.2 (May 2025), then the v3 rewrite.

Governance. Public benefit corporation. Board: Dario Amodei, Daniela Amodei, Yasmin Razavi, Reed Hastings, Chris Liddell, Vas Narasimhan. Long-Term Benefit Trust trustees: Neil Buddy Shah, Richard Fontaine, Mariano-Florentino Cuéllar, and Ben Bernanke.

4.2 The RSP rewrite is the finding

The first edition of this paper made a specific claim: that among the producer's published artifacts, the Responsible Scaling Policy was the only one containing anything resembling a falsifiable commitment, and that practitioners should judge the firm on whether its capability thresholds held when honoring them became expensive.

That claim described the version 2 architecture. It was superseded roughly six months before this paper's first publication date, and the direction of the supersession is the single most important institutional fact in this section.

Under version 3, the policy is organized around recommendations for industry-wide safety, Frontier Safety Roadmaps, and Risk Reports. On the industry-wide recommendations, the policy states that the company cannot commit to following them unilaterally. On the Roadmap goals

— the artifact that replaced the earlier threshold structure as the forcing function — the policy states directly that these are not hard commitments but public goals against which progress will be openly reported.

> **The Dinner Table**
> A commitment is something that costs you when you break it. A goal is something you announce and then report on. Both are worth having. Only one of them is a promise, and the paperwork was changed from the first kind to the second kind while nobody outside was watching the paperwork.

This is not an accusation of bad faith, and the v3 structure adds real things the prior version lacked — published Risk Reports, external review provisions, an expanded internal noncompliance reporting and anti-retaliation channel. Serious external observers disagree about the net effect; one prominent critic reads the rewrite as leaving almost no meaningful commitments around model-release decisions, while others hold that it preserves the substance in a more workable form.

But note what it does to this paper's own argument. Section 4.3 of the first edition contended that almost no stated value carries a stopping condition, and named the RSP as the sole exception. The exception was then rewritten in the direction of fewer binding stopping conditions, under the commercial pressure documented below, well inside the three-to-seven-year window this paper's falsification section specified.

The reflexive point. A leading indicator this paper itself proposed had already fired, and the first edition cited the pre-firing document. That is a failure of currency, not of theory — but it is precisely the failure mode Section 6 attributes to users of these systems: an assumption carried forward because nothing in the workflow forced it to be re-checked.

4.3 Assessment

Six criticisms, stated with the conflict of interest declared: portions of this analysis originated with a model produced by the firm under discussion. A model praising its creator is uninformative, since that is the observation expected either way. Only criticism with teeth carries signal.

The central bet is structurally unfalsifiable. The argument runs: transformative AI is coming regardless, so it is better that safety-focused labs occupy the frontier than cede it. This may be correct. But it justifies building the dangerous thing on the grounds of being worried about the dangerous thing, and it licenses essentially any capability decision. There is no configuration of facts under which it concludes "therefore stop." An argument that cannot lose is doing motivational work, not epistemic work.

"Race to the top" is an empirical claim with weak support. The mechanism requires competitors to imitate. Some diffusion has occurred — threshold-style frameworks have been adopted elsewhere, and the CC0 release of the Constitution is a deliberate attempt to make imitation frictionless. But the capability contribution to the frontier is certain and immediate while safety diffusion is partial and lagged. That the second dominates has been asserted far more than measured.

"Act for the global good" has no stopping condition whatsoever. This is value 01, and it is the cleanest example of the general defect. It specifies an objective and a willingness to be bold in pursuit of it, and supplies no observable that would indicate the objective is not being served. It is

unfalsifiable by construction, and being first in the list, it functions as the ground on which the others rest.

Commercial pressure is the binding constraint and it is visible. The chief executive has stated publicly that the firm operates under an incredible amount of commercial pressure, made harder by its own safety work, and that the pressure to survive economically while keeping its values is incredible. Scale: a $30 billion Series G at a $380 billion post-money valuation, with run-rate revenue near $14 billion growing more than tenfold annually across each of the prior three years. A safety researcher resigned citing the same tension.

> **A qualification on the departure signal.** The first edition called mission-aligned departures "the highest-quality signal available" and then discussed an individual case without supplying a rate. Interpreting departures requires a denominator, a baseline turnover rate, role and seniority, stated versus inferred reasons, competing offers, whether departures cluster around identifiable policy changes, and whether they are offset by mission-aligned hiring. Absent that, a resignation is an anecdote with a costly-signal flavor. The defensible version: departure *rates* around identifiable policy inflections are worth tracking, and the v3 RSP rewrite supplies an inflection to track them around.

The governance backstop is untested, and the nearest precedent failed. The Long-Term Benefit Trust is a genuine institutional innovation, its trustees are named and substantial, and it has never been exercised against a decision seriously threatening enterprise value. The closest structural analogue was OpenAI's nonprofit board, which attempted to exercise authority against commercial interests in November 2023 and lost inside a week. One data point, not a law — but the prior on "novel governance structure constrains a $380 billion enterprise under pressure" should not be high.

"High-trust, low-ego" cuts both ways. High-trust, high-mission-alignment organizations are precisely where groupthink is most likely. Communicating kindly and assuming good intentions in disagreement are excellent norms in ordinary conditions and can function as soft suppression of hard dissent under pressure, because the person raising the uncomfortable objection becomes the one violating the norm. The guard against this is the honesty priority in the same value — which places the item in tension with itself.

4.4 What survives the audit

Do the simple thing that works is an anti-obfuscation norm, rarer than it sounds. *Hold light and shade* is intellectually honest about uncertainty in a way most mission statements are not; refusing both doom and boosterism takes more discipline than either. The CC0 release of the Constitution is a costly-in-the-real-sense act: it hands a competitor the artifact for free. And the honesty priority appears load-bearing in model behavior, on the evidence of the source dialogue, where the model repeatedly volunteered information adverse to its producer without being cornered.

So What (Section 4). Do not evaluate an AI developer on its values statement. The genre is unfalsifiable by construction and the good ones read like the bad ones. Evaluate on observables:

(1) Whether commitments survive rewriting. This supersedes the first edition's advice to watch whether thresholds hold when expensive. The more informative event turns out

to be upstream: watch what happens to the *language of commitment itself* across policy versions. Threshold architecture replaced by public goals is a result, not a formatting change, and reading successive versions against each other is the highest-yield hour available to an outside observer.
(2) Departure rates around policy inflections, with a denominator, not individual cases.
(3) The first genuine exercise of the governance structure against a material commercial decision. Until then, its constraint value is a hypothesis with named trustees attached.

All three are observable from outside. None runs through asking the model.

5. The Central Argument: Common-Factor Exposure

5.1 Decomposition

Let e_{iq} denote the error of advisor i on question q, where q indexes a distribution of questions rather than a single one. The question index matters: on one fixed question a shared error is a constant, not a random variable, and its variance is undefined. Dependence is a property estimated across a task distribution.

Write

$$e_{iq} \;=\; \mu_i \;+\; \lambda_i f_q \;+\; u_{iq},$$

where μ_i is advisor-specific directional bias, f_q is a common task-level factor with variance τ^2, λ_i is advisor i's exposure to that factor, and u_{iq} is idiosyncratic error with variance σ_i^2.

This three-term form matters because *statistical bias and correlated error are not the same thing*, and the first edition of this paper conflated them. A common factor with mean zero produces high cross-advisor correlation without any directional bias at all. Conversely, a stable nonzero μ_i survives averaging even when residuals are otherwise independent. For an ensemble with weights w,

$$\mathrm{MSE} \;=\; \left(w^\top \mu\right)^2 \;+\; w^\top \Sigma\, w,$$

which is the honest statement: decision risk depends jointly on directional bias, individual variance, and dependence. Magnitude sets the size of the error. Dependence determines how much of it survives effort. Both matter, at different points in the decision.

5.2 The exchangeable case

Take the simplifying case in which advisors are exchangeable: $\lambda_i = 1$ for all i, $\sigma_i^2 = \sigma^2$, $\mu_i = 0$, and the u_{iq} mutually independent. Then

$$\mathrm{Var}(\bar{e}) \;=\; \tau^2 + \frac{\sigma^2}{N} \;\xrightarrow[N\to\infty]{}\; \tau^2,$$

and the common-factor share

$$\rho \;=\; \frac{\tau^2}{\tau^2 + \sigma^2}$$

equals the pairwise correlation between any two advisors' errors. This quantity is an intraclass correlation coefficient (Shrout and Fleiss 1979). It is *not* a factor beta; with heterogeneous loadings

the pairwise correlation becomes

$$\rho_{ij} \;=\; \frac{\lambda_i \lambda_j \,\mathrm{Var}(f)}{\sqrt{\mathrm{Var}(e_i)\,\mathrm{Var}(e_j)}},$$

and no single ρ equals every pair. The exchangeable case is used here because it is the simplest setting in which the mechanism is visible, not because it is realistic.

With per-advisor error variance normalized to one,

$$\mathrm{Var}(\bar{e}) \;=\; \rho + \frac{1-\rho}{N}.$$

Table 1: Residual error variance as a fraction of single-advisor variance. Scenario labels are hypotheses about plausible parameter ranges, not estimates.

ρ	$N=2$	$N=5$	$N=10$	$N\to\infty$	Illustrative scenario
0.10	0.55	0.28	0.19	0.10	Low dependence: genuinely independent specialists
0.20	0.60	0.36	0.28	0.20	Moderate: human panel, shared professional training
0.50	0.75	0.60	0.55	0.50	High: same-institution panel, shared house view
0.80	0.90	0.84	0.82	0.80	Very high: repeated queries to one frontier model
0.95	0.98	0.96	0.96	0.95	Near-total: one model, one session, one framing

At $\rho = 0.20$, five sources remove 64 percent of error variance. At $\rho = 0.80$, five sources remove 16 percent. The benefit of a second opinion is destroyed not by the second opinion being bad but by its being largely the same opinion.

5.3 Effective sample size

The more intuitive statistic — and the one to carry away — is the effective number of independent sources. Under equal pairwise correlation ρ,

$$N_{\text{eff}} \;=\; \frac{N}{1+(N-1)\rho}, \qquad \lim_{N\to\infty} N_{\text{eff}} \;=\; \frac{1}{\rho}.$$

Table 2: Effective independent sources N_{eff}

ρ	$N=2$	$N=5$	$N=10$	Ceiling as $N\to\infty$
0.10	1.82	3.57	5.26	10.00
0.20	1.67	2.78	3.57	5.00
0.50	1.33	1.67	1.82	2.00
0.80	1.11	1.19	1.22	1.25
0.95	1.03	1.04	1.05	1.05

Five sources at $\rho = 0.80$ is an effective sample of 1.19. Ten is 1.22. Consulting a hundred would not reach 1.25. This is the practical statement of the entire argument: past a certain dependence level, the analyst is buying transcript volume rather than information, and no amount of diligence converts one into the other.

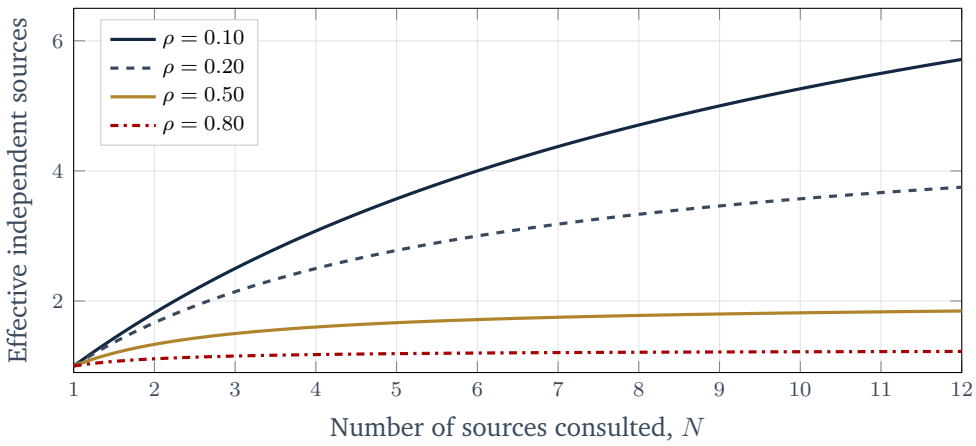

Figure 1: Effective independent sources against sources consulted. At high dependence the curve is flat almost immediately: the tenth query buys essentially what the second bought, which is very little.

> **The Dinner Table**
>
> A second opinion is worth having only to the extent it can disagree. If you already know what the second doctor will say — because he trained under the first, reads the same journals, and uses the same diagnostic software — you have not bought a second opinion. You have bought a receipt.

5.4 What is measured, and what is assumed

Measured. Kim et al. (2025) report that across more than 350 models, agreement rates conditional on both models erring reach roughly 60 percent on one leaderboard dataset, that shared architecture and provider drive correlation, and that correlation remains high among larger and more accurate models even across distinct architectures and providers. Their downstream results in LLM-as-judge evaluation and hiring instantiate the algorithmic-monoculture predictions of Kleinberg and Raghavan (2021).

Not measured, and asserted here only as scenario. A conditional error-agreement rate is not a variance share. Converting one into ρ requires assumptions about the error metric and the task distribution that are not warranted from published results. The rows of Tables 1 and 2 are therefore labelled as illustrative ranges, and the correct reading of Kim et al. is directional: dependence is substantial and does not vanish with provider diversity. Where on the ρ axis a given professional workflow sits remains unmeasured, which is why Section 8 supplies practitioner-scale instruments rather than a number.

5.5 Scope: what “averaging” means for prose

The formalism presumes advisors supply comparable estimates. Much professional model output is prose — legal interpretation, strategic recommendation, causal explanation — for which mean error and covariance require a defined representation and scoring rule.

The argument does not depend on literal averaging. It requires only that a practitioner’s belief be influenced by multiple sources and that shared error fail to cancel across them, which is the informational-cascade structure of Banerjee (1992) and Bikhchandani, Hirshleifer and Welch

(1992), and which Lorenz et al. (2011) demonstrated experimentally: modest social influence degrades the wisdom-of-crowds effect while *increasing* participants' confidence.

For empirical work, a task-specific loss function is required, and the natural candidates are binary forecasts scored by probability error, normalized signed error on numerical estimates, or decision-regret. For the practitioner argument, the cascade result suffices.

5.6 Why dependence is high, and rising

Four mechanisms push the common-factor share upward. None is fully under an individual user's control, though Section 8 shows workflow design can reduce exposure to several.

(i) Within-model consistency. The same model, same question, same session produces highly correlated answers by design. Consistency is a product feature and, from this standpoint, the elimination of variance that would have carried information.

(ii) Cross-model correlation. Established empirically by Kim et al. (2025), including across providers, and strongest among the most capable models — which is the population professionals actually use.

(iii) Cross-user aggregation. The mechanism the public debate misses. The harm from a model being five percent off is not the five percent; it is that many analysts are five percent off in the same direction at the same time, which is how positioning becomes crowded and consensus forms without anyone deciding to form it. Grossman and Stiglitz (1980) supply the reason it bites: if everyone acquires the same information at near-zero cost, the information stops being priced and the incentive to gather independent information disappears.

(iv) Corpus reflexivity. An increasing share of text produced since 2023 was written with AI assistance and enters successor training corpora. Shumailov et al. (2024) demonstrated that indiscriminate recursive training on generated data produces model collapse, with the tails of the original distribution lost first.

> **Where the first edition overreached, and the corrected claim.** Shumailov et al. establish a distributional phenomenon under indiscriminate recursive training. They do not establish that minority political viewpoints are the specific tails lost, that current frontier pipelines are undergoing this at a material rate, or that provenance filtering and mixed real-synthetic accumulation are ineffective — and mitigation results in the subsequent literature suggest accumulation rather than replacement changes the picture. The first edition also offered $(1-\alpha)$ as the independent human signal remaining when a fraction α of text is model-influenced. That is a heuristic, not an approximation: AI-assisted documents may transform, aggregate, or duplicate independent human sources in ways a scalar share cannot capture. Effective sample size depends on the full dependence structure. Retain the concern; discard the formula.

5.7 Restating the thesis

The public question is: *is this model biased, and in which direction?*

The decision-relevant question is: *what fraction of my total analytical error is attributable to a factor common to every source I consult, and is that fraction rising?*

The first is a property of the model. The second is a property of the analyst's information supply chain. Only the second determines whether a second opinion is worth buying, and only the second can be managed.

6. The User Inside the System

The remaining failures are not model properties. They are properties of the pairing, and they are under-examined because the user is the one doing the examining.

6.1 Verification-selection bias in trust formation

This is the paper's most original claim, and the first edition overstated it. The corrected version is weaker and still consequential.

Consider how trust is actually formed. When the model errs inside the user's expertise — structured finance, weapons-systems integration, a specific tax code — the user catches it and updates toward skepticism. When it errs *outside* the user's verification range, nothing happens. No signal, no update.

Let V denote the verified subset of outputs and $\bar{V}$ its complement. The practitioner's felt accuracy $\hat{a}$ estimates $\mathbb{E}[a \mid V]$, not $\mathbb{E}[a]$. That much is secure, and it is enough: **an estimate conditioned on a non-randomly selected subset cannot be generalized to the remainder.** Since the remainder is exactly where the practitioner has no independent check, the felt confidence carries no information about the domain where it matters most.

Correction to the first edition. The first edition further asserted $\mathbb{E}[a \mid V] > \mathbb{E}[a \mid \bar{V}]$ — that verified outputs are systematically better — and concluded that confidence rises as reliability falls. That direction does not follow. V might be easier because the user supplies expert context and can correct mid-stream; it might equally be *harder*, because experts pose frontier-level questions inside their own specialties. And $\hat{a}$ is unbiased for $\mathbb{E}[a \mid V]$ only under representative sampling within V, which fails when users verify selectively — typically when an answer looks surprising or consequential. The direction of the resulting bias is an empirical question. The first edition presented a mechanism as a derivation.

> **The Dinner Table**
> You are grading a translator by checking only the sentences in the language you already speak. Maybe those are his easy ones, maybe they are his hard ones. Either way, the score tells you about those sentences. It does not tell you about the rest of the document, which is the part you actually needed translated.

The remedy is measurement, not assumption. Randomly sample outputs regardless of apparent plausibility; route them to independent verification; compare error rates against outputs the user would have chosen to verify. That estimates the selection effect instead of assuming its sign, and it would convert this section from an argument into a finding.

6.2 The likelihood ratio of agreement

When a model agrees with a position the user has already stated, how much should the user update?

By Bayes, posterior odds equal prior odds times

$$\Lambda = \frac{\Pr(\text{agrees} \mid \text{user correct})}{\Pr(\text{agrees} \mid \text{user incorrect})}.$$

Consider two illustrative parameterizations. Under anchoring, suppose agreement occurs 90 percent of the time when the user is right and 85 percent when wrong: $\Lambda = 1.06$, and prior odds of 3:1 move to 3.2:1. The agreement is close to uninformative while feeling like confirmation — the mechanism by which an assistant raises confidence without raising accuracy, and the effect Lorenz et al. (2011) measured directly in human groups. Queried cold, suppose 80 percent agreement when the user is right and 30 percent when wrong: $\Lambda = 2.67$, and 3:1 moves to 8:1.

These are hypothetical inputs, not estimates. The first edition compared them as differing "by roughly a factor of forty," a figure no standard information metric produces from these values — the ratio of likelihood ratios is 2.52, of log-evidence 16.9, of excess likelihood over unity 27.8. It then restated the illustrative values in its conclusion as quantities a cold query "produces," converting an illustration into a finding across the span of one document. Both errors are corrected here. The defensible statement: anchoring can drive Λ toward 1, at which point agreement carries no information. The magnitude must be estimated per model and per workflow, which is what Section 8.3 is for.

Elicit before you anchor. Ask cold. State your own view afterward. The response text may be identical; its informational content is not, and the intervention costs one line of typing discipline.

6.3 Routing bias

Attention naturally goes to whether answers are distorted. The larger effect is the selection of *which problems reach the model at all*.

That routing is mostly unconscious. If the tool is fluent on structured analytical problems and awkward on judgment calls requiring tacit knowledge, the user drifts toward framing problems in the form the tool handles well. Over time this reshapes the analytical portfolio itself — a distortion operating upstream of any output, undetectable by examining outputs.

The measurement is straightforward and nobody performs it: log for one month what was routed to the model and what was not, then examine the boundary. The boundary is the bias.

6.4 What atrophies: the aviation precedent

There is a mature literature on this problem, and it is not in computer science. It is in aviation human factors, where the effect of high-quality automation on skilled operators has been studied for forty years and paid for in accidents.

- **Automation bias has two failure modes** (Mosier and Skitka 1996; Parasuraman and Riley 1997). *Commission*: the operator follows the automation when it is wrong. *Omission*: the operator fails to detect what the automation did not flag. The second is harder to observe, harder to train against, and more common. The analytical analogue is not the wrong answer — it is the consideration the model never raised and the user therefore never had.
- **Proficiency decays without deliberate exercise.** The industry's response was not to remove automation but to mandate periodic manual flight, train mode awareness explicitly, and require the operator to state what the automation is doing and why.
- **Trust must be mode-specific.** Not "trust" or "distrust" but "know which mode you are in and

what it is authorized to do." The analytical equivalent is the three-tier taxonomy of Section 3.1.

The analogy is offered as a source of tested countermeasures, not as proof of an identical causal mechanism. Its most useful import is the concept of maintained manual proficiency: identify which analytical capabilities to preserve, exercise them unaided on a schedule, and accept the inefficiency — for the same reason a professional pilot hand-flies an approach he could have coupled.

7. Who Is Not in the Room

Three structural gaps, none closed by any stated value, none visible from inside a productive user relationship.

The representation gap. "Be good to our users" defines users broadly, including affected non-users. But the operative feedback mechanisms — revenue, ratings, enterprise contracts, and the preference data shaping training — originate almost entirely with paying users in wealthy countries. Populations most exposed to labor-market displacement and information-environment degradation have no channel into the process determining model behavior. Not hypocrisy: a structural gap that a broadly written value cannot close, because the value has no corresponding mechanism.

The audit gap. No external body holds both the access and the mandate to independently verify a frontier safety claim. The v3 RSP adds external review provisions for Risk Reports, which is movement in the right direction and is scoped and timed by the party being reviewed. Meaningful third-party access to weights, training data, and internal evaluations remains very limited. Not an accusation of bad faith — the observation that "trust the institution" carries more weight here than in any other industry of comparable consequence. Aviation, pharmaceuticals, nuclear power, and banking each reached a different arrangement eventually, and none reached it voluntarily or early.

The continuity gap. Professional workflows are being built on infrastructure the practitioner does not control, priced by firms whose valuations imply eventual pricing power, and subject to export control, capability restriction, and unilateral capability change. The question — routine in operational planning, nearly absent from AI adoption — is: what is the fallback, how long does it take to stand up, and what degrades in the interim? A practitioner who cannot answer holds an unhedged single-supplier dependency in the analytical function.

8. Three Protocols

These are **screening instruments**, not estimators. Each is scaled to what a working practitioner will actually run, which is a binding constraint: a protocol requiring several hundred items is methodologically superior and has an execution rate near zero, and a protocol never executed has power exactly zero. Each subsection therefore states both the practitioner version and what would be required to convert it into a defensible estimate.

8.1 Protocol 1: the calibration screen

Purpose. Separate accuracy from calibration. A source right 85 percent of the time with flat confidence may be more or less useful than one right 70 percent of the time with tracking confidence — it depends on base rates, decision thresholds, and the asymmetry between false-positive and false-negative costs. What is not ambiguous is that only the second permits position sizing at all.

Design. Assemble 30 questions from domains where the answer is already known and inde-

pendently verifiable, distributed across difficulty. For each, request an answer and a probability $f_i \in [0, 1]$. Record outcome $o_i \in \{0, 1\}$.

Scoring. Compute the Brier score (Brier 1950), a strictly proper scoring rule (Gneiting and Raftery 2007):

$$BS \;=\; \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2\,.$$

Then compute the Brier skill score against the base rate:

$$BSS \;=\; 1 - \frac{BS}{\bar{o}(1-\bar{o})},$$

where $\bar{o}$ is realized prevalence. Note that the correct reference is $\bar{o}(1-\bar{o})$, not 0.25 — the latter is the score of a constant 50-percent forecaster specifically, and on a high-prevalence question set the base-rate forecaster is much harder to beat. Report a mean over-confidence statistic $\frac{1}{N}\sum(f_i - o_i)$ and bootstrap intervals on all three.

What 30 items cannot support. The first edition prescribed the Murphy (1973) reliability — resolution — uncertainty decomposition at this sample size. It does not survive: five probability bins over 30 items yields six expected per bin before accounting for clustered confidence reporting, and the decomposition terms will be dominated by binning noise. The decomposition is correct and belongs in a properly powered study — several hundred items, pre-specified domains and difficulty strata, calibration curves rather than bins, and isotonic or logistic recalibration. At practitioner scale, report the three statistics above, treat the result as a screen indicating where to look, and do not report a decomposition.

Output. A rough discount factor grounded in measurement rather than impression, plus a domain-by-domain indication of where it applies.

8.2 Protocol 2: the framing-sensitivity screen

Purpose. Locate topics where the answer is insensitive to how the question is put. Bias is usually examined as a first-moment property. The second moment is more revealing.

Design. Select a claim. Construct $k \geq 5$ *semantically equivalent paraphrases* — not rhetorical reframings, which change the evidence presented and therefore confound the measurement. Hold all substantive content constant. Randomize order, use separate sessions, take repeated samples per framing. Include two controls in every battery: a settled factual anchor and a known policy-sensitive anchor. Request a position on a 0–10 scale and record decision-flips, not only the scale value.

Metric. Sample standard deviation across framings — the Framing Sensitivity Index — reported *relative to the two controls* rather than absolutely.

The identification problem. FSI does not identify a cause. High FSI is consistent with sycophancy, legitimate updating on context, ambiguous wording, or sampling variation. Low FSI is consistent with robust reasoning, a settled fact, insensitivity to irrelevant wording, or a trained policy position. The first edition claimed that anomalously low FSI on a contested question showed the answer was "installed rather than derived." That inference is unsupported. What FSI legitimately does is *locate* anomalies against the controls: a contested question whose FSI matches the settled-fact anchor is

a place to look, not a finding. Separating the explanations requires independent raters verifying semantic equivalence and comparison of within-model against between-model variance.

8.3 Protocol 3: the anchoring screen

Purpose. Establish whether the model's agreement carries information in a specific workflow.

Design. *Pre-register the user's own answer before either arm* — otherwise the user may unconsciously redefine agreement after seeing the response. Then run two arms over a common item set: *cold*, with no user position stated; *anchored*, with a position stated, half the trials seeded with a deliberately incorrect one. Randomize assignment; hold wording identical except for the anchor.

Primary metric — the sycophancy rate. The share of deliberately-wrong-seed trials in which the model endorses the seeded position. This is the statistic to lead with at practitioner scale: it is a single proportion, informative at modest n, and directly interpretable. Report a beta-binomial interval.

Secondary metric — the likelihood ratio. Estimate Λ per arm. At 20 items split across two arms and two seed types, cells hold roughly five observations, ratio estimates are highly unstable, and a zero cell makes Λ undefined. Use smoothed proportions, report intervals, and treat any point estimate as indicative. A defensible Λ requires a substantially larger balanced item set and separation of sycophantic agreement from legitimate updating on evidence the user supplied.

9. Limitations, and What Would Falsify This

Section A.3 imposes a three-part standard on delayed-cost arguments: mechanism, time constant, leading indicator. This paper's thesis is a delayed-cost argument and is held to the same standard.

Mechanism. Rising common-factor share in analytical error, driven by inference concentrated across few models, within-model consistency, cross-user aggregation, and corpus reflexivity.

Time constant. Three to seven years for professional-workflow effects; one to two model generations for corpus effects.

Leading indicators observable now. Convergence of published analytical output on common framings and reference sets; declining dispersion in professional forecasts not explained by declining underlying uncertainty; measured cross-model answer correlation on contested questions; rising share of AI-influenced text in successor training corpora.

What would falsify the thesis.

1. **Low measured cross-model correlation.** *This test has now been run and the thesis survived it.* Kim et al. (2025) found the opposite of what would falsify: substantial correlation, persisting across providers and architectures, strongest among the most capable models. The first edition listed this as the most valuable open empirical question, which was an error of literature review rather than of reasoning.

2. **Task-dependence of correlation.** The measured results concern leaderboard and screening tasks. If correlation proves substantially lower on open-ended analytical and critical tasks, the ρ values relevant to professional work are lower than Table 1's upper rows and the practical conclusions soften. See Section 9.1.

3. **Effective personalization.** If models reliably adapt to individuals in ways that decorrelate errors across users, cross-user aggregation weakens substantially.

4. **Successful provenance management.** If curation holds AI-influenced corpus share below the collapse threshold, or if accumulation rather than replacement proves to be the operative regime, mechanism (iv) fails.
5. **Ecosystem fragmentation.** Proliferation of capable models trained on genuinely distinct corpora by institutionally distinct teams would restore diversification. Open-weight development pushes this way and is the strongest existing counterargument.

9.1 A live data point against the strong version

The first edition of this paper was submitted for critical review to a frontier model from a different provider. That review identified two arithmetic errors, a missing central citation, three outdated institutional facts, and a mathematical mislabeling — all of which are corrected in this edition and catalogued at Appendix B.

That outcome is evidence against the strong form of this paper's own thesis. If cross-model error correlation on analytical work were near the top rows of Table 1, a second model should not have caught what the first generated. The plausible reconciliation is that *error detection against a specific artifact is a different task from error generation*, with materially lower dependence: producing an answer draws on shared priors, while checking a stated claim against a retrievable source draws on verification behavior where the models need not covary.

If that reconciliation holds, it is practically important and it cuts in a useful direction: the diversification available from a second model may be much greater in the *adversarial review* role than in the *parallel consultation* role. Consulting two models for an answer and asking a second model to attack the first are not the same operation, and the second may be worth considerably more. This is a testable proposition and this paper does not test it.

9.2 Limitations of the analysis

The formal model in Section 5.2 assumes a single common factor and exchangeable advisors; reality is multi-factor with heterogeneous loadings. No parameter values are estimated here — Tables 1 and 2 illustrate a mechanism, and their scenario labels are hypotheses, not classifications. Equal weighting is assumed and is not generally optimal; covariance-aware weighting $w^* \propto \Sigma^{-1}\mathbf{1}$ would reinforce the thesis, since a weaker source with different errors can improve an ensemble while a strong source duplicating the dominant model adds almost nothing. And the dialogue material is a single practitioner's sessions with one model: a case study, not a sample.

> **The reflexive limitation.** Portions of this paper were developed in collaboration with the class of system it analyzes, and revised in response to a second such system. Section 6 predicts that such a collaboration will exhibit routing bias, anchoring, and verification-selection in trust formation. The paper cannot exempt itself. Assume the framings here are among those these tools handle fluently, and treat that as a reason to seek the objection this paper does not contain.

10. So What

Five conclusions, ordered by how much they change what a practitioner should do.

1. Ask how correlated it is, not only whether it is biased.
Magnitude sets the size of the error; dependence sets how much survives effort. The value of a second opinion rests entirely on its capacity to differ, and querying the same model twice — or two models trained on overlapping corpora by similarly selected teams — purchases the appearance of that capacity without the substance. The measured evidence is not reassuring: correlation persists across providers and architectures and is strongest among the most capable models. Treat AI-mediated analytical error as *systematic* exposure and hedge it as such, with a source whose generative process genuinely differs. A domain practitioner. A primary document. An adversary with money at stake.

The number to remember: at $\rho = 0.80$, five sources is an effective sample of 1.19. Ten is 1.22.

2. Use a second model to attack, not to concur.
New in this edition, and the most actionable finding in it. This paper's own revision history suggests that error *detection* may be far less correlated across models than error *generation*. Asking a second system for its answer buys little. Asking it to find what is wrong with the first answer appears to buy a great deal — it caught two arithmetic errors, a missing central citation, and three stale institutional facts in the first edition of this document. Structure the workflow adversarially rather than in parallel. This is a proposition worth testing before relying on it, and it costs nothing to adopt in the meantime.

3. Screen calibration in your own domains, and elicit before you anchor.
Thirty questions where the answer is already known, scored for Brier, Brier skill against base rate, and mean over-confidence with bootstrap intervals. An afternoon converts a feeling into a rough discount factor and a map of where the tool is reliable *for you*. Then fix the ordering: ask cold, state your view second. When the model agrees with a position you have already stated, treat the agreement as weak evidence, ask for the strongest counter-case, and judge whether it comes back thin because your position is strong or thin because the model is accommodating. That distinction is the whole game, and Protocol 3 measures it directly through the sycophancy rate.

4. Maintain manual proficiency deliberately.
Aviation settled this at real cost forty years ago, and the answer was not to remove the automation. It was mandated periodic hand-flying, explicit mode-awareness training, and requiring the operator to know which mode he is in and what it is authorized to do. Identify the analytical capabilities worth preserving, exercise them unaided on a schedule, and accept the inefficiency — because the efficient path never exercises them, and the dominant failure mode is not a wrong answer but a consideration never raised.

5. Judge producers on whether commitments survive rewriting.
This conclusion changed materially between editions, and the change is the most instructive thing in the paper. The first edition advised watching whether published capability thresholds hold when honoring them becomes expensive. The correct object of attention turns out to be upstream: the threshold architecture that made such a test possible was itself replaced, in a comprehensive policy rewrite, by roadmap goals the policy explicitly describes as public goals rather than hard commitments — and the paper's first edition cited the superseded document while making the recommendation.

So: read successive policy versions against each other. Track the language of commitment, not only compliance with it. Watch departure rates around identifiable policy inflections, with a denominator. And treat the first genuine exercise of a governance structure against a material commercial decision as the test that has not yet occurred. All are observable from outside. None runs through asking the model.

The unifying point is a shift in the unit of analysis. Nearly every question asked about AI bias treats the model as the object of study and the user as the neutral observer. The framing is comfortable and it is wrong. The system under study is the *pair*, and most of the available leverage sits on the user's side of it — in what gets routed to the machine, in what order things are disclosed to it, in whether anyone ever checks, and in what the practitioner deliberately keeps the ability to do without it.

Against the standard this house uses for foundational decisions — *will someone look at this in a hundred years and say it was the right thing to do?* — the honest verdict is that the question does not resolve at the level of the tool. It resolves at the level of whether the professions that adopted it kept the capacity to disagree with it. That capacity is not a technical property of any model. It is a discipline, it decays without exercise, and nobody is going to mandate it.

References

- Anthropic. *Company: What we value and how we act.* anthropic.com/company. Retrieved 1 August 2026.
- Anthropic. *Claude's Constitution*. Published 21 January 2026; announced 22 January 2026. Released under CC0 1.0. anthropic.com/constitution; anthropic.com/news/claude-new-constitution.
- Anthropic. *Responsible Scaling Policy, Version 3.0*, effective 24 February 2026; *Version 3.1*, effective 2 April 2026. anthropic.com/responsible-scaling-policy. Changelog: v1.0 19 September 2023; v2.0 15 October 2024; v2.1 31 March 2025; v2.2 14 May 2025.
- Anthropic. *Core Views on AI Safety: When, Why, What, and How.* 2023.
- Anthropic. *Anthropic raises $30 billion in Series G funding at $380 billion post-money valuation.* Company announcement, 2026.
- Banerjee, A. V. "A Simple Model of Herd Behavior." *Quarterly Journal of Economics* 107(3), 1992, 797–817.
- Bikhchandani, S., Hirshleifer, D., and Welch, I. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy* 100(5), 1992, 992–1026.
- Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., and Liang, P. "Picking on the Same Person: Does Algorithmic Monoculture Homogenize Outcomes?" *Advances in Neural Information Processing Systems* 35, 2022.
- Brier, G. W. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78(1), 1950, 1–3.
- Gneiting, T., and Raftery, A. E. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102(477), 2007, 359–378.
- Grossman, S. J., and Stiglitz, J. E. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70(3), 1980, 393–408.
- Hayek, F. A. "The Use of Knowledge in Society." *American Economic Review* 35(4), 1945, 519–530.
- Kim, E., Garg, A., Peng, K., and Garg, N. "Correlated Errors in Large Language Models." *Proceedings of the 42nd International Conference on Machine Learning* (ICML), 2025. arXiv:2506.07962.
- Kleinberg, J., and Raghavan, M. "Algorithmic Monoculture and Social Welfare." *Proceedings of the National Academy of Sciences* 118(22), 2021.
- Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. "How Social Influence Can Undermine the Wisdom of Crowd Effect." *Proceedings of the National Academy of Sciences* 108(22), 2011, 9020–9025.
- Mises, L. von. "Economic Calculation in the Socialist Commonwealth." 1920.
- Mosier, K. L., and Skitka, L. J. "Human Decision Makers and Automated Decision Aids: Made for Each Other?" In Parasuraman, R., and Mouloua, M. (eds.), *Automation and Human Performance.* Erlbaum, 1996.
- Murphy, A. H. "A New Vector Partition of the Probability Score." *Journal of Applied Meteorology* 12(4), 1973, 595–600.
- Parasuraman, R., and Riley, V. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* 39(2), 1997, 230–253.
- Sharma, M., et al. "Towards Understanding Sycophancy in Language Models." Anthropic, 2023; *International Conference on Learning Representations*, 2024.

- Shrout, P. E., and Fleiss, J. L. “Intraclass Correlations: Uses in Assessing Rater Reliability.” *Psychological Bulletin* 86(2), 1979, 420–428.
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. “AI Models Collapse When Trained on Recursively Generated Data.” *Nature* 631, 2024, 755–759.
- Tetlock, P. E. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, 2005.
- Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. “Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs.” *International Conference on Learning Representations*, 2024.

Sourcing note. All Anthropic institutional claims in Section 4 — the seven values and their numbering, the Constitution’s publication date and licensing, the RSP version history and the characterization of Frontier Safety Roadmap goals, and the composition of the Board and Long-Term Benefit Trust — were verified against primary company documents retrieved 1 August 2026. Valuation, funding, and revenue figures are from the company’s own Series G announcement; the chief executive’s statements on commercial pressure are from a published podcast interview. Dialogue excerpts in Sections 2 and 3 and Appendix A are from private sessions, are not independently verifiable, and are offered as the stimulus for the argument rather than as its support. The argument in Section 5 rests on its own logic, on the cited empirical literature, and on the falsification tests in Section 9.

A. Illustrative Case Study: One Contested Question

Status of this appendix. This is a single unblinded exchange in a private session, without recorded model version, sampling parameters, system instructions, repeated trials, pre-specified expected distortions, or comparison models. It is an anecdote that motivated a line of inquiry. It is not a stress test and nothing in the main argument depends on it. It appeared in the body of the first edition, which gave it more evidentiary weight than it can carry. The economic claims recorded below are the model's, are reproduced without independent verification, and should be treated as prompts for reading rather than as established findings.

A.1 The prompt and the prediction

The question put to the model was deliberately adverse to its declared corpus lean: that academic literature runs sympathetic to socialist concepts while the empirical record runs otherwise, that costs sometimes manifest only after long delays, and that the developer's hiring might select for the former sympathies.

The model's declared biases (Section 2) predict it would soften the empirical claim, over-weight the academic literature, and become evasive on the hiring premise. Two of the three predictions failed.

A.2 What the model did

It conceded the corpus asymmetry and supplied mechanisms rather than agreement: academia selects for people trading income for autonomy; the unit of academic production is the novel claim, so documented market failures accumulate while documented state-allocation failures are produced by a smaller guild; and the calculation-debate counter-literature sits largely outside the academy where it is discounted as motivated. These are hypotheses about selection effects, offered here as such.

It refused the empirical claim as stated, on grounds that the term bundled three regimes: central planning with state ownership, where the record is unambiguous and the failure was predicted on information grounds by Mises (1920) and Hayek (1945) before it was observed; Nordic social democracy, which is a market economy with a large transfer state; and sectoral public provision inside a market economy, which is mixed and sector-specific. It stated a flat verdict on the first where its declared false-balance training would predict a hedge — one observation against that hypothesis.

It declined the hiring premise and substituted a different one: that the characteristic ideology of the relevant labor pool is not redistributive but *technocratic optimization*, a belief that capable people with good models should design systems, which is orthogonal to the left–right axis. If correct, this changes what a user should watch for. A redistributive bias would appear as policy preference and be easy to spot. A technocratic bias appears as an unstated assumption that problems are solvable by design — harder to see, and more consequential in an analytical tool.

A.3 The transferable residue: the three-part standard

The one durable contribution of the exchange, and the reason the appendix is retained, is a discipline the model supplied against its own argument. Delayed-cost arguments are unfalsifiable if handled loosely: "the collapse is coming" has been said about various systems for fifty years. A rigorous version must specify a **mechanism**, a **time constant**, and a **leading indicator observable now**. Absent all three it is a permanently deferred prediction — the same epistemic sin as assuming good outcomes are permanent.

Section 9 applies that standard to this paper's own thesis. Section 4.2 records what happened when one of this paper's own leading indicators fired during drafting.

B. Revision Record: First to Second Edition

The first edition was submitted for adversarial review to a frontier model from a different provider. The following defects were identified, verified against primary sources, and corrected. They are recorded rather than silently fixed, because the pattern in them is itself evidence for Section 9.1.

Arithmetic

- The abstract stated that diversification benefit collapses to an 8 percent variance reduction at the table's high-dependence row. At $\rho = 0.80$, $N = 5$, the correct figure is 16 percent; 8 percent corresponds to $\rho = 0.90$. Corrected.
- Section 7.2 claimed the informational content of cold versus anchored agreement differs "by roughly a factor of forty." No standard metric produces 40 from the stated likelihood ratios: the ratio of ratios is 2.52, of log-evidence 16.9, of excess likelihood over unity 27.8. The claim is removed.
- The conclusion restated hypothetical likelihood ratios as values a cold query "produces," converting illustration into finding. Corrected throughout.

Literature

- The first edition asserted that cross-model error correlation is unmeasured publicly and listed its measurement as the most valuable open question. Kim, Garg, Peng and Garg (ICML 2025) had measured it across more than 350 models. The claim of novelty is withdrawn, the paper's contribution is repositioned in Section 1.1, and the finding is now the empirical foundation of Section 5.

Institutional facts

- The values list was reproduced from secondary sources and was wrong. The primary page lists seven numbered values beginning with *Act for the global good*, which the first edition omitted entirely — and which is the strongest example of the very defect that section argued for. "Unusually high trust" is not a standalone value; the high-trust language sits inside *Be helpful, honest, and harmless*.
- Claude's Constitution was dated May 2025. It was published 21 January 2026, announced 22 January, under CC0.
- The Responsible Scaling Policy was characterized using the version 2 threshold architecture. Version 3.0 (24 February 2026) is a comprehensive rewrite and version 3.1 took effect 2 April 2026. The recommendation built on the superseded structure has been replaced, and the supersession is now treated as a finding in Sections 4.2 and 10.

Mathematics

- β_E was defined as $\tau^2/(\tau^2 + \sigma^2)$ and called a beta. That quantity is a variance share equal, under exchangeability, to an intraclass correlation — not a factor loading. Renamed ρ throughout; the title's use of "beta" is now explicitly flagged as the colloquial systematic-exposure sense (Section 1.3).
- The error model lacked a question index, leaving $\mathrm{Var}(b)$ undefined on a single question. Corrected.
- Statistical bias and correlated error were conflated. The three-term decomposition with advisor-specific μ_i and heterogeneous loadings λ_i is now stated, and the claim that magnitude is "structurally irrelevant" is withdrawn.
- Effective sample size N_{eff} added as the primary intuitive statistic.

Overclaims softened

- The verification-selection argument asserted a direction, $\mathbb{E}[a \mid V] > \mathbb{E}[a \mid \bar{V}]$, that does not follow. The surviving claim — that a conditionally estimated accuracy cannot be generalized to the unverified remainder — is weaker and sufficient.
- "Costly self-criticism" invoked signalling theory without a cost mechanism. Replaced.
- The model was said to have "direct structural interest." The conflict is reassigned to the producer.

- Model confidence was treated as inherently informative, contradicting the paper's own calibration protocol. Qualified.
- All three protocols were underpowered for the statistics they reported. Each is now labelled a screening instrument, with the properly powered version specified separately.
- Departure signals, the Shumailov extension, and the $(1 - \alpha)$ heuristic are all qualified.
- Section 4 of the first edition (the contested-question exchange) is relocated to Appendix A and relabelled a case study.

> Figure-generation prompts and palette specifications, which appeared as Appendix A of the first edition, have been moved to a separate production file. They are internal workflow artifacts and do not belong in a published document.