The Verification Boundary

Fluency, Expertise, and the Endogenous Construction of Epistemic Reliability

Flyxion
Independent Researcher

2026

Abstract

A synchronic theory of epistemic risk asks whether a user's present knowledge is sufficient to evaluate a claim. A diachronic theory asks whether the knowledge performing that evaluation was itself acquired through processes sufficiently independent of the claims now being evaluated. This essay argues that language models make the second question unavoidable. Language models weaken the coupling between the appearance of epistemic competence and the possession of epistemically warranted belief, not because fluency and truth are independent, but because the coupling between them is insufficiently reliable for a user to treat surface signals as evidence in the way such signals function when produced by a competent, accountable human interlocutor. This produces a verification boundary: a region immediately beyond a user's competence in which an answer remains plausible while independent verification becomes difficult, so that vulnerability to unwarranted acceptance is non-monotonic in knowledge, greatest not at the extremes of ignorance or expertise but somewhere between them. The boundary is not only dangerous but opaque, since a user's estimate of their own competence is produced by the same constraint system whose reliability is in question. Worse, the boundary is not fixed: because a user's future constraint system can itself be constructed through AI-assisted learning, an early plausible error can cease to function as a claim awaiting verification and become part of the machinery by which later claims are verified, a diachronic failure mode in which verification failure compounds rather than remaining a single, correctable mistake. Part II develops a formal sketch of these claims, separating actual from estimated competence, and constraint quantity from constraint fidelity, to show how a user can simultaneously know more, track their domain less faithfully, and grow more confident in their own judgment.

Part I

1 Introduction: The Wrong Problem

The familiar concern that language models hallucinate is not wrong so much as it is too narrow. Factual fabrication is the most visible symptom of a deeper problem, not the problem itself. The signals by which people ordinarily estimate another speaker's reliability, including fluency, specificity, confidence, appropriate qualification, technical vocabulary, explanatory depth, and acknowledgement of alternatives, can all be generated with a degree of independence from whether the underlying proposition is true, a point related to but distinct from the observation that large language models can produce fluent text without the understanding that fluency ordinarily signals [2]. This is the essay's starting thesis: language models weaken the coupling between the appearance of epistemic competence and the possession of epistemically warranted belief. The resulting problem is not simply that models can be wrong, since humans are wrong constantly and this alone would not be remarkable. It is that many of the ordinary cues by which listeners recognize wrongness stop functioning reliably once the cost of producing those cues drops to nearly zero.

This starting thesis, however, is not where the essay ends up, and it is worth saying so at the outset. Decoupled epistemic signals are old news in a weaker sense: human bluffing, credentialism, and institutional authority have always allowed the appearance of competence without its reality, so what changes with language models is in the first instance a matter of scale, speed, and the absence of any accountable person behind the utterance, rather than a wholly new phenomenon. The more distinctive argument, developed across the sections that follow, is that the boundary separating what a user can and cannot verify is not a fixed line whose location is simply unknown to the user, but a boundary that moves, that is opaque from the inside, and that can be actively reshaped, for better or worse, by the very system whose outputs it is meant to guard against. Decoupling explains why a verification boundary exists at all; what happens to that boundary over time, once it is understood to be neither static nor self-announcing, is the essay's real subject.

2 Why Traditional Reliability Proxies Break Down

Ordinary social epistemology rarely requires independently verifying every statement another person makes. Instead, listeners rely on compressed indicators of competence: whether a speaker can answer detailed follow-up questions, distinguish neighboring concepts, hedge appropriately under uncertainty, produce a coherent causal explanation, respond to objections without contradiction, and maintain consistency across an extended exchange. These indicators are imperfect on their own terms. Experts sometimes bluff,

novices sometimes sound confident, and sophisticated arguments are sometimes wrong despite their sophistication. What kept the system functional regardless of its imperfections was that producing the appearance of deep competence traditionally carried a real cost: sustaining a detailed and internally coherent explanation across pressure generally required possessing at least some of the underlying knowledge, because bluffing degrades quickly under specific questioning. Language models change this by radically reducing that cost, and the claim required to establish the point is a modest one. It is not that fluency and truth are statistically independent, since better-trained models are frequently both more fluent and more accurate, and training can deliberately couple expressions of uncertainty to estimated correctness. It is that the coupling between surface epistemic signal and underlying correctness is insufficiently reliable for a user to treat those signals as evidence of truth in the way they treat them when produced by a competent, accountable human interlocutor. A model that sounds more careful may simply have become better at generating the linguistic form of carefulness, without a correspondingly legible increase in the reliability of what it asserts, so confidence calibration cannot be inferred straightforwardly from tone, however natural that inference has always been in ordinary human conversation. This is not a claim that calibration research is worthless; models can be trained so that expressed uncertainty correlates with accuracy under evaluation [4], and such calibration can genuinely help. The point is narrower: tone alone, absent independent evaluation of that correlation, is not evidence a user can rely on.

The remedies people reach for instinctively do not repair this. Blatant fabrication is comparatively tractable, since a nonexistent book, an impossible date, or a fabricated URL can usually be checked directly against an external record; hallucination in this narrow sense is the easy case. The more dangerous failures are subtler: a real source subtly misrepresented, a valid theorem applied outside the conditions under which it holds, a causal explanation that omits one decisive variable, a correct calculation that answers the wrong mathematical question, a critique that correctly identifies several genuine weaknesses before confidently asserting a central objection that is false. These failures survive ordinary plausibility checks precisely because most of the surrounding structure is correct, and a single false step embedded in an otherwise sound argument is far harder to catch than a claim that is wrong throughout. Asking a model to explain itself does not solve this, because the property that makes explanation informative in human exchange, that a person is forced to expose more of their actual understanding under the demand to justify a claim, does not automatically transfer to a system whose conclusion and explanation originate from the same generative mechanism; a false proposition can acquire an internally coherent false explanation as readily as a true one acquires a true explanation, producing a structure that is coherent throughout without being externally validated anywhere. Consulting multiple models does not solve it either, since agreement among them does not automatically constitute independent confirmation: models can share training data, conceptual assumptions, and characteristic blind spots, so several outputs may represent repeated sampling from related epistemic machinery rather than genuinely independent encounters with the world. Five correlated guesses are not five pieces of evidence, no matter how differently they are phrased, and this connects to the broader distinction between recording, admitting, and committing a claim: multiplicity of judgment is not equivalent

to multiplicity of evidence.

3 Expertise as a Constraint System

Expertise is best understood not as the possession of more facts but as a dense network of constraints against which new claims are automatically tested. A programmer recognizes that a claimed API cannot return the object being described. A physicist notices incompatible units before finishing the sentence. A historian recognizes an impossible chronology on sight. A mathematician sees that a quantifier has silently changed scope between two steps. A biologist recognizes that a proposed mechanism is operating at the wrong biological scale to produce the claimed effect. Much of this verification happens without explicit rederivation, because existing knowledge constrains the space of admissible answers before conscious checking even begins.

It is important to state at the outset, rather than concede only in passing later, that expertise in this sense is not synonymous with truth. What a dense constraint system supplies is density, not automatically fidelity: an expert possesses a great many constraints against which to test new claims, but nothing about possessing many constraints guarantees that all of them correctly track the domain they concern. A physicist trained from a flawed derivation, a historian working from a since-revised chronology, a programmer who has internalized a subtly wrong mental model of a system's concurrency behavior, all have dense constraint networks that are nonetheless locally corrupted, and their expertise can make them confidently and systematically wrong for exactly the structural reason developed later in this essay, not despite their density of constraint but partly because of it. Expertise improves verification only insofar as density is accompanied by sufficient fidelity, and the two should not be assumed to travel together.

With that qualification in place, the ordinary case for AI assistance among experts still holds: experts can benefit enormously from such assistance even while the underlying model remains generally unreliable, because the model generates candidate possibilities cheaply and the expert supplies a constraint system the model's own prose cannot certify on its own. Modern systems do possess substantial learned constraints and can additionally draw on retrieval, execution, external verification tools, and formal systems; the relevant asymmetry is not that the model lacks constraints altogether, but that a user cannot infer which constraints actually operated merely from the quality of the resulting prose. That formulation is the stronger one, because it applies without qualification even to future, more capable systems.

4 The Verification Boundary

This suggests a spatial way of thinking about epistemic risk, organized into three regions. Inside the boundary of one's own competence, model errors are comparatively detectable, because verification is inexpensive when the user already possesses many of the relevant constraints. Far outside that boundary, the user's own ignorance tends to become visible to them, and they generally know that specialist verification is required before acting on the answer. The most dangerous region lies immediately beyond the frontier of competence, where the user knows enough for an answer to appear intelligible and consistent with what they already understand, but not enough to detect the one decisive error embedded within it. Most of the surrounding propositions fit existing knowledge, the reasoning resembles legitimate reasoning in every respect the user can check, and only the crucial inferential step lies outside their ability to verify. This region can be called the verification boundary, and it implies that epistemic vulnerability is non-monotonic with respect to knowledge: knowing somewhat more about a subject can make AI assistance safer, but knowing just enough to find an answer plausible without being able to independently verify it creates a particularly dangerous trust condition, more dangerous in some respects than knowing almost nothing at all.

5 Boundary Opacity

The picture so far treats the verification boundary as something a user could in principle locate, if only they had accurate insight into where their own competence ends. This assumption deserves scrutiny rather than being taken for granted. Accurate knowledge of one's own competence is itself an epistemic achievement, and there is no reason built into the preceding argument to expect a user to possess an independent meter telling them where their competence actually ends. The estimate a user forms of their own competence is produced by the same constraint system whose limits are in question, which means that estimate is subject to exactly the same risk of corruption as any other judgment that system produces.

The verification boundary is not a line a user knowingly approaches. Its location is itself something the user must estimate using the same imperfect constraint system whose limits the boundary describes. The dangerous condition is therefore not simply intermediate competence; it can involve a divergence between actual and estimated competence, particularly acute around the specific claim or subdomain in question, and this divergence can run in either direction. A user who underestimates their competence forfeits assistance they could have safely used. A user who overestimates it treats an output as within their competence to evaluate when it is not, which is the more dangerous error, since it removes the caution that would otherwise have prompted independent verification.

Because expertise is domain-local rather than global, this opacity is best understood not as

a single number a person either has or lacks, but as something that varies claim by claim and domain by domain. A molecular biologist's constraint density in molecular biology does not transfer to statistical methodology, historical inference, software security, or the internal epistemology of the AI system they happen to be using, and their confidence in evaluating claims in those adjacent domains can easily exceed what their actual constraint density in those domains would justify. Expertise, on this account, does not place a person globally inside or outside the boundary. It changes the boundary's shape, and it does nothing on its own to guarantee that a person's estimate of that shape is accurate outside the specific region where the expertise was actually earned.

6 The Diachronic Verification Boundary

The verification boundary, as described so far, is treated as a feature of a given user at a given moment, fixed by whatever constraint density and estimated competence that user currently possesses. But constraint density is not fixed, and one of the primary ways it changes is by using the same generative system the boundary is meant to guard against. A user who asks a model to explain a subject they do not yet understand is not merely receiving an answer; they are acquiring the constraint machinery they will later use to evaluate that system's future answers in the same domain. This creates a channel absent from the earlier picture, in which the source of the propositions and the source of the standard for judging propositions become the same system.

Where this channel is benign, it simply pushes the boundary outward: a user who has genuinely learned a field, even partly through AI-assisted study, now possesses more of the constraint density that made expert use safe in the first place, and their vulnerable zone shrinks accordingly. But the channel is not guaranteed to be benign, because nothing about it filters for correctness. If an early plausible error is accepted and incorporated as background knowledge, it stops functioning as an answer that might later be checked and starts functioning as part of the apparatus that performs the checking. Later, correct answers can then be misjudged as errors, and subtly wrong answers consistent with the earlier error can pass entirely unchecked, not because the user's judgment has degraded in some general sense, but because the judgment now includes the contamination.

This produces a recursive structure the earlier, static picture cannot capture: generation, acceptance, incorporation into constraint, and future verification performed by that now partially corrupted constraint system. It implies that verification failure is not merely a per-instance risk to be weighed each time a claim is checked, but a potentially compounding one, in which an early failure to catch an error at the boundary degrades the very instrument that later checking depends on. The most dangerous version of the verification boundary, on this account, is not a single crossing but a standing feedback loop: it is not enough to ask whether a given answer falls inside or outside current competence, because current competence is itself partly a product of past answers that were never independently checked against anything but their own plausibility. An accepted error, put

most directly, can cease to be an object of verification and become part of the machinery by which verification is subsequently performed.

A synchronic theory asks whether the user's present knowledge is sufficient to evaluate a claim. A diachronic theory asks whether the knowledge performing that evaluation was itself acquired through processes sufficiently independent of the claims now being evaluated. The second question is considerably harder to answer than the first, because introspection alone cannot resolve it. A false constraint that has been successfully integrated into a user's model will ordinarily feel like an unremarkable part of what that user simply knows, precisely because its function is to participate silently in judgments about what follows from what, including judgments about itself. The deeper danger, on this reading, is not only that error lies just beyond the frontier of what a user knows. It is that the frontier separating what is known from what is not may itself have been constructed partly from propositions the user never independently verified.

7 The Verification-Cost Paradox

This picture has an economic consequence worth developing on its own terms. AI provides its greatest apparent labor savings precisely in the cases where independent verification is most expensive. If a user already knows the answer to a question, checking the model's response is easy, but the informational value the model adds is correspondingly small. If the user does not know the answer, the apparent informational value is large, but verifying that value may require acquiring much of the expertise whose absence motivated using the model in the first place. This produces a genuine paradox: the value of delegated cognition increases precisely as the ability to verify that delegation decreases, so the situations in which the tool looks most useful are structurally the same situations in which trusting it is least safe. This places a hard ceiling on naive accounts of AI-driven productivity, because generation costs can approach zero while verification costs remain substantial and, in the cases that matter most, largely unchanged from what they were before the tool existed.

The economic claim and the allocation claim that follows it are worth separating rather than treated as restatements of one another. The present section makes the cost claim: generation costs collapse considerably faster than verification costs. The following sections make a distinct allocation claim: because verification capacity is unevenly distributed, this divergence disproportionately increases the productive leverage of those who already possess it. The two claims are related but not identical, and the second does not follow from the first without the additional premise that the capacity to verify cheaply generated claims is not itself evenly distributed across users to begin with.

8 Evidence Exists Before It Becomes Legible

This connects to a broader distinction between the existence of evidence and its legibility to a given observer. Contradictory information can already be present in a model's output without functioning as evidence for a particular user, because the distinctions required to recognize its significance are simply absent from that user's knowledge. A novice and an expert can inspect the exact same answer and effectively receive different information from it: the expert sees the answer together with a large collection of constraints it satisfies or violates, while the novice sees primarily the answer itself, stripped of the context that would reveal its weaknesses. AI can therefore accelerate the production of information without proportionately accelerating the human capacity to determine what that information actually warrants, which means output volume and epistemic gain can diverge sharply depending entirely on who is reading the output.

9 Generated Plausibility and Externally Constrained Belief

Rather than treating this as grounds for simple distrust, it is more useful to arrange epistemic checks along a rough hierarchy running from fluency through coherence, consistency, independent corroboration, and finally direct verification, where each level provides something the previous one cannot. Fluency establishes only readability. Coherence establishes relationships among the claims within a single response. Consistency establishes compatibility across multiple claims or observations. Independent corroboration connects a claim to evidence produced by some genuinely separate process. Direct verification constrains a claim through observation, execution, measurement, formal proof, or whatever domain-appropriate test applies.

A distinction worth making explicit at this point separates a bare generative model from an epistemic system built around one. A bare generative model exhibits the decoupling problem in its cleanest form, since nothing in its output is directly constrained by anything outside its own training and sampling process. A system coupled to search, code execution, formal verification, sensors, or provenance-tracking databases can partially restore the missing connection to external constraint, and in doing so can genuinely lower verification costs for particular classes of claim. But its reliability in that case derives partly from those external systems rather than from the persuasiveness of its generated explanation, and a user who cannot distinguish which parts of an answer were externally checked from which parts were merely generated retains the original problem in a disguised form. The purpose of the hierarchy above is not to eliminate model-generated reasoning from use, which would be neither possible nor desirable, but to make clear at which point additional epistemic machinery, human or otherwise, becomes necessary before a given claim can be acted on with confidence.

10 The Paradox of Safe Use

This produces an apparently paradoxical conclusion: the less one needs a language model to know something on one's behalf, the more safely one can exploit what it knows. Experts can use models aggressively and productively precisely because they can reject malformed or subtly wrong outputs cheaply, almost as a byproduct of reading them. Novices, by contrast, must treat apparently authoritative answers with the most caution exactly where those answers would be most valuable to them, since the value of an answer to a novice scales with their ignorance while their capacity to check it scales with the opposite. This explains why expert productivity gains and general epistemic unreliability are not contradictory findings, despite often being reported as though they were in tension. Expertise supplies the verification infrastructure that the generative system itself does not possess, and productivity gains appear only where that infrastructure is already in place in the human using the tool, and where that infrastructure is itself reasonably faithful to the domain it concerns.

11 Beyond "Trust but Verify"

The common advice to trust but verify understates the difficulty of the situation, because verification itself has costs and requires competence that is unevenly distributed and cannot simply be willed into existence at the point of use. A more adequate approach distinguishes between different classes of output according to what kind of check they actually admit. Some outputs can be tested computationally. Some can be checked directly against primary sources. Some permit formal proof. Some permit experimental replication. Some depend on genuinely independent expert judgment that cannot be shortcut. Others remain intrinsically interpretive and therefore cannot be cheaply converted into externally constrained knowledge at all, no matter how much verification effort is applied. The appropriate verification mechanism should follow the structure of the claim being made, rather than a generic instruction to the user to be more careful, which offers no guidance on what carefulness would actually consist of in a given case.

12 Conclusion to Part I: The New Scarcity

Language models make plausible explanation extraordinarily cheap. They make candidate hypotheses cheap. They make argument, criticism, counterargument, synthesis, and the surface appearance of expertise all cheap in a way nothing before them was. What they do not do, and what improving generative fluency alone cannot do, is make reality correspondingly cheap to consult. The more defensible and more interesting claim is not that nothing can lower verification costs, since tool use manifestly can, but that im-

provements in generation alone cannot eliminate a verification problem whose solution requires information not contained in generation.

The resulting scarcity is therefore no longer primarily the production of plausible propositions, since plausible propositions are now abundant and nearly free. It is the capacity to discriminate among plausible propositions under some form of external constraint. Once the diachronic argument is taken into account, this scarcity acquires a further layer: the constraints used to perform that discrimination are themselves increasingly produced by the same economy that produces the propositions being discriminated among, which means the scarce resource is not simply discrimination but discrimination whose own foundations were independently earned, using an estimate of one's own competence that is no more guaranteed to be accurate than any other judgment the same system produces. In an environment saturated with generated plausibility, epistemic competence increasingly consists not in producing answers, which any sufficiently fluent system can now do on demand, but in possessing constraints, externally constrained, sufficient to reject the wrong ones, and in possessing some independent basis for knowing where that capacity actually ends. The central problem of language-model epistemology is therefore not hallucination taken in isolation. It is the widening gap between the cost of generating something that looks like knowledge and the cost of establishing that it actually is knowledge, compounded by the fact that the instrument used to establish it can itself be shaped, unnoticed, by what it is trying to check.

Part II: A Formal Sketch of the Verification Boundary

The quantities introduced below are latent structural variables rather than presently calibrated measurements. The formalism specifies relationships that an operational theory would need to preserve without claiming that constraint density, fidelity, or self-estimated competence currently admits a unique empirical metric. Its purpose is therefore to distinguish hypotheses and expose possible trajectories, not to assign numerical values to cognitive states. This is not a concession that the formalism is decorative. Separating constraint density from constraint fidelity prohibits treating knowledge accumulation as automatically truth-preserving. Separating detection from unwarranted acceptance prohibits counting epistemic restraint as successful error detection. Introducing an estimated competence distinct from actual competence prohibits assuming users know their own position relative to the boundary. Each separation eliminates a conceptual mistake that remained available in the purely prose version of the argument, which is the sense in which the notation does real work rather than merely restating Part I in symbols.

12.1 Detection versus Unwarranted Acceptance

The verification-boundary hypothesis developed in Part I requires separating two quantities that are easy to conflate in prose. Let k denote a user's constraint density in a given domain: informally, how much of the admissible and inadmissible structure of that domain the user has internalized. Let $d(k)$ denote the probability that a user who is shown an erroneous output correctly identifies it as erroneous. Let $u(k)$ denote the probability that an erroneous output is accepted, believed, or acted upon without adequate external verification. These are not complementary quantities: a user can fail to detect an error yet still decline to act on it out of general caution, and a user can detect a weaker version of an error while still acting on a subtler one.

Novice epistemic restraint should not be modeled as detection. A novice who declines to trust an unfamiliar claim has not identified anything false about it; they have recognized their own inability to warrant it, which is a different achievement from identifying the specific decisive error. The verification-boundary hypothesis is therefore stated principally in terms of $u(k)$, not $d(k)$:

$$u(k_{\text{novice}}) < u(k_{\text{intermediate}}) > u(k_{\text{expert}}).$$

This does not require that novices outperform intermediates at detection. The hypothesis is compatible with, and is strengthened by, simultaneously observing

$$d(k_{\text{novice}}) < d(k_{\text{intermediate}}) < d(k_{\text{expert}}),$$

that is, with detection rising monotonically with competence while unwarranted acceptance follows an inverted U. Such a combination would indicate that intermediates genuinely know more than novices and correctly identify more errors than novices do, yet remain more vulnerable to the specific errors they fail to detect, because their partial competence renders the remaining, undetected outputs credible enough to act upon. This is a considerably stronger and more surprising claim than the trivial one that intermediates detect fewer errors than experts. One design tests this directly by matching error difficulty to each population's own frontier of competence rather than presenting a single fixed stimulus to all three groups, since a subtlety calibrated to an intermediate user's blind spot is not the same error, functionally, for a novice or an expert. A fixed battery deliberately spanning difficulties is an equally viable alternative, and arguably a more informative one, since it allows vulnerability to be estimated as a function of the relationship between measured competence and independently measured item difficulty rather than requiring categorical populations at all; a natural variable for future operationalization is a distance $\delta_{ij} = k_i - h_j$ between a user's competence k_i and an item's constraint demand h_j, with the hypothesis predicting heightened unwarranted acceptance in some region of slightly negative δ, immediately beyond rather than arbitrarily far beyond a given user's frontier. That reformulation is not pursued further here, but it matters conceptually: it treats the boundary as relational rather than as a property of fixed populations, so that an expert in one subdomain can sit at negative δ minutes later when evaluating a neighboring specialty.

12.2 Boundary Opacity

Until this point k has denoted both a user's actual constraint density and, implicitly, something available to the user in forming judgments about their own reliability. These should be separated explicitly. Let

$$k_t(x) = \text{actual constraint density with respect to domain or claim } x \text{ at time } t,$$

$$\hat{k}_t(x) = \text{the user's estimate of that same quantity.}$$

Indexing both by x rather than treating k as a single global scalar absorbs the observation that expertise is domain-local: a user can possess high $k_t(x)$ in one domain and substantially lower $k_t(x')$ in an adjacent one, without any change to how confidently they hold $\hat{k}_t(x')$. Nothing in the formalism up to this point ever assumed a single global competence per user, so no separate patch is required to accommodate an expert's blind spots outside their own subdomain; the domain-indexing already entails it.

The dangerous condition is not simply intermediate $k_t(x)$. It can involve a divergence

$$\hat{k}_t(x) > k_t(x),$$

and the refined prediction is that unwarranted acceptance should increase with this gap, holding actual competence fixed: for two users i, j with $k_i(x) = k_j(x)$ but $\hat{k}_i(x) > \hat{k}_j(x)$, the hypothesis predicts $u_i(x) > u_j(x)$, with u increasing monotonically in the miscalibration gap $\hat{k}_t(x) - k_t(x)$ under otherwise comparable conditions. This is empirically distinguishable from the original, undifferentiated hypothesis: it requires measuring actual domain competence independently of self-assessed competence, exposing participants to carefully constructed errors, and testing whether positive miscalibration predicts unwarranted acceptance after controlling for actual competence. Vulnerability, on this refinement, depends on at least three quantities that can move independently of one another: actual constraint density, constraint fidelity, and a user's knowledge of their own constraint density. This is what it means to say the verification boundary is not a line a user knowingly approaches. Its location is itself something the user must estimate using the same imperfect constraint system whose limits the boundary describes.

12.3 Constraint Quantity and Constraint Fidelity

The recursive concern developed in Part I requires separating how much constraint structure a user has accumulated from how faithfully that structure tracks its domain. Let k_t denote constraint density at time t, evolving as

$$k_{t+1} = k_t + \Delta k_t,$$

where Δk_t increases with learning regardless of whether what is learned is correct. Let $q_t \in [0, 1]$ denote constraint fidelity, understood as a property of the constraint system as

a whole, $q_t = Q(K_t)$, rather than as an accumulating score. No recurrence is specified for q_t, deliberately: fidelity should depend on the relative weight of correct and incorrect constraints already present, not simply increase with each true proposition learned and decrease with each false one, and specifying such a recurrence would risk an unbounded quantity standing in for what is meant to be a bounded property of a system. What the argument requires is only that incorporating a new proposition changes fidelity according to some unspecified update

$$q_{t+1} = Q(q_t, T_t, w_t),$$

where $T_t \in \{0, 1\}$ indicates whether the incorporated proposition is true and w_t denotes its structural influence, that is, how much of the existing admissibility function it is positioned to alter.

Structural influence and detectability are best treated as two independent dimensions rather than folded into a single notion of error severity. Let detectability denote how readily a given error is caught by ordinary plausibility checks, and let downstream consequence denote how much of the admissibility function applied to later claims is altered if the error is incorporated unchecked. These are not positively ordered: a spectacular hallucination can have extremely high detectability and negligible downstream consequence, while mistaking a foundational rule of inference can have extremely low detectability and enormous downstream consequence once incorporated into K_t. This gives the earlier claim that hallucination is the easy case a firmer basis than conspicuousness alone; what makes an error dangerous is not primarily how visible it is but where it sits on this second, largely independent axis.

Separating k_t from q_t makes visible a trajectory the earlier, undifferentiated picture could not represent:

$$\Delta k_t > 0, \qquad \Delta q_t < 0.$$

A user can be learning, in the ordinary sense of accumulating terminology, distinctions, causal stories, and relationships among concepts, while becoming epistemically worse, in the sense that an increasing share of that accumulated structure is false. Sufficiently developed k does not merely coexist with corrupted q; it can actively stabilize the corruption, since a more elaborated constraint system offers more inferential pathways through which an anomalous observation can be reconciled with an existing, mistaken commitment rather than forcing revision of that commitment. This has a structural cousin in the Quine–Duhem observation that a sufficiently interconnected belief system can typically absorb an anomaly by revising some peripheral element rather than its core [6, 3], with the number of available peripheral elements increasing precisely as the system grows more elaborate; it also has a family resemblance to confirmation bias and belief perseverance in the psychological literature [5, 1], though the mechanism developed here is offered as a structural property of the constraint network itself rather than as a claim about motivated attention.

The cost of an incorporated error, on this account, is not adequately represented as a single false proposition p stored among many true ones. Let K_t denote the user's accumulated

constraint system at time t, and let $A_t(x) = A(x \mid K_t)$ denote the admissibility function it applies to a newly encountered proposition x. If an erroneous proposition e enters K_t, its cost is not confined to the falsity of e itself:

$$e \in K_t \quad \Rightarrow \quad A_{t+1}(x_1), A_{t+1}(x_2), \ldots, A_{t+1}(x_n) \text{ may change.}$$

An accepted error can therefore cease to be an object of verification and become part of the machinery by which verification is subsequently performed, altering the classification of many later, otherwise unrelated propositions rather than remaining a single isolated mistake.

12.4 Synchronic and Diachronic Boundaries

This licenses a formal distinction between two boundaries rather than one. Let

$$B_s(t) = B\big(k_t, q_t, \hat{k}_t\big)$$

denote the synchronic verification boundary: a user's vulnerability given their present constraint density, fidelity, and self-estimated competence, independent of how that configuration was formed. Let

$$B_d(t) = \big(B_s(0), B_s(1), \ldots, B_s(t)\big)$$

denote the diachronic verification trajectory: the ordered sequence, not a set, describing how earlier admissions have shaped the boundary presently in force, since the object of interest is precisely the history and not merely the collection of values it has taken. The synchronic question, given what I currently know, asks which claims I am poorly equipped to evaluate. The diachronic question asks to what extent the knowledge with which I now evaluate a claim was itself produced through earlier, unverified claims. The second question resists introspective resolution in a way the first does not: a false constraint successfully integrated into K_t ordinarily feels like an unremarkable part of what one simply knows, since its function is precisely to participate silently in judgments about what follows from what, including judgments about itself, and including the user's own estimate $\hat{k}_t$ of their competence to make that judgment in the first place.

Once q and $\hat{k}$ are separated from k, the implicit assumption underlying the earlier, purely synchronic picture, that increasing competence eventually guarantees increasing safety, no longer holds unconditionally. The undifferentiated picture assumes something like $k \uparrow \Rightarrow u(k) \downarrow$ once the intermediate danger zone has been crossed. Once fidelity and self-estimation are treated as independent quantities, a considerably more pathological trajectory becomes representable:

$$k \uparrow, \qquad q \downarrow, \qquad \hat{k} \uparrow .$$

A user can gain genuine knowledge, acquire a foundational misconception, and grow more confident in their ability to evaluate the domain, simultaneously and as a single

coherent trajectory rather than as three unrelated events. A plausible empirical signature of this trajectory is that AI-assisted learners might increasingly outperform novices at catching ordinary, unrelated errors while simultaneously growing more resistant to correcting specific foundational misconceptions introduced early in their learning, and more confident than their actual competence in the affected subdomain would warrant, precisely because their growing competence has generated more machinery for reconciling anomalies with those misconceptions, and more material from which to construct an inflated estimate of their own reliability, rather than for exposing either problem.

12.5 Verification Capital

The productivity argument developed in Part I can now be restated more precisely than as a claim about expertise alone. Let verification capital be a structural quantity

$$V = f(k, q)$$

left deliberately unspecified as to functional form; no multiplicative or additive structure is claimed, since none is required by the argument. What is required is only

$$\frac{\partial V}{\partial q} > 0$$

under ordinary conditions, together with the weaker and conditional claim

$$\frac{\partial V}{\partial k} > 0 \quad \text{only under suitable conditions on } q,$$

since it is precisely the possibility that increasing constraint density can degrade verification capacity at sufficiently low fidelity that the contamination argument above establishes. On this formulation, AI leverage does not depend on expertise understood simply as accumulated knowledge; it depends on verification capital, which consists not merely in how many constraints a user possesses but in how reliably those constraints track the domain being evaluated. This also forecloses a romantic reading of human expertise as a stable baseline against which AI-induced contamination is to be measured. Human constraint systems can themselves be incomplete, locally mistaken, obsolete, or otherwise poorly calibrated prior to any interaction with a generative system; AI introduces a new and considerably faster mechanism for altering constraint systems, but it does not introduce epistemic contamination as a category, since human constraint systems were never immune to it.

12.6 Conclusion to Part II

Two further implications follow from the notation once it is in place. The first concerns the object of study: the phenomenon described throughout is not confined to machine-generated text. AI-assisted learning can also make the surface signals of human expertise cheaper to produce, since a person can acquire vocabulary, canonical distinctions,

explanatory fluency, and a densely interconnected conceptual picture considerably faster than the historical economics of learning permitted, without q necessarily scaling to match k, and potentially with $\hat{k}$ scaling faster than either. Where such fluency has genuinely been earned, however imperfectly, this is the same k-outpacing-q trajectory formalized above, and the old heuristic that sustained fluency reliably signals expertise weakens for structurally the same reason it weakened for models, rather than for the separate and considerably older reason that a person might simply borrow talking points without acquiring any constraint structure at all. The second implication concerns the essay's framing as a whole. What began as a question of whether language-model outputs should be trusted becomes, once the boundary is understood to be diachronic and opaque rather than merely synchronic and knowable, a question about what happens to an epistemic culture in which plausibility becomes abundant, self-assessment becomes unreliable in the same measure as the judgments it assesses, and verification, along with the externally constrained constraint systems verification depends on, remains scarce.

References

[1] Anderson, C. A., Lepper, M. R., and Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, 39(6), 1037–1049.

[2] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623.

[3] Duhem, P. (1906). *La théorie physique: son objet et sa structure*. Paris: Chevalier et Rivière.

[4] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1321–1330.

[5] Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.

[6] Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, 60(1), 20–43.