

Adversarial Epistemic Closure

Independent Criticism, Correlated Error, and the Conditions of Theoretical Revision

Flyxion

Independent Researcher

October 2026

Abstract

The availability of criticism does not guarantee the possibility of correction. A research system can generate large quantities of feedback while remaining closed to evidence that would require its foundational assumptions to change, and a comparatively isolated researcher can sustain a productive correction process if the surrounding system supplies independent tests, explicit failure conditions, and a way of absorbing negative results. This paper develops a formal account of the difference between criticism as an available activity and correction as an institutional capability. It separates feedback from information relative to a stated hypothesis space, shows that the number of reviews does not fix their incremental evidentiary contribution and that, when reviews are mediated by a shared source, review volume cannot lower the error probability below a floor set by that source, and decomposes correctability into three properties: adversarial accessibility, adversarial discriminability, and adversarial responsiveness. Responsiveness is defined as a graded, stratified quantity rather than a universal condition. The paper distinguishes revision reachability from revision realization, separates legitimate dependence on a shared background framework from vicious justificatory circularity in the adjudication of objections, proves that circular adjudication makes revision unreachable for the challenges it targets, and shows that no self-governed standard can guarantee correction across a sufficiently broad class of errors. It then states how the account differs from belief-revision theory, Bayesian updating, and social epistemology, presents a pair of processes that coincide on conventional measures of criticism and differ on the proposed ones together with a preregistrable retrospective protocol for testing whether the difference matters, and closes by stating what observations would count against its own claims.

1 Introduction

Research communities describe themselves as self-correcting, and the description is usually offered as an explanation of why error does not persist. The explanation presupposes that criticism, once available, converts into correction. That presupposition is rarely examined, and it is false in general [10]. Criticism is an activity. It can be performed in abundance, received with courtesy, answered at length, and archived in public, while the claims under criticism

remain exactly as they were. Correction is a different kind of thing. It is a change in the epistemic status of a claim, in its scope, or in its formulation, brought about by evidence that discriminates between the claim and its rivals. A system may have a great deal of the first and very little of the second.

The practical significance of the distinction has grown. Journals, preprint servers, open repositories, comment threads, and automated reviewers have lowered the cost of producing objections to nearly nothing. If the correctability of a research process were measured by the volume of feedback it receives, correctability would now be at an all-time high. Few people believe this, and the reasons for disbelief are worth stating exactly. Feedback in bulk can be homogeneous, so that many voices repeat one error. It can be non-discriminating, so that praise and hostility alike leave the relative standing of competing explanations unchanged. It can be procedurally honored and substantively ignored. And it can be adjudicated by standards that already presuppose the conclusion under challenge.

The aim here is to give these observations a formal structure whose claims are no wider than its assumptions. The object of study is neither artificial intelligence nor independent research as such. Both appear below as applications, and both make the general problem unusually visible, but the theory concerns any system in which beliefs about consequential claims are meant to be revised in response to challenge: scientific institutions, peer review, formal verification, organizational learning, and collective belief revision more broadly. The central notion is adversarial epistemic closure, defined in Section 6. The term requires care, because closure can suggest either an undesirable impermeability to criticism or a desirable completion of the correction loop. The paper uses the word in the second sense and gives the first a separate name, epistemic insulation.

The contribution is deliberately modest in one respect and demanding in another. It does not claim that research in general lacks the capacity for correction. That would be a consequential sociological claim requiring its own evidence, and the framework developed here would demand that it be exposed to challenge. It claims instead that the conditions under which a system is correctable can be stated precisely, that they are not reducible to the presence of reviewers, the prestige of reviewers, or the intensity of disagreement, that two of them yield results with definite mathematical content, and that apparently rigorous systems can fail them in identifiable ways. Section 9 says what the account adds to existing frameworks, and Section 10 gives the test by which that addition should be judged.

2 Feedback Is Not Information

Praise, hostility, consensus, and repetition are all forms of feedback. None of them necessarily discriminates between hypotheses. A reviewer who declares a manuscript brilliant has provided a signal about the reviewer’s disposition and perhaps about the manuscript’s polish, but has not supplied an observation that bears differently on the manuscript’s central claim than on its best rival. A reviewer who declares it worthless has the same limitation. The relevant measure of an observation is how much it changes the relative support for competing explanations.

Two meanings of the symbol for a theory must be kept apart. Let Θ be a set of candidate theories and let T be a random variable taking values in Θ , representing uncertainty about which candidate is correct. Let $\theta_0 \in \Theta$ denote the incumbent theory, the one currently accepted by the process. Let E denote the body of evidence already available and R a review procedure that, when run, yields an observation e . The expected information gain of the procedure, relative to the existing evidence, is

$$\mathcal{I}(R \mid E) = \mathbb{E}_{e \sim P(e \mid E, R)} \left[D_{\text{KL}}(P(T \mid E, e, R) \parallel P(T \mid E, R)) \right] = I(T; e \mid E, R), \quad (1)$$

where the expectation is taken over the predictive distribution of the review outcome given the existing evidence, and the last expression is the conditional mutual information between the candidate theory and the review outcome. When this quantity is zero, the review outcome is independent of which candidate is correct, given what is already known, and the procedure produces feedback without information about T .

The qualification in that last sentence is essential. Expression (1) is defined only relative to the hypothesis space Θ and the probability model that assigns likelihoods to review outcomes. Two limits follow. First, it measures belief change and not truth-directed improvement. Under a misspecified likelihood, a misleading review can produce substantial information gain, because the posterior moves a great deal in response to an observation whose bearing has been computed incorrectly. Second, and more consequentially for a paper concerned with foundational criticism, a review may reveal a flaw that the existing hypothesis space cannot represent: an unconsidered alternative, a mislocated boundary of the domain, a concept that should not have been posited. Such a review can have zero measured information gain about T and still be the most valuable observation the process receives, because its effect is to enlarge or restructure Θ . The formalism therefore requires two channels. Within-space discrimination is measured by (1). Space revision is an event in which Θ changes, and it cannot be measured by any quantity defined on a fixed Θ . The remainder of the paper treats the first channel quantitatively and the second through the revision graph of Section 7, where a vertex is a formulation of the theory and its scope, and so can change the hypothesis space itself.

A second criterion is needed for either channel, concerning the evidentiary reliability of the review and the calibration of the revisions it prompts. A revision policy is a map from evidence and incumbent theory to a revised theory. The policy is calibrated, informally, when revisions of a given strength are warranted by evidence of corresponding strength, as judged by an external standard that can be consulted after the fact, such as later proof, replication, or resolution by independent means. The paper does not assign numerical values to calibration, because no defensible way of doing so is available at this level of generality. It uses the concept as a requirement that any claim of correctability must eventually be checked against a record of how past revisions fared.

3 Correlated Criticism

Multiple reviewers are not necessarily independent. They may share intellectual traditions, institutional incentives, reference works, training data, search engines, or generative models. When they do, agreement among them can create an illusion of confirmation without supplying much additional evidence. The point is old in the statistical literature on pooled judgments and has been given sharper form in work on diversity and collective problem solving [9] and on the way social influence degrades aggregate judgment [16]. The question here is what can be said precisely.

The right quantity is the incremental contribution of an individual review. For reviews $e_1, \dots, e_n$ taken in any order, define

$$\Delta I_i = I(T; e_i \mid E, e_1, \dots, e_{i-1}). \quad (2)$$

By the chain rule for mutual information, the joint information is the sum of the increments,

$$I(T; e_1, \dots, e_n \mid E) = \sum_{i=1}^n \Delta I_i, \quad (3)$$

and each increment is nonnegative. Two facts follow immediately. The increments measure what a review adds given what the earlier reviews already supplied, so a review that repeats earlier ones has an increment near zero however informative it would have been alone. And, because increments can be positive even when reviewers share substantial background assumptions, shared training or tradition does not by itself eliminate independent information. The correct claim is therefore narrower than the intuition that correlated reviewers are worthless. It is that the number of reviews does not determine their incremental evidentiary contribution. That contribution depends on how much each review adds to what the others already carry, and this is a property of the joint distribution, not of the count.

When the dependence among reviewers is complete mediation through a shared source, much sharper statements become available. Let S be a shared upstream variable, which may be a common literature, a common school of thought, or the output of a single model that every reviewer consulted. Complete mediation means that, given the existing evidence, reviews depend on the theory only through S , so that $T \rightarrow S \rightarrow (e_1, \dots, e_n)$ is a Markov chain conditional on E . Complete mediation is a strong assumption, and real reviewers typically violate it by running their own calculations, searches, checks, and experiments. It is stated here as a limiting case that fixes what review volume can and cannot do, and the departures from it are exactly where independent contributions enter.

Lemma 1 (Source cap with private contributions). *For arbitrary reviews $e_1, \dots, e_n$ and an arbitrary shared variable S ,*

$$I(T; e_1, \dots, e_n \mid E) \leq I(T; S \mid E) + \sum_{i=1}^n I(T; e_i \mid E, S, e_1, \dots, e_{i-1}). \quad (4)$$

Under complete mediation every term of the sum vanishes, and the bound reduces to $I(T; e_1, \dots, e_n | E) \leq I(T; S | E)$, the data processing inequality [2].

Proof. Adding S to the observations cannot decrease information, so $I(T; e_{1:n} | E) \leq I(T; e_{1:n}, S | E)$. The chain rule gives $I(T; e_{1:n}, S | E) = I(T; S | E) + I(T; e_{1:n} | E, S)$, and a second application of the chain rule expands the last term into the sum in (4). Under complete mediation each summand is zero because e_i is conditionally independent of T given S and E . $\square$

The summand in (4) is the private contribution of review i , the information it carries about T that is not available through the shared source. The lemma states that a collection of reviews can exceed the information in the shared source only by the sum of such private contributions. This is the formal content of the distinction between reviews that merely restate a common background and reviews that execute independent checks.

The consequence for correction is a floor on error that review volume cannot lower. Suppose Θ is finite, and suppose the process identifies the correct theory by some decision rule applied to the reviews.

Theorem 1 (Volume floor). *Let Θ be finite with $|\Theta| \geq 2$, and let $\hat{T}$ be any estimator of T based on E and the reviews $e_1, \dots, e_n$. If reviews are completely mediated by S , then for every n the error probability satisfies*

$$\mathbb{P}(\hat{T} \neq T | E) \geq \frac{H(T | E) - I(T; S | E) - 1}{\log_2 |\Theta|}, \quad (5)$$

with entropies measured in bits. In particular, if $I(T; S | E) \leq H(T | E) - 1 - \delta$ for some $\delta > 0$, then the error probability is at least $\delta / \log_2 |\Theta|$ at every review volume. In addition, the increments ΔI_n tend to zero as $n \rightarrow \infty$, so that the marginal contribution of an additional review vanishes.

Proof. Fano's inequality gives $\mathbb{P}(\hat{T} \neq T | E) \geq (H(T | E, e_{1:n}) - 1) / \log_2 |\Theta|$. By definition of mutual information, $H(T | E, e_{1:n}) = H(T | E) - I(T; e_{1:n} | E)$, and the Lemma under complete mediation bounds the subtracted term by $I(T; S | E)$, which gives (5). For the second claim, the increments are nonnegative and, by (3), sum to at most $I(T; S | E) \leq H(T | E) < \infty$, so they form a convergent series and tend to zero. $\square$

The theorem is conditional, and its conditions are the point. It does not say that large review volumes are useless. It says that when reviews are informationally mediated by a common source, volume cannot reduce the probability of mistaking the true theory for another below a floor fixed by what the source carries, and that this holds even if each individual review is informative taken alone, since ΔI_1 may be large. The second statement, that increments vanish, would hold in any finite Θ because total information is bounded by the prior entropy. The distinctive content of the theorem is that mediation can place the cap far below that bound. Review volume improves correctability only to the extent that reviews carry private contributions, which is why Section 13 classifies reviews by private contribution and not by number.

A standard illustration of the same phenomenon in a special case is the effective number of independent reviewers. For n reviewers whose errors have equal variance and a common pairwise correlation r , the variance of the mean error is that of $n_{\text{eff}} = n/(1 + (n - 1)r)$ independent reviewers, and n_{eff} approaches $1/r$ as n grows. This is a property of equicorrelated estimators and is invoked only to show the shape of the effect, which is saturation. It relies on a Gaussian-like error model that the information-theoretic results above do not need.

Independence, finally, should not be treated as an intrinsic epistemic virtue. Independent criticism is not necessarily reliable criticism. An independent reviewer can be incompetent, poorly informed, or committed to an unsound inference rule, and a collection of independent bad reviewers does not converge on anything useful. The value of independence lies in what errors it separates. Two reviewers who share a training lineage but differ in access to data may be independent with respect to evidentiary error and dependent with respect to interpretive error. Two reviewers who disagree sharply may still be dependent if both rely on the same flawed benchmark. The appropriate question is therefore which sources of error a panel's composition decouples, and whether those are the sources most likely to afflict the claim under review.

4 A Taxonomy of Objections

Objections differ in what they challenge, and treating them as points on a single severity scale conceals the differences. Five kinds are distinguished here. They are not degrees of seriousness, since a local correction to a central lemma may matter more than a comparative objection to a peripheral remark. They are different attacks on different aspects of a theory.

A local correction identifies an error in an individual step, calculation, citation, or statement, and challenges internal correctness. Its proper outcome is repair, and the theory's structure is typically untouched. A structural objection challenges the validity of an inference, showing that a conclusion does not follow from its stated premises even when each premise is granted. Its proper outcome is either a repaired argument with additional premises, which changes the theory's assumptions, or a demotion of the conclusion's status. An external counterexample challenges empirical adequacy, exhibiting a case within the theory's stated domain on which it gives the wrong answer. The proper outcome is typically a narrowing of the domain or a modification of the model, and only rarely abandonment. A comparative objection challenges explanatory distinctiveness, showing that an existing account already explains the phenomena equally well, so that the theory's claim to originality or to unique explanatory power is overstated. A necessity objection challenges conceptual economy, arguing that some posited structure does no work and that the theory's predictions survive its removal.

The taxonomy matters because the appropriate revision differs by kind, and because a system can be responsive to some kinds and sealed against others. An institution that reliably accepts local corrections but never accommodates necessity objections is correctable in a restricted sense that may leave its most consequential assumptions untouched. The measure of a system's correctability must therefore be indexed by the kinds of objection it can absorb, and

a system should not be certified as correctable in general on the strength of its performance on the easiest kind. The definitions below carry this index, and Section 10 shows that it is what allows the framework to separate processes that conventional measures cannot.

5 The Economics of Criticism

Serious review is costly. It requires time, expertise, and the willingness to understand a theory on its own terms before attacking it. The benefits of a successful review are disproportionately captured by the author, who receives a better theory, and by later readers, who receive a more reliable literature, while the reviewer receives little that can be priced. This is a public-goods structure [11]. The predictable result is that valuable criticism is undersupplied, while inexpensive praise, which costs little and purchases goodwill, and performative opposition, which costs little and purchases visibility, proliferate.

Remedies exist, and each introduces its own difficulty. Financial incentives for finding errors make review profitable but invite reviewers to optimize for what is rewarded. Reputation systems make careful review socially valuable but tend to reward the appearance of rigor, for example the length and confidence of a review, over its discriminating content. Adversarial markets for objections, in which challengers stake something on the failure of a claim, align incentives more sharply but require a settlement procedure, and the settlement procedure becomes the new site where correlated error and circular adjudication can enter. Each mitigation is subject to the pattern commonly summarized as Goodhart's law, that a measure used as a target ceases to measure what it measured [7, 18]. The economic analysis therefore leads back to the structural question: whatever incentive scheme is adopted, correctability depends on whether the quantity being rewarded tracks discriminating evidence.

6 Adversarial Epistemic Closure

Closure and insulation must be distinguished at the outset. Epistemic insulation is the condition in which counterevidence cannot affect foundational commitments, whether because the commitments are immunized by definition, because the relevant tests are unobtainable, or because the community that would conduct them has no standing. Adversarial closure, in the sense intended here, is the condition in which consequential claims have specified routes by which they can be challenged and revised, and in which those routes are in fact used. The first is a failure. The second is an achievement that must be maintained. The use of the same English word for both is unfortunate, and the terminology should be read with the contrast in view.

A definition cannot require that every consequential claim be falsifiable by an independent observer. Some claims are mathematical and are challenged by proof or by counterexample rather than by observation. Some are interpretive and are challenged by comparison with rival readings that the evidence does not decisively adjudicate. Some depend on evidence that cannot presently be obtained. A definition that ignored these cases would classify most of

mathematics and the humanities as insulated, which is absurd, and would classify an untestable empirical conjecture as closed merely because a test is conceivable. The definition therefore indexes the admissible forms of challenge to the kind of claim. Let $\mathcal{C}$ be a class of consequential claims. For each claim $c \in \mathcal{C}$ let $K(c)$ be the set of the five kinds of objection, drawn from the taxonomy above, that the nature and maturity of c make admissible, so that a conjecture whose decisive test is not yet possible has a smaller $K(c)$ than a mature empirical claim. For each kind $k \in K(c)$ let $\text{Ch}_k(c)$ denote the specified challenge procedures of that kind, which may include proof checking, formal verification, empirical replication, comparative analysis, or deferred tests with stated conditions.

A procedural account will not suffice, and the reason is instructive. Imagine an institution that publishes every objection, maintains a low-friction submission process, recruits independent reviewers, and keeps a public revision ledger. Every criticism receives a reasoned response. Yet the institution never changes its central theory. This is not necessarily a failure, because the objections may simply be unsuccessful. Conversely, the institution might accept frequent superficial revisions while preserving an indefensible central assumption. The existence of a revision mechanism cannot establish the quality of its operation. Three properties must be distinguished.

For accessibility, the difficulty of a challenge must be separated into two parts. The intrinsic cost of a challenge is the cost that any process would impose given the maturity of the problem and the state of the art: a conjecture that needs decades of research to resolve has a high intrinsic cost of challenge, and that fact alone does not make the community that studies it insulated. The imposed cost is the additional cost the process places on challengers beyond the intrinsic one, through restricted access to data or venues, gatekeeping of eligibility, delays, fees, or exposure to retaliation. Let r denote a stated resource level for a competent challenger.

Definition 1 (Adversarial accessibility). *A claim c is adversarially accessible with respect to kind $k \in K(c)$ at resource level r if the imposed cost of formulating, submitting, and having evaluated a challenge of kind k does not exceed what a competent challenger at level r can bear.*

The index removes two ambiguities. Accessibility is a property relative to a kind, so the question of whether one admissible kind or every admissible kind must be accessible is answered by stating the set of kinds over which closure is claimed, as in the definition of closure below. And it is relative to imposed cost, so that difficult problems are not classified as defective processes merely because they are difficult.

Definition 2 (Adversarial discriminability). *A challenge procedure is adversarially discriminating with respect to c and an alternative c' if its expected information gain between c and c' , in the sense of (1) restricted to the pair, is positive, or, for a challenge that restructures the hypothesis space, if it exhibits a candidate absent from the space that is not dominated by c on the stated criteria. In either case the reliability of the procedure must have been independently assessed.*

Responsiveness is where an all-or-nothing formulation fails. An institution that responds appropriately to nine hundred ninety-nine consequential objections and mishandles one would

fail a universal requirement completely, and an institution may be unable to act on an objection because its warrant cannot be established until much later. The normative requirement and its empirical measurement should therefore be separated. The normative requirement is that warranted objections ought to produce proportionate revision. The empirical question is how reliably, and with what delay, a given process does so. Warrant is judged by an external standard κ , such as later proof, replication, or independent resolution, applied after the fact and specified independently of the process being assessed. Let $t_w(x)$ be the time at which the warrant of a challenge x becomes establishable under κ , so that delay is not charged to a process for evidence that did not yet exist.

Definition 3 (Adversarial responsiveness). *For a process Π , a kind k , and a horizon τ , the responsiveness of Π is*

$$\mathcal{R}_\tau(\Pi, k) = \Pr\left(\text{proportionate revision within } \tau \text{ of } t_w(x) \mid x \text{ is a warranted challenge of kind } k \text{ received by } \Pi\right), \quad (6)$$

where proportionate revision means a change to the claim, its scope, or its epistemic standing lying within the range of defensible responses to a challenge of that kind and strength.

Responsiveness is thereby a graded institutional property that can be compared across institutions and across time, and that is stratified by kind, so that a high value for local corrections cannot conceal a low value for necessity objections. It is a probability over a population of challenges, which means it can only be estimated from a track record of challenges that were in fact received, and a process without such a record has an undefined value, not a high one.

Definition 4 (Adversarial epistemic closure). *A research process Π exhibits adversarial epistemic closure of degree at least ρ at horizon τ and resource level r , with respect to a class of consequential claims $\mathcal{C}$ and a set of kinds $\mathcal{K}$, if every claim $c \in \mathcal{C}$ is adversarially accessible at level r for each $k \in \mathcal{K} \cap K(c)$, the challenge procedures for those kinds are adversarially discriminating, and $\mathcal{R}_\tau(\Pi, k) \geq \rho$ for each such k .*

Closure in this sense is graded, and the ideal binary notion is the limiting case $\rho = 1$ over all admissible kinds. The three properties remain related but not interchangeable. Accessibility without discriminability yields noise, a heavy traffic of objections that settles nothing. Discriminability without accessibility yields criticism that exists in principle and is unavailable in practice. Accessibility and discriminability with low responsiveness yield institutional theater, in which the forms of revision are performed and the substance is withheld.

The word proportionate carries real weight. A single counterexample may invalidate a universal theorem without destroying the research program in which it appeared. A comparative objection may reduce a theory's claim to originality without touching its internal correctness. An empirical anomaly may require narrowing a model's domain rather than discarding its explanatory structure. Proportionality is not a license for minimal revision. It is a requirement

that the response lie within a defensible range for the kind and strength of the challenge, and that both under-reaction and over-reaction count as failures [13]. The range cannot be specified in advance for every case, and Section 8 explains why its specification must be explicit enough to be audited.

7 Reachability and Realization

A graph-theoretic formulation separates two questions that discussions of correctability tend to run together. Represent a research process as a directed graph $G = (V, E)$ in which each vertex is a publicly identifiable theoretical state, meaning a specific formulation of a claim set together with its stated scope, its hypothesis space, and its epistemic status, and each edge is an admissible transition induced by evidence or criticism, labeled by the kind of challenge that licenses it.

Definition 5 (Revision reachability). *A claim set at vertex v has revision reachability with respect to kind k if there exists a path in G from v to a vertex v' whose claim set differs substantively from that at v , where every edge on the path is labeled by a challenge of kind k .*

Reachability is a statement about the structure of the process. It says that a route exists. It does not say that the route is ever taken. A system can possess theoretically available transitions that are never followed, because the evidence that would trigger them has not arisen, because it has arisen and been dismissed, or because the standards of dismissal are themselves untouchable.

Realization is the empirical counterpart, and is measured by $\mathcal{R}_\tau(\Pi, k)$ in (6). It can only be assessed by observing what happens when warranted evidence actually arrives. The distinction explains why inspection of a system's rules cannot certify its correctability. Rules establish at most reachability. Realization requires a track record, and a track record requires cases in which a warranted challenge was in fact presented. A young system, or a system whose claims have not yet been seriously challenged, can claim reachability and cannot yet claim realization. Treating the first as the second is a recurring error in assessments of institutional reliability.

Three observations follow. First, reachability is cheap to assert and costly to falsify, which is why it is the typical claim of systems that wish to be thought correctable. Second, realization can fail in a manner that is invisible from within, since the process may regard each of its non-revisions as justified. Third, realization is observable only relative to a standard κ of what was warranted, and that standard must be specified independently of the process being assessed. The next section concerns the standards by which the process itself decides.

8 Adjudication Without Vicious Circularity

The problem of the author as ultimate judge is the oldest form of self-sealing. If the only party entitled to determine whether an objection is warranted is the party whose theory is challenged, then the theory may be protected by construction. The opposite remedy, surrendering judgment

entirely to critics, is equally defective, since critics can be mistaken, hostile, or confused. The task is to identify a requirement that neither privileges the author nor dissolves authorial judgment, and the requirement must not prohibit what is legitimate.

Some dependence of adjudication on prior commitments is unavoidable and proper. A proof is evaluated within an axiomatic framework, and an objection to a theorem is assessed using other propositions from that framework. An experimental result is evaluated using accepted background theory about the instruments. Let B denote the background framework: the body of premises, methods, and evidentiary standards accepted as common ground by the parties to a challenge, and not themselves the subject of that challenge. Let σ denote the standards by which a process decides whether a challenge x to a claim c is warranted, and let $\text{Prem}(\sigma, x)$ be the set of premises on which the verdict of σ on x depends.

Definition 6 (Vicious circularity). *The adjudication of a challenge x to a claim c under standards σ is viciously circular if $c \in \text{Prem}(\sigma, x)$ and $c \notin B$, so that the contested proposition itself, and not merely the common background, is used as a premise in deciding whether the challenge to it is warranted.*

The definition draws the line between an assumption being used as part of an accepted framework, which is legitimate, and the specific contested proposition being used to dismiss an objection to itself, which is not. A theorem is not circularly defended by the axioms from which it follows, since the axioms are in B and the theorem is not among the premises of its own adjudication. A theory that excludes any evidence incompatible with its central claim, by classifying such evidence as artifact or noise on the basis of the claim, is circularly defended.

The harm done by circularity can be stated exactly. Say that a challenge x is incompatible with c if the content of x , taken as warranted, entails that c is false.

Lemma 2 (Circularity blocks incompatible challenges). *Suppose the adjudication standards σ are consistent and viciously circular with respect to c , that is, $c \in \text{Prem}(\sigma, x)$ and $c \notin B$ for every challenge x to c . Then no challenge incompatible with c can be adjudicated warranted under σ .*

Proof. To adjudicate x warranted is to accept its content. If x is incompatible with c , accepting its content together with the premise c yields a contradiction. A consistent standard that treats c as a premise therefore cannot accept the content of x , and the verdict is that x is not warranted. $\square$

The consequence for the framework is that revision reachability, for every kind of challenge whose successful outcome would contradict c , is vacuous under circular adjudication, however many edges the graph nominally contains. This replaces the earlier, looser claim that standards must be free of all dependence on the contested assumption.

Proposition 1 (Non-circularity is necessary, not sufficient). *Let $\mathcal{R}_\tau(\Pi, k) > 0$ for some kind k of challenge incompatible with c . Then the adjudication of such challenges is not viciously circular with respect to c . The converse fails: non-circular adjudication does not imply adversarial*

epistemic closure, since a process may adjudicate without circularity and still lack accessibility, discriminability, or sufficient responsiveness.

Proof. The first claim is the contrapositive of the Lemma, since positive responsiveness for incompatible challenges requires that some such challenge be adjudicated warranted. For the converse, take a process whose standards are explicit and non-circular and that imposes prohibitive costs on challengers; it fails accessibility and hence closure. $\square$

The condition has an operational form. It requires that the grounds of adjudication be explicit enough that a third party can identify circular or inconsistent applications. If a process accepts a given kind of evidence against a rival's claim and rejects the same kind of evidence against its own, a third party can detect the asymmetry only if the standard is written in a form that permits comparison. In interpretive fields where full explicitness is unattainable, the requirement weakens to this: the grounds on which a challenge was rejected must be stated in terms that do not presuppose the interpretation under challenge, so that a competent third party who does not share the interpretation can evaluate whether the rejection was reasonable. Standards drift, in which criteria for acceptance are adjusted after a challenge arrives so as to exclude it, is the characteristic violation, and it is detectable only when earlier standards were recorded.

A further question is whether circularity can be avoided simply by having the incumbent adopt good standards. It cannot be guaranteed, for a structural reason. Let a challenge be externally warranted if it is warranted under κ . Let Σ^* be the class of standards that a process could adopt.

Proposition 2 (No self-governed guarantee). *Assume (F), the fallibility of standards: for every $\sigma \in \Sigma^*$ there exists a possible challenge x_σ that is externally warranted and that σ adjudicates as unwarranted. Assume (X), self-governance: the standard in force at each stage is revised only through challenges that the standard then in force adjudicates as warranted. Then, at every stage, there exists an externally warranted challenge that the process rejects. Hence no process satisfying (X) can guarantee correction across the whole class of possible challenges.*

Proof. Let σ_t be the standard in force at stage t . By (F) there is a challenge x_{σ_t} that is externally warranted and rejected by σ_t . By (X) the standard changes only through challenges that σ_t accepts, and x_{σ_t} is not among them, so at stage t it remains rejected. The argument applies at every t . $\square$

The proposition is close to a diagonal argument and should be read as such. Its force lies in the assumption (F), which is plausible for any explicit standard that is itself a possible source of error and cannot be established as an unconditional truth. What it shows is limited and useful: a guarantee of correction over a broad class of possible errors requires that standards be revisable by some procedure not wholly governed by the standards in force, that is, by an input from outside the incumbent. This is the structural reason why independent criticism matters, and it holds regardless of the incumbent's good faith. It does not show that self-governed processes always fail, only that none can certify its own success.

9 Position Relative to Existing Frameworks

An account of this kind must say why it is not an existing framework in new vocabulary. Three neighbors are closest.

Belief-revision theory in the AGM tradition [1] studies revision operators on belief sets and the rationality postulates such operators should satisfy, and relates revision to an ordering of epistemic entrenchment that determines which beliefs yield first when a new input conflicts with the old ones [4]. The framework developed here is compatible with that apparatus and addresses what it takes as given. AGM theory presupposes that a conflicting input reaches the operator and that the entrenchment ordering is fixed. The questions of whether a challenge can reach the operator, who controls the entrenchment ordering, and whether the ordering was set by the proposition it protects are exactly what accessibility and vicious circularity formalize. A circular standard is, in these terms, an entrenchment ordering that ranks the contested claim above every input that could contradict it, set by the party who holds the claim. The new content is the assignment of control over the ordering and the audit of whether it is applied symmetrically.

Bayesian updating gives the canonical account of how evidence should shift belief given a fixed hypothesis space and fixed likelihoods, and the framework uses it in the quantity (1). Its known limitation is the problem of new theories, in which evidence is best explained by a hypothesis absent from the space [5]. The distinction drawn in Section 2 between within-space discrimination and space revision is a response to that limitation, and the treatment of standards as objects of challenge addresses the further problem that likelihoods and priors are themselves supplied by the party being assessed.

Social epistemology supplies the closest body of work. It treats the communication structure of epistemic communities [19], the division of cognitive labor [11, 12], the social conditions under which criticism is transformative [15], the dependence of knowers on one another [8], and the institutional epistemology of testimony and expertise [6]. Its models of network structure and diversity bear on the correlated-criticism results of Section 3, which can be seen as an information-theoretic account of one mechanism within that literature. What it does not typically provide is the separation of imposed from intrinsic cost of challenge, the stratification of responsiveness by kind of objection, the circularity condition on adjudication, or the requirement that a theory of correctability state its own failure conditions.

The distinguishing contribution is therefore a combination: a graded, kind-stratified measure of realized revision judged against an external standard, an audit condition on the standards by which challenges are adjudicated, and an information-theoretic account of when added reviewers add evidence. Whether the combination can be reduced to existing frameworks is an empirical and conceptual question the paper does not claim to have closed, and the next section proposes how it should be tested.

10 A Discriminating Test

The framework earns its place only if it distinguishes cases that existing accounts cannot. Two processes, Π_A and Π_B , are constructed to look equally correctable by every conventional measure of criticism. They receive the same volume of reviews from reviewers of the same independence, in the sense that the joint distributions have the same ΔI_i profiles. They acknowledge and respond to the same fraction of objections, maintain the same public ledger, accept the same proportion of local corrections, and update beliefs by the same Bayesian rule over the same hypothesis space. They differ in one respect. In Π_A the standards of admissibility for evidence are written down, recorded before challenges arrive, and independent of the central claim c . In Π_B the standards classify as admissible only evidence compatible with c , which is to say that $c \in \text{Prem}(\sigma, x)$ for every challenge x .

Every conventional measure coincides across Π_A and Π_B . The proposed quantities do not. By the Lemma of Section 8, $\mathcal{R}_\tau(\Pi_B, k) = 0$ for every kind k whose successful outcome would contradict c , which includes structural, external, and necessity challenges directed at c , while $\mathcal{R}_\tau(\Pi_A, k)$ can be high for the same kinds. The stratified responsiveness and the circularity audit therefore separate processes that the aggregate acceptance rate, the volume of review, and the existence of a ledger cannot. This pair establishes distinguishability. It is constructed, and it shows that the framework contains a distinction that the other measures lack. It does not by itself show that the distinction matters.

Whether the distinction predicts a consequential outcome is an empirical claim, and it must be stated so that it can fail and so that its evaluation does not depend on the framework's own definitions. This section therefore specifies a retrospective comparative study in enough detail to be preregistered and replicated. Its design principle is that every quantity the framework proposes as a predictor is measured from materials that existed before the challenge it is meant to predict, while every outcome is determined afterward by a procedure that cannot see the predictors and does not rely on what the challenged party eventually said.

10.1 Estimand and unit of analysis

The unit of analysis is a claim-challenge pair: a single consequential claim, such as a theorem, an empirical result, or a stated design or modeling assertion, together with one documented challenge to it. It is not a paper, a journal, or an institution, because the framework's quantities are defined claim by claim and kind by kind, and because aggregating to the institutional level would let strong performance on easy challenges conceal weak performance on deep ones. Several challenges can attach to one claim and several claims to one community, so all analyses treat pairs as clustered within claims and within communities. The primary estimand is the stratified responsiveness $\mathcal{R}_\tau(\Pi, k)$ of (6), estimated over challenges that an independent panel judges warranted, for a fixed horizon τ set in the registration, and its dependence on the pre-challenge process characteristics defined below.

10.2 Sampling frame and rules

The frame consists of domains in which challenges and responses are publicly documented, dated, and archived with verifiable timestamps. Candidates include mathematics, where errata, counterexamples, and published refutations of claimed results are recorded; fields with versioned preprints and public comment histories; journal exchanges of comment and reply; and open software projects whose issue trackers and design documents record challenges to stated design claims. Within each chosen domain and period, all documented challenges meeting the inclusion rules are enumerated by a systematic search whose queries and dates are recorded. A challenge is included if it is public, dated, specific enough to identify the claim it addresses, and states its grounds. It is excluded if it is spam, abuse, an exact duplicate of an earlier challenge to the same claim, or a question without an asserted objection. Sampling is stratified by community and by challenge kind from the taxonomy of Section 4, and it is conducted on the basis of the documentary record at the time of challenge and never on the basis of what happened afterward. This last rule is the main guard against selection on the outcome. A sample drawn from claims known to have been corrected would overstate the association between process characteristics and correction, since it would exclude every case in which a warranted challenge failed. The sample must therefore include challenges that were rejected, ignored, or left unresolved.

10.3 Pre-challenge measures

For each pair, coders measure the characteristics of the process in force on the challenge date, using only materials dated before that date, which are verified from archival timestamps and truncated at the challenge date before being given to coders. Coders are blind to the outcome and to the panel judgments. Five groups of measures are coded against a written codebook. The explicitness of adjudication standards records whether admissibility and evidentiary standards were written down and dated before the challenge. The circularity audit records whether the standards in force presupposed the contested claim, applying the definition of vicious circularity in Section 8 to the recorded standards, with the shared background framework B identified from the documents. Accessibility records the imposed cost of challenge at a stated resource level, separated from the intrinsic cost, with indicators such as eligibility restrictions, data access, fees, and documented retaliation. Independence and private contribution record the provenance of the reviews the process had received on related claims, namely which sources and which executed checks contributed, from which the increments of Section 3 are approximated. Finally, the conventional measures are coded in the same way: review volume, a rating of review quality, institutional resources, and the existence and use of established correction practices such as erratum and replication norms. Two coders work independently on every pair, disagreements are resolved by a third, and agreement statistics are reported by measure.

10.4 Independent adjudication of warrant

Whether a challenge was warranted must be decided without reference to how the challenged party responded, or the study would be circular in exactly the way the framework warns against.

A panel of at least three domain experts assesses each sampled challenge. Panelists are not affiliated with the originating group or the challenger and have no history of collaboration with either, and they work independently before reconciliation. They receive the claim, the challenge, and the evidence available on the challenge date, together with any evidence that became available later and was established by sources independent of the challenged party, since this is part of the external standard κ and fixes t_w . The challenged party’s response, and every later revision or refusal to revise, is redacted from their materials. They apply a rubric written in advance, based on the evidence and on the standards of the field, which classifies the challenge as warranted, not warranted, or indeterminate, rates its kind and its severity, and records the date at which its warrant became establishable. Reconciliation follows a prespecified rule, under which a split verdict is classified as indeterminate and the pair is excluded from the primary estimand. Reliability is estimated by having a second panel assess a random subsample, and if agreement falls below the threshold fixed in the registration, the study is declared uninformative and not supportive.

10.5 Outcomes and observation window

Outcomes are determined after the panel’s judgment, from the documentary record, by coders who see the panel’s classification but not the pre-challenge measures. The response of the process is classified into one of five categories: substantive revision of the claim, scope restriction, demotion of epistemic status, justified rejection, and unresolved dispute. The panel’s rubric supplies the range of proportionate responses for each pair, and the response is coded as proportionate, insufficient, or excessive relative to that range, so that over-reaction is recorded and not treated as success. The primary outcome is whether a proportionate revision occurred within the fixed horizon τ measured from t_w , among challenges the panel judged warranted. A secondary analysis uses time to proportionate revision with right-censoring at the end of follow-up. Every pair must have at least τ of follow-up from t_w , and pairs with less are excluded by that rule and not by their outcomes, so that older challenges do not appear more correctable merely because they have had longer to accumulate revisions. Responses to challenges the panel judged not warranted are analyzed separately, as a measure of calibration: a process that revises as readily in response to unwarranted challenges as to warranted ones is pliable and not responsive.

10.6 Matching and comparison models

Within strata, pairs are matched on claim type, disciplinary context, challenge kind, panel-rated severity, evidentiary availability on the challenge date, meaning whether a decisive test was feasible then, the period in which the challenge arose, and institutional resources. Matching is performed by coarsened exact matching with regression adjustment as a robustness check, and both results are reported. The comparison that decides the question is predictive. A baseline model predicts the primary outcome from the conventional measures only: review volume, review quality, institutional resources, and established correction practices. An augmented

model adds accessibility, discriminability as assessed from the recorded challenge procedures, and the explicitness and circularity measures. The two are compared by out-of-sample predictive performance under a proper scoring rule, using cross-validation that holds out entire communities. The minimum improvement that would count in favor of the framework is fixed in the registration before any outcome is coded. No value is proposed here, because none can be justified without a pilot study that estimates the noise in the baseline.

10.7 Negative controls and sensitivity analyses

Four controls are specified. First, for local-correction challenges, which can be absorbed by almost any process, the framework predicts a weak association with the explicitness and circularity measures, and a strong association would suggest that the measures are tracking general institutional quality. Second, for challenges the panel judged not warranted, the framework predicts no association with revision, and one would indicate that the measures track pliability. Third, placebo exposures drawn from unrelated features of the standards documents, such as length and formatting, should not predict outcomes. Fourth, permutation of the exposure codes within matched strata gives the distribution of the performance gain under the null. Sensitivity analyses address differences between communities that the matching does not capture. They include community fixed effects, leave-one-community-out replication, and a bounding analysis that reports how strong an unmeasured confounder, associated with both the design of standards and the propensity to revise, would have to be to remove the estimated association.

10.8 Inference, limits, and prespecified failure criteria

The design can establish predictive associations. It does not identify causal effects without additional assumptions, in particular that processes with explicit and non-circular standards do not differ from others in unmeasured propensity to revise. A causal reading would require a natural experiment, such as the adoption of a recorded-standards policy at a venue at a known date, analyzed by comparing venues before and after adoption, and the paper proposes this only as an extension.

The registration states in advance the results that would count against the framework. It has failed its comparative test if the augmented model does not improve on the baseline by the preregistered margin, so that accessibility, discriminability, and independently assessed responsiveness add nothing to review quality, institutional resources, and established correction practices. It has failed its distinctive prediction if, in the stratum of challenges incompatible with the claim, processes coded as circular show the same responsiveness as processes coded as non-circular. It is not supported if the negative controls show associations comparable in size to the primary association, or if the result is not stable across communities. An inconclusive result, arising from low panel reliability or insufficient follow-up, is reported as inconclusive. The registration, the codebook, the coder training materials, the frozen archival records of each pair, the panel rubric, the coded data, the analysis code, and a log of deviations from the plan

are to be published in a versioned public repository before outcome coding begins, so that the study can be replicated and so that departures from the plan are visible. No figures are asserted in this paper, and none should be inferred from the constructed pair of processes above.

11 The Ways Rigorous Systems Become Self-Sealing

The framework yields a diagnostic vocabulary. A system can fail to be correctable in several distinct ways, and the usefulness of the theory depends on its capacity to identify which failure is operating rather than merely labeling the outcome.

Procedural theater occurs when accessibility is high and responsiveness is low for consequential kinds. Objections are solicited, logged, answered with courtesy, and left without effect. A diagnostic marker is a ledger in which the number of recorded objections grows while the number of recorded revisions to consequential claims remains near zero over a long period, together with the absence of any stated reason why none of the objections was warranted.

Superficial revision occurs when responsiveness is high for local corrections and low for structural, comparative, and necessity kinds. The process revises typographical errors, citations, and minor lemmas readily, while deeper objections receive replies that defend the original formulation. The marker is a skewed profile of $\mathcal{R}_\tau(\Pi, k)$ across the five kinds.

Scope inflation occurs when a theory responds to a counterexample by enlarging its claimed domain to absorb the case, converting a refutation into an apparent confirmation. Retreat of scope is the proportionate response to anomaly, and expansion without new support is its inverse.

Standards drift is the adjustment of acceptance criteria after a challenge has been presented, so as to exclude it. It is detectable only when earlier standards were recorded.

Reviewer pool curation occurs when the set of eligible critics is shaped, deliberately or through selection effects, so that the reviewers most likely to detect the relevant error are absent. Accessibility is formally present and materially empty, because the imposed cost is carried in eligibility.

Corpus-level correlation occurs when reviewers who appear independent are in fact conditioned on a common source, so that the cap of Section 3 applies to the degree that private contributions are absent.

Circular adjudication occurs when the contested claim is used as a premise in deciding challenges to it, and is the failure to which the circularity Lemma of Section 8 applies.

Reflexive exemption occurs when a framework for evaluating criticism declines to apply its own standards to itself. This is the failure that the present paper must be most careful to avoid, and Section 14 takes up the matter.

These are not exhaustive and they overlap. Their value is that each corresponds to a failure of one or more of the three properties, so that the framework offers a way to say where in the correction loop a given process fails and not merely that it fails.

12 Where Existing Mechanisms Succeed and Where They Fall Short

It would be an error to treat the framework as an indictment of existing practice. Several research traditions possess mechanisms that approximate parts of the proposed conditions, and the interesting question is how closely and where the approximation breaks.

Mathematics has a strong form of adversarial accessibility in that a proof is public and a counterexample, if it exists, can in principle be produced by anyone competent. Discriminability is also strong, since a valid counterexample decisively separates the claim from its negation. Responsiveness is typically strong for local errors, and the history of proofs and refutations shows how a community absorbs counterexamples by reformulating definitions and narrowing theorems [14]. The shortfall lies in the middle of the chain. For very long proofs, accessibility degrades because few people can check them, and agreement among those few may be correlated through shared training and shared intuitions about what is plausible. Formal verification addresses precisely this weakness by replacing human checking of derivations with mechanical checking against an explicit kernel, which supplies a private contribution that does not pass through the shared intuition. Even there, a gap remains between the formal statement that was verified and the informal claim that was intended, so that verification establishes correctness relative to a specification and leaves the fidelity of the specification to ordinary human judgment.

Experimental science has strong discriminability in domains where decisive experiments exist and weak discriminability where the Duhem–Quine difficulty applies, that is, where a failed prediction can be absorbed by adjusting auxiliary assumptions [3, 17]. The difficulty is also a place where circular adjudication can enter, since the choice of which auxiliary assumption to blame is a standard of adjudication and may be made in favor of the central claim. Replication practice improves accessibility, though replication is itself costly and therefore subject to the public-goods problem of Section 5. Responsiveness varies with how deeply the challenged result is embedded in other commitments, since results embedded in many downstream claims are costly to revise and therefore tend to be defended.

Software engineering offers an instructive partial model. Test suites make certain objections cheap and mechanical, code review makes others routine, and continuous integration enforces some responsiveness by blocking changes that fail specified conditions. What the practice illustrates is the importance of explicit failure conditions that are checked automatically, so that responsiveness does not depend on goodwill. Its limitation is that tests encode the author’s anticipations, and an error outside the anticipated space passes silently. A test suite is a standard that the author controls, and Proposition 2 applies to it: it cannot certify its own completeness.

Collaborative research that distributes subproblems among independent groups, with agreed public criteria for what each result must show, approximates the full set of conditions more closely than any individual institution, though it too is vulnerable to correlated error when the groups share methods. The pattern across these cases is consistent. The conditions are partly met in practice, partly by design and partly by accident, and what remains missing is generally

either private contribution from independent sources of error or explicitness of the standards by which warrant is judged.

13 AI-Assisted Scholarship

Generative systems alter the economics of criticism sharply. A review that once required hours of expert attention can be produced at negligible marginal cost, in any quantity, in any style, from any stated point of view. In the vocabulary above, this lowers imposed cost of challenge nearly to zero, and so raises accessibility to its ceiling. An author anywhere can obtain a detailed critique of a manuscript at any hour. The question is what it does to discriminability and to the private contributions on which, by the results of Section 3, the value of additional reviews depends.

The principal hazard is correlated error at scale. If a large number of reviews are generated by the same model, or by models trained on overlapping corpora with overlapping objectives, then the reviews are likely to be largely mediated by a shared source, and the cap of (4) bounds the information they jointly carry to the degree that they contain no private contribution. The claim is not that a thousand such reviews equal one, since individual reviews can carry private contributions through computation, retrieval, or independent derivation. It is that their number does not determine their increments, and that the increments are likely to be small precisely on the dimensions that matter most, namely shared assumptions about what a good argument looks like and which positions are plausible. The effective number of independent reviewers saturates quickly in the presence of high correlation.

A second hazard is simulated disagreement. A system can be prompted to produce objections from multiple perspectives, and the resulting text has the surface features of a diverse panel. The diversity is rhetorical, since each perspective is generated by the same underlying competence under different instructions. Treating such output as an adversarial process mistakes variation in style for variation in source of error. Agreement among simulated perspectives, and equally their staged disagreement, is evidence about what the model finds natural to say, and bears only indirectly on whether the claim is true.

A third hazard is artificial consensus. Authors who consult the same systems, and reviewers who use them to draft their evaluations, may converge on a shared set of objections and a shared set of accepted responses, producing an appearance of settled opinion whose agreement is manufactured by a common instrument. The convergence has the form of independent confirmation and lacks its content.

The question for this setting is whether increasing the quantity of criticism without increasing its independence can reduce epistemic reliability by producing false confidence. The framework suggests that it can. False confidence arises when the perceived weight of agreement is computed as if the reviews were independent while the true effective sample is small. An author who observes that many reviews raised no objection to a claim, and infers that the claim is robust, has made a calculation that is valid for independent reviewers and invalid for reviewers mediated by a common source. The more reviews are cheaply obtained, the larger the potential discrepancy

between the apparent and the actual evidentiary weight.

None of this implies that automated critique is useless. It can be a strong detector of local errors, inconsistencies, and gaps in exposition, which are the cases where shared sources are least harmful because the errors are not shared. The recommendation is that automated reviews be classified by the private contribution they supply and not by the volume they supply. Reviews that rely on mechanical checking against an explicit specification, such as proof assistants, test suites, and numerical replication, carry private contributions that do not pass through the model's training and should be weighted accordingly. Reviews that rely on the model's general judgment should be treated as a single largely mediated source, however many times they are sampled. Provenance, meaning a record of which sources and which executed checks contributed to each review, becomes a first-class requirement, because without it the mediation structure cannot be assessed and the increments cannot be estimated.

14 The Self-Referential Requirement

A theory of adversarial epistemic closure must itself specify what would count against it. Without such a specification the theory risks becoming another elegant account of why existing criticism is inadequate while exempting its own assumptions from comparable scrutiny, and this would be a clear instance of reflexive exemption. The requirement is therefore imposed on the framework as a condition of its own acceptability, and the conditions follow.

The framework would be undermined if systems that score well on accessibility, discriminability, and responsiveness, as these are operationalized here, failed to correct later-recognized errors at a rate comparable to systems that score poorly. The claim that the three properties jointly constitute correctability is a claim about their predictive relation to correction outcomes, and it can fail by that route. It would be undermined, in a different way, if systems lacking one of the three nevertheless corrected reliably, which would show that the property is not necessary, and the decomposition would need to be revised accordingly. It would be undermined if the retrospective protocol of Section 10 met any of its prespecified failure criteria, for instance if the explicitness and independence of standards added no predictive power for proportionate revision beyond review quality, institutional resources, and established correction practices. It would be undermined if the quantities could not be operationalized without circularity, for instance if every attempt to determine whether an objection was warranted relied on a judgment that already presupposed the answer, since the external standard κ would then be unavailable in principle. It would be undermined if the proportionality requirement turned out to admit, in each concrete case, any response whatever, so that it constrained nothing. The volume floor would lose practical significance if empirical study showed that real reviews almost always carry private contributions large enough that complete mediation is never approached, and the paper would then have shown a limiting case with few instances. And Proposition 2 depends on the fallibility assumption (F), so that a demonstration that some explicit standard is complete over a broad class of challenges would restrict its scope.

The framework also inherits a limitation that it cannot dissolve. The standard κ by which

warrant is judged *ex post* must be specified by someone, and the background framework B of Section 8 may itself embed assumptions. The conditions of the framework do not eliminate this regress. They bound it by requiring that standards be explicit and that they not take the contested claim as a premise, which moves the locus of possible error to the standard and makes that locus inspectable. A further level of review of the standard is possible and generates the same structure one level up. The regress ends in practice where independent communities with different commitments find the standards acceptable, and the theory makes no claim to end it more decisively than that.

A further instruction applies to the paper’s own sociological assertions. Statements about how often research systems fail to meet the conditions are empirical claims about institutions and require evidence of the kind the framework itself demands. The most demanding test is a case where a theory satisfies every procedural requirement and still resists a valid objection. If the framework can identify precisely why this happens, by locating the failure in discriminability, in responsiveness for a particular kind, in circular adjudication, or in mediation of the evaluators by a common source, rather than merely redescribing the failure afterward, it has begun to earn its status as a distinct theoretical concept. If it can only relabel such cases, it has not.

15 Conclusion

The epistemic quality of a research system is not adequately measured by the volume of feedback it receives, the prestige of its reviewers, or the intensity of its disagreements. It is measured by its demonstrated capacity to undergo warranted revision in response to discriminating evidence. The framework developed here makes that statement precise by separating accessibility, discriminability, and responsiveness, by defining responsiveness as a graded, kind-stratified quantity judged against an external standard, by distinguishing the existence of a revision path from its realization, by separating legitimate dependence on a shared background from vicious circularity in adjudication, and by treating independence as valuable only to the extent that reviews carry private contributions beyond what a shared source supplies.

Two results carry mathematical content. When reviews are mediated by a shared source, review volume cannot reduce the probability of error below a floor set by what the source carries. And no self-governed standard can guarantee correction across a sufficiently broad class of errors, which is why an input from outside the incumbent is structurally required and not merely desirable. The shift in emphasis is from a sociology of criticism toward a formal theory of correctability under correlated observation. The arrival of inexpensive, abundant, and highly correlated machine criticism raises the stakes of getting this structure right. When criticism is scarce, its mere presence is informative. When it is abundant, only its independence, its discriminating power, and its consequences for the claims it addresses can distinguish a process that learns from one that merely accumulates commentary.

References

- [1] Alchourrón, C. E., Gärdenfors, P., and Makinson, D. (1985). On the logic of theory change: partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2), 510–530.
- [2] Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley.
- [3] Duhem, P. (1954). *The Aim and Structure of Physical Theory*. Princeton University Press.
- [4] Gärdenfors, P., and Makinson, D. (1988). Revisions of knowledge systems using epistemic entrenchment. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufmann.
- [5] Glymour, C. (1980). *Theory and Evidence*. Princeton University Press.
- [6] Goldman, A. I. (1999). *Knowledge in a Social World*. Oxford University Press.
- [7] Goodhart, C. A. E. (1984). Problems of monetary management: the U.K. experience. In *Monetary Theory and Practice*. Macmillan.
- [8] Hardwig, J. (1985). Epistemic dependence. *The Journal of Philosophy*, 82(7), 335–349.
- [9] Hong, L., and Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46), 16385–16389.
- [10] Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6), 645–654.
- [11] Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy*, 87(1), 5–22.
- [12] Kitcher, P. (1993). *The Advancement of Science*. Oxford University Press.
- [13] Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In Lakatos and Musgrave (eds.), *Criticism and the Growth of Knowledge*. Cambridge University Press.
- [14] Lakatos, I. (1976). *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press.
- [15] Longino, H. E. (1990). *Science as Social Knowledge*. Princeton University Press.
- [16] Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22), 9020–9025.
- [17] Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, 60(1), 20–43.

- [18] Strathern, M. (1997). “Improving ratings”: audit in the British university system. *European Review*, 5(3), 305–321.
- [19] Zollman, K. J. S. (2007). The communication structure of epistemic communities. *Philosophy of Science*, 74(5), 574–587.