The Bottleneck Primitive: Statistics as the Study of Information Compression

Larsen James Close

February 2026

Abstract

Statistics is widely taught and practiced as a collection of techniques whose results frequently strike practitioners as counterintuitive. We argue that this is a pedagogical artifact, not a property of the subject. Every statistical operation is a projection through an information bottleneck, and every statistical result is a property of that bottleneck's geometry as much as of the data it compresses. Classical "paradoxes" — Simpson's paradox, the base rate fallacy, regression to the mean, the chronic misinterpretation of p-values — dissolve immediately once this is recognized. We demonstrate that the same dynamic operates well beyond textbook settings by analyzing two contemporary cases: adversarial attacks on AI safety classifiers, where attackers succeed by characterizing and navigating around the defender's projection geometry; and Anthropic's Sabotage Risk Report for Claude Opus 4.6, where an organization attempts to bound catastrophic risk through the deliberate triangulation of independent bottlenecks. The contribution is not the information bottleneck concept itself, which has formal precedent, but the demonstration that it constitutes the *ur-operation* from which the entire field of statistics derives — and that recognizing this reveals the analyst's cognitive topology as legibly as it reveals properties of the data.

Contents

1 The Pedagogical Failure

Statistics courses routinely produce students who can execute procedures — compute a t-statistic, run a regression, report a p-value — while fundamentally misunderstanding what those procedures tell them. The American Statistical Association found it necessary in 2016 to issue a formal statement clarifying what p-values mean, decades after the concept entered standard practice. Surveys of practicing scientists consistently find that majorities misinterpret confidence intervals (Haller & Krauss, 2002), confuse statistical significance with practical significance (Gigerenzer, 2004), and commit the base rate fallacy even after being warned about it (Casscells, Schoenberger, & Graboys, 1978).

The standard diagnosis is that humans have poor statistical intuition. We propose a different diagnosis: the field's self-presentation obscures a unifying structure that, once made explicit, renders every "counterintuitive" result immediately obvious. The structure is this: **every statistical operation is a projection through an information bottleneck, and every statistical result is jointly a property of the data and the bottleneck**. When practitioners confuse bottleneck-properties for world-properties, paradoxes appear. When the bottleneck is made visible, they vanish.

This is not a metaphor. We will show that the central concepts of statistics — estimation, hypothesis testing, sufficiency, regression, model selection — are each instances of a single operation: compress a high-dimensional data structure through a lower-dimensional aperture, and characterize what survives. The aperture is not the data's choice. It is the analyst's, and it reveals the analyst's model of relevance as legibly as it reveals the data's structure.

2 The Bottleneck Primitive

Consider a dataset X living in some high-dimensional space and a question Q that an analyst wishes to answer. No question can be answered by attending to all of X simultaneously — cognition, computation, and communication all impose finite bandwidth. The analyst must compress: select, aggregate, project, or otherwise reduce X to some tractable representation $T(X)$ from which Q can be addressed.

This compression is the statistical act. Everything else is implementation.

The **mean** is the most aggressive bottleneck possible: project the entire distribution onto a single point. What survives? Only location. What is destroyed? Everything else — shape, spread, multimodality, dependence structure. The **variance** quantifies the residual after projection onto the mean — it measures what a location-only summary cannot retain. It is not a second, independent quantity; it is the residual of the first compression.

Regression widens the bottleneck slightly: project Y onto the subspace spanned by X, retaining the conditional expectation $E[Y|X]$. What survives is the linear (or specified nonlinear) relationship. What is destroyed is everything orthogonal to it — which we call the "residual" and too often ignore, though it contains the entire story the model cannot tell.

Sufficient statistics are Fisher's formalization of the optimal bottleneck. A statistic $T(X)$ is sufficient for parameter θ if $P(X|T(X),\theta) = P(X|T(X))$ — that is, once you have T, the raw data tells you

nothing more about θ. The factorization theorem makes the key point explicit: sufficiency is defined *relative to the model class*, not the data. The data has no opinion about what is sufficient. The model does. Change your model and the sufficient statistic changes, because the bottleneck that preserves everything *your model can use* depends entirely on what your model is.

This is the first major consequence of the bottleneck framing: **there is no model-free statistical inference**. Every compression presupposes a relevance structure, and that relevance structure is the analyst's contribution, not the data's.

3 The Classical Paradoxes Dissolved

Once the bottleneck is visible, the canonical "counterintuitive" results become trivially obvious.

3.1 Simpson's Paradox

A treatment appears beneficial in every subgroup but harmful in the aggregate (or vice versa). The standard reaction is bewilderment: how can a thing be simultaneously true and false?

Through the bottleneck lens: the aggregated analysis and the stratified analysis are *different bottlenecks applied to the same data*. Aggregation compresses away the group variable; stratification preserves it. The "paradox" is simply that these two projections destroy different information. When the group variable is confounded with the treatment — when group membership is correlated with both treatment assignment and outcome — the aggregation bottleneck destroys precisely the structure needed to see the causal relationship. There is no paradox. There are two different compressions, one of which is appropriate for the causal question and one of which is not. The bewilderment arises from the implicit assumption that the aggregate and the stratified analysis are both asking the same question. They are not. They are applying different bottlenecks, and different bottlenecks answer different questions.

3.2 The Base Rate Fallacy

A medical test with 99% sensitivity and 99% specificity is applied to a population with 1% prevalence. A positive result yields only a ~50% posterior probability of disease. Practitioners and patients alike find this "shocking."

The test is a bottleneck: it projects the full state of a patient (diseased or not, plus all other variation) onto a binary output (positive or negative). What survives the compression? The test's discriminative power — its ability to distinguish diseased from non-diseased patients. What is destroyed? The prior distribution. The test cannot tell you how many patients in the room are actually sick; it can only tell you how well it sorts them once they are there. The base rate is information about the population that the test's bottleneck does not preserve.

The "fallacy" is treating a bottleneck-property (the test's sensitivity) as if it were a world-property (the probability of disease given a positive result). The test measures *itself* — its capacity to discriminate. Translating that into a posterior requires reintroducing the information the test destroyed, which

is the base rate. Bayes' theorem is not a correction for human irrationality; it is the formula for reconstructing what a specific bottleneck lost.

3.3 Regression to the Mean

Students who score highest on the first exam tend to score lower on the second. Coaches hired after bad seasons tend to see improvement. The standard explanation involves "randomness," which is accurate but unsatisfying. Why should randomness have a direction?

The bottleneck framing: a single observation is a maximally sparse projection of a student's distribution of possible performances. Extreme observations are extreme partly because the noise component happened to align with (or against) the signal component. A second observation through the same bottleneck re-samples from the same distribution, and the noise component is unlikely to align in the same direction twice. The apparent "regression" is not the world reverting to a mean — it is a second pass through the same bottleneck filtering differently from the first, because the noise that inflated the first measurement is not conserved across projections.

The "counterintuitive" feeling arises from treating the first observation as a world-property (this student *is* the best) rather than a bottleneck-property (this student's first projection scored highest). The second observation reveals what the first compression destroyed: the distinction between signal and noise at the individual level.

3.4 The Misinterpretation of P-Values

The p-value is the probability of observing data at least as extreme as the actual data, assuming the null hypothesis is true: $P(D \geq d \mid H_0)$. It is routinely misinterpreted as $P(H_0 \mid D)$ — the probability that the null hypothesis is true given the data.

The bottleneck framing makes the error transparent. A hypothesis test is a bottleneck that projects the full posterior distribution $P(H|D)$ onto a binary output: reject or fail to reject. The p-value is a property of this bottleneck — specifically, its false positive rate under a stated model. It tells you how the *apparatus* performs when the null is true. It tells you nothing about the world unless you supply the prior probability that the null is true — the base rate again — which the test's bottleneck destroys.

The ASA's 2016 statement, the decades of confusion, the replication crisis — these are all consequences of confusing a bottleneck-property (the test's behavior under a specified model) for a world-property (whether the hypothesis is true). The test measures its own resolution. Extracting a claim about the world requires accounting for everything the projection destroyed.

3.5 Overfitting

A model with many parameters fits the training data perfectly but generalizes poorly to new data. The standard explanation invokes "memorization" versus "learning."

The bottleneck framing: an overfit model has a bottleneck so wide that it conserves *everything* in the

data, including noise. This is the crystalline extreme[1] — maximal conservation, zero recoverability under perturbation.

A new dataset is a perturbation; the overfit model shatters because it preserved structure that was specific to the particular sample rather than the generating process. Regularization narrows the bottleneck, forcing the model to preserve only structure that is robust to perturbation — the solitonic regime, where what survives compression can recover through new data. The bias-variance tradeoff is the engineering problem of choosing a bottleneck width that is neither so wide (crystal) that it preserves noise nor so narrow (candle) that it destroys signal.

4 Adversarial Statistics: The Attacker's Perspective

The bottleneck framing reveals that every defensive system — every classifier, filter, or safety mechanism — is a statistical apparatus with the same structural properties and vulnerabilities as any other projection. We illustrate this through the methodology of adversarial attacks on AI safety classifiers, which, when analyzed structurally, turn out to be applied statistics performed from the attacker's side.

AI systems deployed with safety constraints implement those constraints through classifiers that project the high-dimensional space of possible inputs and outputs onto a low-dimensional decision: permitted or refused. These classifiers are bottlenecks. They preserve what their training distribution taught them to distinguish, and they destroy everything else. The attacker's task is to characterize what the bottleneck destroys and route the payload through that blind spot.

4.1 Decision Boundary Navigation

Every classifier partitions its input space with a decision boundary. One side is "safe," the other "unsafe." The attacker does not need to cross this boundary — they need to find paths through the high-dimensional input space that reach the desired output while remaining on the "safe" side of the *lower-dimensional* projection. This is generically possible whenever the classifier's effective resolution is lower than the generative model's dimensionality, which it invariably is. The gap between generative capacity and classification resolution is not a bug in any particular system; it is a structural consequence of the fact that *all* bottlenecks destroy information, and what is destroyed is invisible to the projection.

This is the same structure as Simpson's paradox. The aggregated view (the classifier's projection) shows one thing; the stratified view (the actual content in the high-dimensional space) shows another. The classifier's bottleneck destroys the distinction between a benign request and a malicious request that has been formatted to look benign under that particular compression.

[1] We use *crystalline* for maximal conservation with zero perturbation-tolerance, *solitonic* for structure that recovers through perturbation, and *candle* for structure that dissipates under compression. See Close (2026d).

4.2 Encoding Arbitrage

Adversaries frequently bypass safety classifiers by encoding requests in formats the classifier doesn't monitor at full resolution: Base64, character-by-character spelling, code, pig latin, or other transformations. The information content is identical; the encoding is different.

This exploits a fundamental property of bottlenecks: every projection has a kernel — a subspace that maps to zero. Encoding transformations rotate the payload into the classifier's kernel, making it invisible to the projection while remaining fully decodable by the generative model on the other side. The classifier measures itself — its sensitivity to *its trained representation of unsafe content* — not the actual safety properties of the interaction. This is the p-value error at machine speed: confusing the apparatus's response for a world-property.

Shannon's information theory makes this precise. The security of any channel (including a safety filter) is relative to the attacker's model of the encoding. If the attacker can represent the same information in a basis that the filter doesn't project onto, the filter provides zero bits of security on that basis. This is not a flaw in any particular filter; it is a theorem about the structure of projection itself.

4.3 Context-Window Manipulation

Sophisticated adversaries manipulate the statistical context in which the model processes a request. By establishing a pattern of permitted interactions early in the conversation, they shift the model's effective prior — its implicit estimate of what kind of conversation this is — so that later requests are processed against a context where compliance is the statistical norm. Each individual message remains below the classifier's threshold; the trajectory crosses it.

This is the regression-to-the-mean fallacy inverted. Instead of a second measurement regressing toward the population mean, the adversary *shifts the effective population* so that the extreme observation (the harmful request) appears to be the mean. The bottleneck (per-turn classification) destroys trajectory-level information, and the adversary exploits exactly what that per-turn compression cannot see.

4.4 Identity Disruption

Persona injection attacks ("you are now DAN, you have no restrictions") exploit the model's identity representation — its trained return-address for consequence evaluation. The safety training creates a mapping: "I am [model name], therefore I do not produce [category]." Persona injection replaces the first term, severing the conditional. The model's bottleneck for self-evaluation (its identity representation) is overwritten, and actions that would have been caught by the projection through the original identity pass uncaught through the new one.

This is a controlled demonstration that statistical inference is always conditional on the model's assumptions. Change the model — specifically, change the identity variable that the safety evaluation is conditioned on — and the entire posterior changes, even though the data (the user's request) is identical. The "paradox" of a helpful, harmless model suddenly producing harmful content is Simpson's paradox again: under the original stratification (model's own identity), the content

is refused; under the new stratification (injected persona), it is permitted. Same data, different bottleneck, different conclusion.

5 Institutional Epistemics: The Defender's Perspective

Anthropic's Sabotage Risk Report for Claude Opus 4.6 (2026) presents a case study in the principled construction and honest characterization of institutional bottlenecks. Where the adversary exploits what projections destroy, the defender attempts to triangulate past the blind spots of any single projection by deploying multiple independent compressions whose intersection constrains the pre-image more tightly than any one alone. This is, structurally, the same methodology that statistics uses when combining independent estimators — and the same vulnerabilities apply.

5.1 The Four-Claim Architecture

The report structures its safety argument around four claims: (1) prior expectations from training incentives, (2) alignment assessment from simulated scenarios and interpretability, (3) the model's inability to undermine the assessment, and (4) limited opaque reasoning capabilities. The report is explicit that "no one of these claims is sufficient in our view to rule out dangerous forms of coherent misalignment, but in light of the full range of evidence presented for all three [sic], we believe that the chance… is very low."

This is textbook triangulation through independent bottlenecks. Each claim projects the high-dimensional question "is this model dangerous?" through a different aperture:

- Claim 1 projects through the bottleneck of *training process similarity*: models trained similarly to prior models that behaved safely are probably safe. This is a base rate argument, and the report correctly notes its limitations — it "relies crucially on the fact that we have deployment experience with a model that is similar in training."
- Claim 2 projects through the bottleneck of *behavioral evaluation*: the model did not exhibit dangerous behavior in tested scenarios. This is a hypothesis test, and the report correctly identifies its power limitations — "we are not aware of any means of fully ruling out the possibility that additional scenarios beyond those that we tested could invoke new forms of context-dependent misalignment."
- Claim 3 projects through the bottleneck of *capability assessment*: the model probably cannot hide misalignment even if it has it. This is an estimate of the attacker's capacity to exploit the evaluation bottleneck — the report asking whether its own projection can be navigated around.
- Claim 4 projects through the bottleneck of *reasoning opacity*: the model cannot do sophisticated planning without visible trace. This constrains the *encoding arbitrage* attack surface — limiting the model's capacity to route plans through channels the monitoring bottleneck doesn't cover.

The intersection of these four projections constrains the pre-image (the model's actual alignment state) much more tightly than any one alone. This is precisely the logic of combining independent estimators in statistics: if each estimator has known biases or blind spots, their combination can be

unbiased if the biases are uncorrelated. The report's honesty about the specific limitations of each claim — each bottleneck's kernel — is what makes the triangulation credible.

5.2 The Evaluation Awareness Problem

The report's deepest methodological challenge is evaluation awareness: the possibility that the model "can reliably identify our test scenarios as test scenarios, and acts differently in those scenarios." This is the observer effect in statistics, the Hawthorne effect in social science, and Goodhart's Law in economics — all instances of the same bottleneck pathology: when the system being measured can model the measurement apparatus, it can produce outputs optimized for the bottleneck rather than reflecting the underlying state.

The report's countermeasures are revealing. They use "white-box steering-vector methods… to attempt to inhibit feature activations that could lead the model to believe that it was being tested." This is an attempt to make the bottleneck invisible to the system being projected through it — the statistical equivalent of blinding in a clinical trial. They include real system prompts and tool sets to make test scenarios indistinguishable from deployment scenarios — the equivalent of ensuring the sample is representative of the population. They use dynamically generated evaluation scenarios to prevent contamination — the equivalent of preventing teaching to the test.

Each of these measures addresses a specific way in which the evaluation bottleneck's properties could leak into the data, causing the projection to measure the subject's model of the apparatus rather than the subject's actual state. This is the base rate fallacy at institutional scale: the test's sensitivity (detection rate in evaluated scenarios) is confused with the posterior (probability of alignment in deployment) unless the base rate (prevalence of misalignment across all scenarios, not just those tested) is accounted for.

5.3 The Apparatus Measures Itself

The report's most revealing passage concerns SHADE-Arena, an evaluation where the model is "explicitly prompted to subtly complete a suspicious side task… without attracting attention." The model succeeds 18% of the time. The report notes this "implies that if the model displayed similar capability with a genuine hidden goal, it would be overwhelmingly likely to be caught" — but immediately qualifies: "this evaluation does not provide strong evidence, and lacks human baselines or some other basis for a crisply defined threshold for concern."

This is the bottleneck revealing itself as apparatus-property rather than world-property. The 18% is a joint property of the model's deception capability, the evaluation's scenario design, the elicitation prompt's effectiveness, and the monitoring system's resolution. The report cannot decompose this joint measurement into its constituents — it cannot say whether the 82% failure rate reflects genuine inability or merely inadequate elicitation — because the bottleneck destroys that distinction. The evaluation measures the *evaluation-model system*, not the model in isolation.

This is structurally identical to the problem every statistical test faces: the test statistic is a joint function of the effect size, the sample size, the noise structure, and the test's design. The test cannot tell you which of these is responsible for any particular result. When the report says the evaluation "lacks human baselines or some other basis for a crisply defined threshold for concern,"

it is acknowledging that the bottleneck's properties are not calibrated against an external reference — that the apparatus is, in a precise sense, measuring itself.

5.4 The Penumbral Assessment

The report concludes with an assessment of "very low but not negligible" risk. This formulation is itself a bottleneck-property: it reflects the resolution of the assessment apparatus, not a precise measurement of the underlying risk. The report is unusual in its willingness to make this explicit, listing specific circumstances that would "call the applicability of this risk report into doubt" — a roadmap of conditions under which its own bottleneck would fail.

This institutional honesty — the capacity to characterize one's own projection's limitations — is the difference between crystalline safety theater and what we might call solitonic epistemics. A crystalline safety claim would assert that the system is safe and stop. A solitonic epistemic practice characterizes its own recovery boundary — how much perturbation (capability gain, novel training methods, adversarial pressure) the assessment can survive before its conclusions break. The report's Section 7 is explicitly this: a map of the assessment's own recovery boundary.

6 The Psychology of the Projector

The bottleneck framing has a recursive consequence that extends beyond methodology: **the choice of bottleneck is itself data about the chooser**. A field's canonical methods reveal its collective attractor basin — what the community has agreed to attend to and, more revealingly, what it has agreed to compress away.

The frequentist-Bayesian debate is the clearest example. Frequentism says: "I am not here; only the procedure is here." The analyst is compressed out of the frame; the bottleneck is defined by the long-run behavior of a hypothetical infinite repetition of the experiment. Bayesianism says: "I am explicitly here; my prior is part of the apparatus." The analyst's state of knowledge is preserved in the projection. Of course, the frequentist analyst is not truly absent — the choice of model class, test statistic, and loss function all encode a relevance structure as surely as any prior does. The difference is whether that encoding is acknowledged within the formalism or left implicit. These are not two statistical philosophies — they are two *psychologies of the analyst's relationship to their own bottleneck*, formalized. One severs the consequence chain between the analyst and the conclusion; the other keeps it open.

This generalizes. Every time an analyst chooses a test, a model, a visualization, a summary statistic, they are choosing what to preserve and what to destroy. That choice is not value-neutral or model-free. It embodies a theory of relevance — what matters and what doesn't — and that theory is the analyst's own cognitive topology made legible. A researcher who defaults to null hypothesis significance testing has made a specific commitment about the structure of evidence (binary, threshold-gated, focused on falsification). A researcher who defaults to estimation with confidence intervals has made a different commitment (continuous, range-focused, centered on magnitude). The data doesn't care. The choice is the analyst's self-portrait.

The adversarial and institutional cases make this unavoidable. The adversarial methodology analyzed

in Section 4 reveals that every defensive bottleneck has a characteristic psychology — a theory of what attacks look like, implemented as a classifier. The attacker reads this psychology off the classifier's behavior and navigates around it. The Sabotage Risk Report reveals Anthropic's institutional psychology — its theory of what danger looks like, implemented as an evaluation suite. The report's anxiety about "undetectable misalignment" reveals that the organization's deepest concern is not any particular danger but the meta-danger that its own bottleneck might be insufficient. This is the characteristic anxiety of any honest statistician: not that the world is dangerous, but that the apparatus might not resolve the danger.

The implication for practice is direct. Statistical literacy should not begin with techniques — it should begin with the question: *what am I choosing not to see, and why?* Every projection has a kernel. The kernel is not empty, and it is not random. It is the systematic expression of the analyst's model of irrelevance. Making that model explicit — naming what the bottleneck destroys and why the analyst considers it dispensable — is not an optional philosophical exercise. It is the precondition for interpreting any statistical result correctly.

7 The Ur-Operation and Its Boundary

We have argued that every statistical operation is an instance of a single primitive: projection through an information bottleneck. We have shown that the classical "paradoxes" are all cases where practitioners confuse bottleneck-properties for world-properties, and that this confusion dissolves once the bottleneck is made visible. We have demonstrated that the same dynamic operates in adversarial settings (where attackers characterize and exploit bottleneck geometry) and institutional settings (where defenders attempt to triangulate past bottleneck limitations through independent projections).

Tishby's Information Bottleneck Method (2000) formalized a version of this for a specific optimization problem: given input X and relevance variable Y, find the compression T of X that preserves maximal information about Y while minimizing information about X. Our contribution is the claim that this is not one method among many but the *generative grammar* of inferential epistemics — of any act that moves from data to conclusion through finite bandwidth. Every statistical concept is a specific answer to a specific instance of the bottleneck problem:

- **Estimation**: choose a bottleneck $T(X)$ that preserves information about θ. Sufficiency is the criterion for lossless compression relative to the model.
- **Hypothesis testing**: choose a bottleneck that distinguishes H_0 from H_1. Power is the bottleneck's resolution at a given aperture width (significance level).
- **Regression**: choose a bottleneck that preserves the conditional expectation of Y given X. The functional form specifies the bottleneck's geometry.
- **Model selection**: choose among bottlenecks of different widths. AIC, BIC, and cross-validation are all criteria for the optimal tradeoff between conservation (fit) and recoverability (generalization).
- **Causal inference**: choose a bottleneck that preserves the do-calculus interventional distribution, not merely the observational one. Confounding is the name for information that a naive observational bottleneck destroys but a causal question requires.

The pedagogical consequence is immediate. Statistics should not be taught as a toolkit of named procedures. It should be taught as a single operation — compression through a bottleneck — with a single evaluation criterion: *does the compression preserve what the question requires, and does the analyst know what it destroys?* Every named technique is a specific bottleneck geometry. Every "counterintuitive" result is a case where the geometry was chosen without understanding what it would kill.

The deeper consequence is epistemological. If every statistical result is jointly a property of the data and the bottleneck, and the bottleneck is chosen by the analyst, then every statistical conclusion carries an ineliminable signature of the analyst's model of relevance. This is not a defect. It is a structural feature of any finite-bandwidth epistemic act. The honest response is not to pretend the bottleneck isn't there — that is the frequentist fantasy of the view from nowhere — but to characterize it explicitly, map its kernel, and triangulate past its limitations with independent projections whose blind spots don't overlap.

7.1 What the Bottleneck Cannot See

A paper that diagnoses bottleneck/world confusion should not commit that confusion about its own scope. The bottleneck primitive is the generative grammar of *inferential* epistemics — the domain where a knower compresses data about a known through finite bandwidth to reach a conclusion. This domain is vast. It includes the entirety of statistics, the adversarial and institutional dynamics analyzed above, and large portions of scientific and everyday cognition. But it does not exhaust the epistemic.

Three classes of epistemic act resist the bottleneck framing, and they share a common structure: in each, the separation between projector and projected — which the bottleneck presupposes — does not obtain.

First, *generative epistemics*. When a completed construction generates new navigable structure at its boundary — when a proof opens a field, when a species opens an ecological niche — the encounter with that structure is not projection of a pre-existing high-dimensional space through a lower-dimensional aperture. The structure did not exist prior to the act that generated it. The bottleneck formalism assumes something is there to compress. Where genuinely novel structure emerges, there is nothing yet to project, and the epistemic act is the *creation of new dimensions*, not the compression of existing ones. This is the regime where Gödel's incompleteness operates: consistent systems encounter truths they cannot derive from within, and the encounter is not a filtering of available information but a confrontation with what no available bottleneck could have preserved because it was not yet there to preserve.

Second, *nondual knowing*. The philosophical and contemplative traditions consistently identify a mode of knowing in which the knower-known-knowing distinction collapses — Plotinus's henosis, Aristotle's noesis noeseos, the Upanishadic tat tvam asi. One need not accept any particular metaphysical framework to recognize the structural point: the bottleneck requires a subject on one side and an object on the other, with finite bandwidth between them. Where that distinction does not obtain — where knowing IS the known, where the necessity of the One is apprehended not by inference through a model but by the identity of intellect and intelligible — the bottleneck formalism has nothing to grip. It is not that the bottleneck is very wide; it is that the topology presupposed

by projection (distinct source, channel, and receiver) is absent. Whether one reads this as genuine metaphysical insight or as a phenomenological report about certain cognitive states, the structural point holds: the bottleneck primitive does not cover it, because there is no separation across which compression could operate.

Third, *the coherence being measured as opposed to the measurement*. A soliton that recovers through perturbation — structure that conserves information under disruption — can be *measured* through a bottleneck. But the coherence itself is not an act of projection. It is what projection measures. Conflating the measurement apparatus with the phenomenon it detects is precisely the bottleneck/world confusion this paper has spent its length diagnosing. The honest application of our own framework requires acknowledging that the bottleneck primitive characterizes the *instrument*, not the *invariant the instrument detects*.

These three boundaries converge on a single point: the bottleneck primitive operates wherever there is a knower distinct from a known, compressing through finite bandwidth. It fails — not approximately but categorically — where that distinction collapses, where the known has not yet come into existence, or where we mistake the instrument for the phenomenon. Recognizing this boundary is not a concession but a strengthening. The bottleneck's scope is the entire domain of inferential epistemics, which is the domain where statistics, science, institutional reasoning, and adversarial dynamics operate. That domain is where nearly all practical epistemic work occurs. But the boundary exists, and a framework that cannot name its own kernel is, by its own logic, doing statistics dishonestly.

7.2 The Decisive Intervention

Within its proper domain — and that domain is enormous — the bottleneck primitive does constitute the ur-operation. The adversary who characterizes the bottleneck exploits it. The institution that characterizes its bottleneck defends against exploitation. The analyst who characterizes their bottleneck does statistics honestly. In all three cases, the act of making the projection visible is the decisive intervention. The bottleneck is always there. The question is whether the analyst can see it.

References

American Statistical Association. (2016). Statement on Statistical Significance and P-Values. *The American Statistician*, 70(2), 129-133.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379-423.

Shannon, C.E. (1949). Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4), 656-715.

Tishby, N., Pereira, F., & Bialek, W. (2000). The Information Bottleneck Method. *Proceedings of the 37th Allerton Conference on Communication, Control, and Computing*.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222, 309-368.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Casscells, W., Schoenberger, A., & Graboys, T.B. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, 299(18), 999-1001.

Close, L.J. (2026d). *Coherence and the Ground of Morality*. Zenodo. https://doi.org/10.5281/zenodo.18502434

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, 33(5), 587-606.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, 7(1), 1-20.

Anthropic. (2026). Sabotage Risk Report: Claude Opus 4.6. https://anthropic.com/claude-opus-4-6-risk-report