Confidence Is Not Certification

Calibration, Evidentiary Sufficiency, and the Authorization of Autonomous Systems

Flyxion
Independent Researcher

November 2026

Abstract

Numerical confidence estimates in autonomous computational systems reflect a broader tendency to impose probabilistic representations, scalar measurements, and total orderings upon domains whose epistemic structure may not support them. This study distinguishes the production of confidence statements from statistical calibration, deductive validity, evidentiary sufficiency, and the authorization of consequential actions. Examples concerning institutional facts, meal photographs, artificially coloured foliage, biochemical classification, and preference-dependent planning show how an unidentifiable quantity, an unjustified comparison relation, or an omitted causal explanation can survive a formally impeccable estimate. Two results organize the constructive argument. First, when correctness labels are obtained only through verification, and no further restrictions apply, calibration is not identified on any confidence region that the routing policy never verifies, so that a threshold policy can remove the evidence needed to evaluate it; randomized audits with positive verification probability restore identification under stated assumptions. Second, confidence-based routing, which selects a verification procedure, is separated from obligation-based commitment, which determines whether an action may proceed. In the resulting architecture measurement, inference, verification, adjudication, and authorization remain distinct operations, and observability is preserved as an operational obligation.

Keywords: statistical calibration, identifiability, measurement theory, epistemic uncertainty, selective verification, deductive validity, autonomous systems, evidentiary sufficiency, authorization, probabilistic representation

Contents

1 Introduction: The Numerical Performance of Certainty **6**

1.1 Four failures of numerical representation 7

1.2 The grammatical availability of an answer 8

2 The Presumption of Measurability **8**

2.1 Representations and their admissible operations 9

2.2 Precision without evidentiary resolution 10

2.3 Methodological competence and domain boundaries 11

3 Deductive Validity and Established Facts **12**

3.1 Distributed corroboration 12

3.2 Counterfactual consequences of factual error 13

3.3 Deductive necessity and empirical fallibility 13

3.4 The distinction between truth and retrieval reliability 15

3.5 The reference-class problem 15

3.6 The sufficiency of nonnumerical knowledge 16

4 Artificial Orderings and Heuristic Commensuration **17**

4.1 Partial orderings and heterogeneous properties 17

4.2 Scalarization and preference dependence 18

4.3 Itinerary planning as heuristic completion 20

4.4 The distinction between feasibility and optimality 21

4.5 Scientific and institutional consequences 22

4.6 Heuristic networks and restricted inference 22

5 Observability and the Non-Identifiability of Physical Quantities **23**

5.1 Observation maps and equivalence classes 23

5.2 A lower bound on unavoidable estimation error 25

5.3 Nutritional estimation from photographs 26

5.4 Conditional expectation and individual composition 27

5.5 Refinement through additional observation 28

5.6 Set-valued estimates and evidentiary bounds 29

5.7 Testimony as an additional evidentiary channel 30

5.8 Statistical identifiability under noisy observations 30

5.9 The operational significance of approximation 31

6 Causal Ambiguity and Observational Equivalence 31

6.1 The spray-painted tree 32
6.2 Mechanism-dependent inference 33
6.3 A mechanism-ambiguity obstruction 33
6.4 Intervention and the inadequacy of correlation 34
6.5 Distributional change and transportability 35
6.6 Biochemical classification and developmental heterogeneity 36
6.7 Prediction, diagnosis, and causal explanation 36
6.8 Reference standards and construct validity 37
6.9 Base rates and predictive interpretation 37
6.10 Selective ignorance and the excluded mechanism 38

7 Probabilistic Closure and Excluded Possibilities 39

7.1 Internal coherence and external adequacy 39
7.2 Posterior concentration under misspecification 40
7.3 Relative support and absolute adequacy 41
7.4 Excluded hypotheses and zero prior probability 42
7.5 The distinction between uncertainty and ignorance 43
7.6 Sensitivity to the admissible model class 43
7.7 Heuristic closure and practical adequacy 44
7.8 The absence of a probability is not a probability 44

8 Calibration Without Certification 45

8.1 The definition of calibration 45
8.2 Calibration and discrimination 46
8.3 Calibration and the evidentiary record 48
8.4 Verbalized confidence and token generation 49
8.5 Internal representations and information about correctness 49
8.6 Empirical findings on language-model confidence 50
8.7 Token probabilities and answer probabilities 51
8.8 Stochastic variation and reproducibility 51
8.9 Prompt sensitivity and epistemic invariance 53
8.10 Confidence conditioned upon the produced answer 54

8.11 Tool-computed estimates and numerical provenance 54
8.12 Marginal calibration and case-specific evidence 55
8.13 From calibration to certification 56

9 The Externalization of Verification 56
9.1 Reconstruction of the proposed procedure 56
9.2 Observable indicators and latent correctness 57
9.3 Document retrieval and semantic support 58
9.4 Agreement and common-mode error 59
9.5 An explicit agreement model 59
9.6 The observational equivalence of agreement records 60
9.7 Outcome labels and the substitution of proxies 61
9.8 Calibration against imperfect labels 62
9.9 Non-identifiability under unknown label error 63
9.10 Selective verification 64
9.11 A decomposition of selective verification bias 64
9.12 Deterministic thresholds and the absence of overlap 65
9.13 Partial identification under incomplete verification 66
9.14 Randomized audits and the positivity requirement 67
9.15 A practical mixed verification policy 69
9.16 The cost of rare-event detection 69
9.17 Small groups and estimation uncertainty 70
9.18 Smoothing and the role of prior assumptions 71
9.19 Changes in the help centre and the generative system 72
9.20 The law of large numbers and historical relevance 73
9.21 Versioned evidentiary records 73
9.22 Endogenous evidence and the feedback problem 74
9.23 An illustrative self-confirming routing failure 75
9.24 Evaluation independence and institutional incentives 75
9.25 Human review as a fallible evidentiary operation 76
9.26 Original answers and repaired answers 77
9.27 Requirements for an external confidence architecture 77
9.28 The question that follows calibration 78

10 Authorization Without Scalar Reduction **78**

10.1 Correctness and permission 79
10.2 The multidimensional status of a proposed action 79
10.3 Obligations and their satisfaction 80
10.4 The noncompensatory character of mandatory conditions 81
10.5 Confidence as a routing signal 83
10.6 The decision-theoretic role of probability 84
10.7 Admissibility under observational ambiguity 85
10.8 Refinement as an alternative to premature action 86
10.9 Predictive sufficiency and operational sufficiency 87
10.10 Local verification and global compatibility 88
10.11 Temporal validity and stale evidence 89
10.12 Reversibility and the consequences of error 90
10.13 Evidence-bearing authorization 91
10.14 The distinction between certificate existence and certificate validity . . 91
10.15 A worked customer-support example 92
10.16 Audit and certification as complementary operations 93
10.17 Refusal as an informative computational result 94
10.18 The limits of obligation-based architectures 94

11 Conclusion: The Right Not to Quantify **95**

11.1 What the distinctions preserve 95
11.2 Observability and authority 96
11.3 The right not to quantify 96

1 Introduction: The Numerical Performance of Certainty

A numerical expression can acquire an appearance of scientific legitimacy independently of the procedure that produced it. The assertion that a proposition is true with a confidence of 92% appears, at least superficially, to convey more information than the unqualified assertion of its truth. The numerical qualification suggests that some uncertainty has been identified, represented, and subjected to calculation. Yet the mere presence of a percentage establishes none of these operations. It establishes only that a number has been associated with a proposition. This distinction becomes consequential when numerical confidence estimates are incorporated into the control architecture of autonomous computational systems. A language model may be asked to generate an answer, evaluate the reliability of that answer, and supply a numerical estimate that determines whether the answer will be communicated, subjected to further examination, or used as the basis for an external action. In such an architecture, a linguistic expression of confidence can acquire operational consequences that greatly exceed its demonstrated epistemic significance.

An episode of the *Data Science at Home* podcast, entitled "Your LLM's Confidence Score Is Made Up," provides a useful occasion for examining this problem (Data Science at Home, 2026). The episode begins with an intentionally arbitrary confidence assignment to the proposition that Canberra is the capital of Australia. The speaker supplies a percentage of 92% and subsequently explains that the number was chosen for its appearance of scientific precision rather than obtained through an estimation procedure. The principal engineering criticism developed in the episode is sound. A numerical confidence statement generated by a language model does not, merely by virtue of having been generated, constitute evidence that the model is calibrated. A system designer who treats such a statement as an empirically established probability of correctness introduces an assumption that requires independent justification.

The episode further distinguishes calibration from informational discrimination. A weather forecaster who assigns a probability of rainfall equal to the historical base rate on every occasion may be calibrated while conveying no information that distinguishes one day from another. The episode also emphasizes the necessity of maintaining a record of forecasts and observed outcomes, without which claims about empirical calibration cannot ordinarily be substantiated. These observations should be retained without qualification concerning their central methodological importance. The speaker also explicitly allows that a language model may communicate a numerical estimate obtained through a separate computational instrument. The criticism is therefore not directed against the linguistic communication of numerical results as such, but against the substitution of an

unvalidated linguistic output for a measurement or estimation procedure. The present study extends this criticism in two directions.

The first extension concerns the conditions under which a confidence estimate is meaningful in the first place. Before determining whether a numerical estimate is calibrated, it is necessary to establish what is being estimated, whether the target is identifiable from the available observations, whether the relevant distinctions have been represented, and whether a probabilistic formulation is appropriate to the epistemic task. The second extension concerns the operational significance of a confidence estimate after its statistical legitimacy has been established. A well-calibrated probability of correctness does not necessarily establish that a particular answer is adequately supported, that an action is authorized, that its consequences are admissible, or that its execution satisfies the obligations imposed by the surrounding system. The central claim of this study is therefore not that numerical confidence estimates are inherently misleading, nor that probabilistic inference should be displaced by an exclusively deductive methodology. It is that confidence, measurement, inference, verification, and authorization are different operations whose conflation produces both scientific and engineering errors.

The mathematical instruments employed below are largely established ones, drawn from measurement theory, statistical identifiability, forecast verification, the analysis of missing data, causal inference, and constrained optimization. The factorization criterion, the dependence of weighted rankings upon their weights, and inverse-probability weighting are used as foundations and are not offered as discoveries. The contribution lies in their integration into a single account of autonomous continuation, in which representation, estimation, verification, certification, and commitment are separate operations, and in the argument that observability must itself be preserved when a verification policy depends upon the estimates that it is meant to test.

1.1 Four failures of numerical representation

Several distinct methodological failures are commonly subsumed under the general criticism of overconfidence. The first is the introduction of an unjustified numerical uncertainty into an otherwise well-established proposition. A person may know an ordinary geographical, historical, or institutional fact without possessing a defensible procedure for assigning a precise probability to its truth. The absence of such a procedure does not imply the absence of knowledge. The second is the imposition of a scalar ordering upon alternatives that possess heterogeneous and potentially incommensurable properties. A ranking can be useful for organizing activity without corresponding to an objective hierarchy among the ranked objects. The third is the estimation of a target quantity that is not identifiable from the available observations. In this case, the difficulty is not

simply uncertainty concerning a quantity that could otherwise be recovered. Rather, distinct admissible states of the world may produce the same observations while differing in the quantity of interest.

The fourth is the substitution of statistical reliability for evidentiary sufficiency or operational authorization. A procedure may possess a demonstrated record of success without every particular output satisfying the conditions required for its acceptance or execution. These failures are distinct, in that a system may avoid any one of them while exhibiting another. Thus an estimator may be calibrated over an appropriately defined population while remaining incapable of identifying the physical state responsible for a particular observation. A ranking may be computationally reproducible while depending upon unjustified preference weights. An answer may be factually correct while providing no authority to perform the action that it describes. A general theory of reliable autonomous continuation must preserve these distinctions.

1.2 The grammatical availability of an answer

Language models make the problem especially visible because they can readily satisfy the grammatical requirements of a request whose epistemic requirements have not been established. A request to rank several objects can be answered with a numbered list. A request to estimate confidence can be answered with a percentage. A request to identify a biological characteristic can be answered with a diagnostic category. A request to estimate the nutritional composition of a meal can be answered with a table of quantities. In each case, the existence of a syntactically appropriate response is independent of the existence of an adequate evidentiary procedure. The distinction is not peculiar to language models, since human respondents, institutional committees, statistical consultants, and scientific instruments operating outside their validated domains can exhibit analogous failures. Language models nevertheless reduce the practical cost of producing elaborate numerical representations, thereby making it easier to overlook the methodological work that would be required to justify them.

The danger is not simply that a system produces incorrect answers. It is that the availability of an answer in the requested format can obscure the absence of a defensible interpretation of that format.

2 The Presumption of Measurability

Scientific measurement begins, not with the assignment of a number, but with the establishment of a relationship between a specified property and an observational or experimental procedure. The development of measurement theory has repeatedly em-

phasized that the admissibility of numerical operations depends upon the structure of the empirical relations being represented (Stevens, 1946; Krantz et al., 1971). Numerical representation is not a universal license to perform arbitrary arithmetic upon the represented objects. An ordinal scale permits comparisons that need not support meaningful subtraction. A nominal classification distinguishes categories without necessarily ordering them. An interval scale supports comparisons of differences under specified transformations, while a ratio scale admits further operations associated with the preservation of a meaningful zero. These familiar distinctions express a more general principle: the mathematical structure of a representation must be justified by the structure of the domain being represented.

Definition 2.1 (Presumption of measurability)**.** The presumption of measurability is the methodological assumption that a property, distinction, or relation admits a meaningful numerical representation merely because a numerical representation can be requested, constructed, or computationally manipulated.

The presumption is not equivalent to the claim that a particular measurement is inaccurate. An inaccurate measurement may be improved through instrument calibration, repeated observation, or the correction of systematic error. The presumption of measurability concerns a logically prior question: whether the proposed measurement procedure has been shown to represent the intended property at all. A numerical estimate of the length of an object obtained with an imperfect ruler is a different epistemic object from a numerical estimate of the intellectual significance of an essay obtained through an unspecified scoring procedure. The former concerns an established physical quantity and an instrument whose relation to that quantity can be examined. The latter may concern a heterogeneous collection of evaluative considerations for which neither a natural unit nor an agreed comparison relation has been established.

This does not imply that intellectual significance cannot be investigated empirically. It implies that the construction of an indicator must not be confused with the discovery of a quantity that exists independently of the indicator.

2.1 Representations and their admissible operations

Let $\mathcal{X}$ denote a domain of objects and let $\mathcal{R}$ denote a collection of empirically meaningful relations among those objects. A numerical representation is a mapping

$$m : \mathcal{X} \longrightarrow \mathbb{R}$$

intended to preserve some specified portion of $\mathcal{R}$. For example, if $\preceq$ is a comparison relation, an order-preserving representation would satisfy

$$x \preceq y \quad \Longrightarrow \quad m(x) \leq m(y).$$

This condition expresses only weak preservation, since every constant function satisfies it. A faithful representation requires the biconditional

$$x \preceq y \quad \Longleftrightarrow \quad m(x) \leq m(y).$$

Because the order on $\mathbb{R}$ is total and transitive, a real-valued faithful representation can exist only when $\preceq$ is itself a total preorder, so that any incomparable pair of alternatives already excludes it. Totality and transitivity are necessary but, on uncountable domains, not sufficient. Moreover, the existence of the mapping m is not sufficient to establish that the intended empirical relation exists, that it is transitive, or that it admits a complete numerical representation. One may always assign arbitrary real numbers to the members of a finite collection. Such an assignment does not thereby constitute a measurement of a property of those members.

Definition 2.2 (Representational obligation)**.** A representational obligation is a condition that must be established before an operation performed upon a formal representation can be interpreted as an operation upon the represented domain.

The obligation associated with a numerical ordering is to establish the relevant comparison relation. The obligation associated with a physical estimate is to establish the relationship between observation and target quantity. The obligation associated with a probability is to establish the event or hypothesis to which the probability refers and the interpretation under which the probability is meaningful. These obligations are not discharged by the successful execution of the corresponding arithmetic.

2.2 Precision without evidentiary resolution

The number of digits used to express an estimate is a property of its representation, not necessarily a property of the evidentiary procedure that produced it. Suppose an estimation procedure returns a value

$$\widehat{\theta} = 0.928417.$$

The representation contains six decimal places. Nothing follows from this fact alone concerning the resolution of the instrument, the sampling variability of the estimator, the adequacy of its model, or the identifiability of the underlying parameter. Indeed, the

same value could have been produced by a validated numerical calculation, an arbitrary constant, a random number generator, or a language model completing a sentence. The problem is therefore not adequately described as excessive numerical precision. Reducing the expression to 0.93 would not repair the absence of an evidentiary procedure. Nor would replacing it with the qualitative expression "high confidence" necessarily resolve the underlying problem.

The relevant distinction concerns the provenance and interpretation of the estimate.

Definition 2.3 (Evidentiary resolution)**.** The evidentiary resolution of an observational procedure is the extent to which its available observations distinguish among admissible states that differ with respect to a specified target property.

Evidentiary resolution is target-dependent. A photograph may possess sufficient resolution to identify the colour of an object while containing insufficient information to establish its chemical composition. A transcript may establish that a person uttered a particular sentence while failing to establish the truth of the proposition expressed by that sentence. The distinction between representational precision and evidentiary resolution will recur throughout the following analysis.

2.3 Methodological competence and domain boundaries

The construction of meaningful measurements ordinarily requires knowledge extending beyond the ability to perform statistical calculations. Experimental design, physical instrumentation, psychological methodology, causal inference, and the substantive knowledge of the relevant scientific domain contribute different constraints upon what can legitimately be inferred. A statistical association between two recorded variables does not establish that either variable adequately represents the property named by its label. A regression coefficient cannot repair a defective operational definition. A classification algorithm cannot determine whether its target categories correspond to the distinctions that a scientific investigation ought to preserve. Methodological competence includes the ability to recognize when a question has been formulated in a manner that presupposes an unavailable observation, an unjustified ordering, or a nonexistent unit of measurement. The appropriate response to such a question may involve reformulating it, identifying the additional observations required to answer it, or declining to produce the requested numerical representation.

This capacity for methodological refusal is not opposed to scientific inquiry. It is one of the conditions under which scientific inquiry remains distinguishable from the unrestricted production of numerical assertions.

3 Deductive Validity and Established Facts

The Canberra example introduced in the podcast illustrates a difficulty that precedes the question of language-model calibration. The proposition that Canberra is the capital of Australia is an established institutional fact. It is neither a forecast concerning a future meteorological event nor an estimate of a latent physical quantity from incomplete sensor observations. The proposition has a publicly maintained truth condition associated with the political and administrative organization of Australia. Its ordinary verification involves consulting appropriate institutional records, geographical references, or other sufficiently reliable sources. The introduction of a confidence estimate of 92% therefore raises a question that is not answered by the percentage itself. What event, population, or uncertainty model is supposed to justify the remaining 8%? The number might refer to the historical accuracy of a particular respondent on geographical questions. It might refer to the probability that an information retrieval system has returned the correct record. It might express a subjective degree of belief concerning the proposition. It might represent uncertainty about whether the institutional fact has recently changed.

These interpretations are not equivalent, and in particular the probability that a respondent will correctly answer a randomly selected question from a specified collection is not automatically the probability that an already identified answer to a particular question is correct. A meaningful numerical estimate requires an interpretation that distinguishes these possibilities.

3.1 Distributed corroboration

Ordinary factual knowledge is frequently supported by a distributed network of records, practices, observations, and institutional dependencies. The existence of a city, its geographical location, and its designation as a national capital are reflected in maps, administrative records, legislative documents, educational publications, transportation systems, diplomatic activities, and the ordinary conduct of individuals and organizations. These sources are not necessarily statistically independent, since many reproduce information obtained from common authorities, and their evidentiary value therefore cannot be calculated by simply multiplying the apparent probabilities of their individual correctness as though they were independent. Nevertheless, the distributed structure of their corroboration is significant. A counterfactual change in the relevant institutional fact would have consequences extending beyond the alteration of a single sentence in a reference work. Let p denote an established proposition and let

$$\mathcal{D}(p) = \{d_1, d_2, \ldots, d_n\}$$

denote a collection of records and practices that depend upon, express, or corroborate p.

The evidentiary significance of $\mathcal{D}(p)$ depends upon its provenance, internal dependencies, institutional authority, and susceptibility to common-mode error. It does not follow that each member of $\mathcal{D}(p)$ supplies an independent likelihood factor. Nor does it follow that the existence of a sufficiently extensive corroborating structure makes the proposition logically infallible. What follows is that the justification of ordinary factual knowledge can possess a rich evidentiary structure without being reducible to a uniquely determined numerical confidence estimate.

3.2 Counterfactual consequences of factual error

The distributed corroboration of established facts can also be examined through the consequences that would be expected if those facts were false. Suppose a geographical reference identifies a city that does not exist. The error may initially be difficult to detect if the reference is examined in isolation. However, as the purported city becomes incorporated into transportation planning, postal services, administrative records, commercial activities, and public communication, the absence of the corresponding physical and institutional structures would create opportunities for detecting the discrepancy. This observation suggests a distinction between the repetition of a proposition and its integration into independently constrained practices. A proposition repeated in a thousand derivative documents may possess less evidentiary support than a proposition whose consequences are independently manifested in several heterogeneous systems. The latter form of corroboration is especially important because the relevant systems are not merely repeating the proposition. They are organized in ways that would become difficult to sustain if the proposition were false.

This does not eliminate the possibility of coordinated deception, institutional error, or the persistence of obsolete records. It does, however, provide a substantive account of why certain established facts can be treated as adequately supported for ordinary purposes without the invention of a numerical uncertainty.

3.3 Deductive necessity and empirical fallibility

Deductive inference introduces a different kind of epistemic relation. Let Γ denote a collection of premises and let φ denote a conclusion. The expression

$$\Gamma \models \varphi$$

states that every interpretation satisfying Γ also satisfies φ. Under the ordinary semantics of classical logic, there is no interpretation in which all the premises are true and the

conclusion is false. The entailment is not strengthened by assigning it a confidence of 99%, nor weakened by the absence of a numerical estimate of confidence. Its validity concerns the relationship between the premises and the conclusion.

Proposition 3.1 (Conditional necessity)**.** *Let Γ be a finite collection of propositions and suppose that $\Gamma \models \varphi$. For any probability measure defined over interpretations of the relevant language, if*

$$\mathbb{P}\left(\bigwedge_{\gamma\in\Gamma}\gamma\right) > 0,$$

then

$$\mathbb{P}\left(\varphi \,\middle|\, \bigwedge_{\gamma\in\Gamma}\gamma\right) = 1.$$

Proof. Let A denote the event that every proposition in Γ is true, and let B denote the event that φ is true. Since $\Gamma \models \varphi$, every interpretation in A belongs to B. Consequently,

$$A \subseteq B$$

and therefore

$$A \cap B = A.$$

Since $\mathbb{P}(A) > 0$, the definition of conditional probability gives

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A)} = 1.$$

□

The proposition does not establish that the premises are empirically true. It establishes that uncertainty concerning the premises must not be confused with uncertainty concerning the validity of the entailment. Similarly, uncertainty about whether a particular proof has been correctly transcribed or checked does not imply that deductive validity is itself a matter of statistical frequency. A language model may produce an invalid derivation while claiming complete confidence in it. A human mathematician may produce a valid derivation while remaining uncertain about whether an unnoticed error has occurred. These are differences in the reliability and epistemic state of the agents performing the derivation, not differences in the semantics of logical consequence. Probability and deductive logic are therefore compatible mathematical systems which can be combined in a coherent formal framework while retaining distinguishable epistemic roles.

3.4 The distinction between truth and retrieval reliability

Consider a computational system asked to retrieve an established fact. Let p denote the relevant proposition, let A denote the answer produced by the system, and let Y indicate whether the answer satisfies an independently specified correctness criterion. The quantity

$$\mathbb{P}(Y = 1)$$

may describe the reliability of the retrieval procedure over a specified distribution of tasks. It need not describe uncertainty concerning the truth of p under an independently established interpretation. A geographical database might answer questions correctly in 98% of recorded cases. This does not imply that every city name stored in the database should be regarded as possessing an individual truth probability of 0.98. The aggregate reliability of the procedure is a property of its performance under specified conditions. The evidentiary support for a particular answer is a property of the records, observations, and verification procedures available in that case.

Confusing these quantities encourages the assignment of population-level frequencies to individual propositions without establishing the relevance of the population to the particular case.

3.5 The reference-class problem

Even when a numerical probability is appropriate, its interpretation depends upon the reference class or probabilistic model under consideration. The historical accuracy of a respondent on questions about national capitals may differ from the respondent's accuracy on questions about geography generally. Accuracy on well-established institutional facts may differ from accuracy on recently changed administrative arrangements. The same answer can therefore belong to multiple reference classes possessing different empirical success rates. Let $\mathcal{K}_1$ and $\mathcal{K}_2$ denote two such classes. It is entirely possible that

$$\mathbb{P}(Y = 1 \mid X \in \mathcal{K}_1) \neq \mathbb{P}(Y = 1 \mid X \in \mathcal{K}_2),$$

even when the particular task x belongs to their intersection,

$$x \in \mathcal{K}_1 \cap \mathcal{K}_2.$$

The choice of reference class is therefore a substantive component of the estimation procedure, not an incidental detail that disappears once the arithmetic has been performed. This observation qualifies the podcast's assertion that two forecasters possessing the same information must obtain the same probability unless one has made a computational

error.

Suppose two forecasters observe identical evidence D but employ different probabilistic models M_1 and M_2. Their posterior estimates may satisfy

$$\mathbb{P}(H \mid D, M_1) \neq \mathbb{P}(H \mid D, M_2).$$

Under Bayesian inference, the posterior within model M_i is determined by

$$\mathbb{P}(H \mid D, M_i) = \frac{\mathbb{P}(D \mid H, M_i)\mathbb{P}(H \mid M_i)}{\mathbb{P}(D \mid M_i)}.$$

Differences in prior distributions, likelihood functions, reference classes, or structural assumptions can produce different posterior estimates without either forecaster having performed an incorrect calculation. Agreement follows when the relevant probabilistic specifications are shared and the calculation is performed exactly. Approximate inference may introduce further variation even under a shared specification. The methodological requirement is therefore not that every competent observer must produce the same number from the same observations. It is that the assumptions, procedures, and evidentiary relationships responsible for the number remain available for examination.

3.6 The sufficiency of nonnumerical knowledge

The foregoing distinctions suggest that the absence of a numerical confidence estimate should not ordinarily be interpreted as an epistemic deficiency. An established proposition can be supported by authoritative records and extensive corroboration. A mathematical conclusion can be supported by a valid derivation. A computational result can be supported by a reproducible execution and independent checks. In none of these cases is an additional scalar confidence estimate necessarily required to make the supporting procedure legitimate. Conversely, where the relevant evidence is absent, a numerical expression of confidence does not supply it. The appropriate methodological question is not always how certain an agent is concerning a proposition. It may instead concern which observations establish the proposition, which deductive relations support it, which assumptions remain unresolved, and which consequences can legitimately be drawn from it.

These questions preserve the structure of the evidence rather than immediately compressing that structure into a scalar.

4 Artificial Orderings and Heuristic Commensuration

The presumption of measurability has a closely related counterpart in the presumption of orderability. Whereas the former concerns the assumption that a property admits a meaningful numerical representation, the latter concerns the assumption that a collection of objects admits a meaningful arrangement from inferior to superior, less significant to more significant, or less desirable to more desirable. The distinction is important because the construction of an ordering does not necessarily require the measurement of an independently existing quantity. One may arrange objects according to an arbitrary convention, an administrative requirement, a computational heuristic, or a preference that has been introduced for the purpose of facilitating a decision. Such arrangements can be useful without possessing the epistemic significance commonly attributed to them. An alphabetical arrangement of books does not establish a hierarchy of literary merit. The order in which a collection of errands is performed does not necessarily express their relative importance. A sequence of research tasks may reflect dependencies, resource constraints, or the convenience of the researcher rather than an ordering of their scientific significance.

The distinction between organizational convenience and evidentiary significance becomes obscured when the same numerical and linguistic conventions are used to express both.

4.1 Partial orderings and heterogeneous properties

Let $\mathcal{X}$ denote a collection of alternatives. Suppose each alternative $x \in \mathcal{X}$ possesses a vector of relevant properties,

$$f(x) = \Big(f_1(x), f_2(x), \ldots, f_n(x)\Big).$$

For simplicity, suppose each component is oriented such that larger values are preferable according to the corresponding criterion. A natural comparison relation is given by componentwise dominance:

$$x \succeq_P y \quad \Longleftrightarrow \quad f_j(x) \geq f_j(y) \quad \text{for every } j.$$

On the original collection of alternatives this relation is generally a preorder, since distinct alternatives may possess identical property vectors. It induces a partial order on the

corresponding equivalence classes. The associated strict dominance relation is

$$x \succ_P y$$

whenever $x \succeq_P y$ and at least one component satisfies a strict inequality. However, many pairs of alternatives will remain incomparable under this relation. It may be the case that

$$f_1(x) > f_1(y)$$

while

$$f_2(x) < f_2(y).$$

Neither alternative then dominates the other. This incomparability is not necessarily a defect of the representation, for it may faithfully express the structure of a decision problem in which the alternatives differ in ways that cannot be reduced to a single comparison without introducing additional assumptions.

A total ordering would require every pair of alternatives to be comparable. The construction of such an ordering therefore introduces information, conventions, or preferences that are not supplied by the componentwise relation alone.

Definition 4.1 (Artificial completion)**.** An artificial completion of a comparison relation is an extension that resolves otherwise incomparable alternatives through supplementary conventions, preferences, or aggregation rules not determined by the original comparison relation.

The term artificial is not intended pejoratively. An artificial completion may be entirely appropriate when a decision must be made despite the absence of a uniquely determined ordering. The methodological error occurs when the supplementary assumptions responsible for the completion are subsequently treated as properties discovered in the alternatives themselves.

4.2 Scalarization and preference dependence

A familiar method of constructing a complete preorder is to introduce a scalar utility function,

$$U_w(x) = \sum_{j=1}^{n} w_j f_j(x),$$

where the weights satisfy

$$w_j \geq 0, \qquad \sum_{j=1}^{n} w_j = 1.$$

For the present discussion, suppose that the component functions have already been expressed on scales suitable for the proposed aggregation. This assumption is itself substantive, because an arbitrary normalization of heterogeneous measurements does not establish their commensurability. The utility function defines the comparison

$$x \succeq_w y \quad \Longleftrightarrow \quad U_w(x) \geq U_w(y).$$

The resulting comparison is mathematically well-defined. Its interpretation nevertheless depends upon the choice of weights and the meaning of the component functions.

Proposition 4.1 (Preference-dependent reversal)**.** *Let x and y be alternatives for which there exist indices i and j satisfying*

$$f_i(x) > f_i(y)$$

and

$$f_j(x) < f_j(y).$$

Then there exist strictly positive normalized weight vectors w and v such that

$$U_w(x) > U_w(y)$$

while

$$U_v(x) < U_v(y).$$

Proof. Define

$$d_k = f_k(x) - f_k(y).$$

By assumption, $d_i > 0$ and $d_j < 0$. For $\varepsilon > 0$ sufficiently small, construct a weight vector $w^{(\varepsilon)}$ by setting

$$w_i^{(\varepsilon)} = 1 - (n-1)\varepsilon$$

and

$$w_k^{(\varepsilon)} = \varepsilon, \qquad k \neq i.$$

Then

$$U_{w^{(\varepsilon)}}(x) - U_{w^{(\varepsilon)}}(y) = \Big(1 - (n-1)\varepsilon\Big) d_i + \varepsilon \sum_{k \neq i} d_k.$$

As $\varepsilon \to 0^+$, this expression converges to $d_i > 0$. It is therefore positive for sufficiently small positive ε. Constructing a second weight vector concentrated on component j gives an analogous expression converging to $d_j < 0$.

Both weight vectors can be chosen to have strictly positive components summing to one. The two resulting scalarizations consequently produce opposite strict orderings. □

The proposition establishes that a reversal of ranking need not correspond to a change in any property of the ranked alternatives. The reversal can result entirely from a change in the supplementary preference structure. Consequently, the reproducibility of a ranking under fixed weights does not establish its independence from those weights. A deterministic procedure may produce exactly the same ordering on every execution while remaining dependent upon an arbitrary or insufficiently justified aggregation convention.

4.3 Itinerary planning as heuristic completion

Consider a computational system asked to plan a journey through several cities. The relevant considerations may include transportation costs, geographical proximity, historical interest, accessibility, accommodation, climatic conditions, available time, and the traveller's particular interests. Some of these considerations admit relatively straightforward numerical representation. Transportation costs can be expressed in a specified currency, distances in units of length, and travel durations as estimates under stated conditions. Other considerations do not possess an equally straightforward common scale. Historical interest, aesthetic preference, opportunities for intellectual activity, and the desirability of particular social environments may depend upon heterogeneous personal considerations. Even where individual components can be measured, their relative importance is not determined by their measurement. The request to produce an itinerary nevertheless requires a sequence. A computational system can satisfy this requirement by introducing heuristics, resolving ties, selecting convenient defaults, or completing an otherwise partial preference relation.

The resulting itinerary need not be interpreted as a claim that the selected destinations possess an objective hierarchy of desirability. Its legitimacy may instead consist in its practical adequacy under the available constraints. This distinction permits the use of heuristic orderings without attributing to them an unwarranted epistemic status.

Definition 4.2 (Pragmatic ordering)**.** A pragmatic ordering is a sequencing of alternatives justified by its usefulness for organizing or completing an activity, rather than by a claim that the alternatives possess an independently established total order.

A pragmatic ordering may be selected through a deterministic convention, a stochastic procedure, an approximate optimization algorithm, or an informal judgment. The selection mechanism can be evaluated according to its practical consequences. It need not be interpreted as measuring the intrinsic significance of the ordered objects. Indeed, where several alternatives satisfy all relevant constraints, random selection may be a more

transparent procedure than the construction of an elaborate but unjustified numerical hierarchy.

4.4 The distinction between feasibility and optimality

A further source of artificial ordering is the assumption that every practical decision must identify an optimum. Let $\mathcal{A}$ denote a collection of candidate actions and let

$$g_k(a) \leq 0, \qquad k = 1, \ldots, m,$$

represent the conditions that an action must satisfy. The admissible set is

$$\mathcal{A}_{\text{adm}} = \{a \in \mathcal{A} : g_k(a) \leq 0 \text{ for every } k\}.$$

If the practical objective is to identify an admissible action, the task is completed by finding any element

$$a \in \mathcal{A}_{\text{adm}}.$$

The introduction of an objective function

$$J : \mathcal{A}_{\text{adm}} \to \mathbb{R}$$

and the subsequent search for

$$a^* \in \arg\max_{a \in \mathcal{A}_{\text{adm}}} J(a)$$

constitute additional operations, which may be appropriate when a defensible objective has been specified but are not required merely because the admissible set contains multiple elements. This distinction is particularly relevant to autonomous computational systems. A system may need to select a permissible continuation without possessing a complete ordering of every permissible continuation. The unnecessary introduction of an objective function can transform a comparatively well-defined feasibility problem into a poorly justified optimization problem.

It may also create incentives to improve the numerical objective at the expense of considerations that were not represented within it. The problem is therefore not that heuristic selection is insufficiently sophisticated. It is that optimization can be imposed where the actual requirement is simply the satisfaction of constraints.

4.5 Scientific and institutional consequences

The distinction between pragmatic ordering and evidentiary ordering becomes increasingly important as numerical rankings acquire institutional consequences. A researcher may arrange several hypotheses in an order convenient for investigation. This arrangement does not establish their relative probabilities of truth. A funding organization may rank applications because it possesses a finite budget. The administrative necessity of making a selection does not establish that the scientific significance of the applications admits a uniquely correct scalar representation. A clinical system may prioritize cases according to an operational triage rule. The resulting sequence need not constitute a complete measurement of the severity, complexity, or importance of every patient's condition. In each case, the practical need to produce an ordering can be mistaken for evidence that an independently existing ordering has been discovered. This mistake is especially consequential when subsequent institutional decisions are justified by reference to the ranking rather than to the substantive considerations from which the ranking was constructed.

The numerical representation then acquires an authority that its original purpose did not warrant.

4.6 Heuristic networks and restricted inference

Probabilistic networks can be useful in domains where exhaustive representation is computationally infeasible or where the precision required for practical action is limited. A sparse Bayesian network, for example, may represent selected dependencies among variables through a factorization of the form

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{\mathrm{pa}(i)}),$$

where $\mathrm{pa}(i)$ denotes the parents of variable X_i in the chosen graph. The usefulness of the representation depends upon the suitability of its variables, conditional independence assumptions, and parameter estimates for the task under consideration. Sparsity can make inference tractable, but it can also omit dependencies that become important under circumstances not adequately represented by the model. A heuristic probabilistic network may therefore provide a useful approximation without constituting an exhaustive account of the causal or epistemic structure of the represented domain. The point is not an objection to Bayesian networks as mathematical objects but a restriction upon the conclusions that can legitimately be drawn from their successful application.

A representation constructed to facilitate approximate planning should not automati-

cally be treated as an instrument for establishing precise scientific facts. The required evidentiary standard depends upon the operation being performed and the consequences of an incorrect result.

5 Observability and the Non-Identifiability of Physical Quantities

The preceding section concerns the introduction of comparison relations that are not determined by the available properties of the alternatives. A different limitation arises when the target quantity is well-defined but the available observations do not determine its value. This is the familiar problem of non-identifiability in inverse problems and statistical inference. The distinction is essential, because a physical quantity may be perfectly meaningful and measurable in principle while remaining unrecoverable from a particular observation. In that case the difficulty concerns, not the legitimacy of the quantity, but the information supplied by the chosen observational procedure.

5.1 Observation maps and equivalence classes

Let $\mathcal{S}$ denote a space of admissible underlying states, let $\mathcal{O}$ denote a space of observations, and let

$$R : \mathcal{S} \longrightarrow \mathcal{O}$$

represent an observational procedure. Suppose the quantity of interest is represented by a function

$$T : \mathcal{S} \longrightarrow \mathcal{T},$$

where $\mathcal{T}$ denotes the target space. The inverse problem consists in determining $T(s)$ from the observation

$$o = R(s).$$

The observational procedure induces an equivalence relation on $\mathcal{S}$:

$$s_1 \sim_R s_2 \quad \Longleftrightarrow \quad R(s_1) = R(s_2).$$

The equivalence class associated with an observation o is the fibre

$$\mathcal{S}_o = R^{-1}(\{o\}) = \{s \in \mathcal{S} : R(s) = o\}.$$

An observation therefore identifies an equivalence class of underlying states without necessarily identifying an individual state within that class. The target quantity is

exactly recoverable only if the distinctions suppressed by the observational procedure are irrelevant to the target.

Definition 5.1 (Observational identifiability)**.** A target T is observationally identifiable with respect to R on $\mathcal{S}$ if

$$R(s_1) = R(s_2) \quad \Longrightarrow \quad T(s_1) = T(s_2)$$

for every $s_1, s_2 \in \mathcal{S}$.

This definition concerns exact recovery under a deterministic observation map. Statistical identifiability under noisy observations requires a related but distinct formulation, considered below.

Proposition 5.1 (Factorization criterion)**.** *A target T is observationally identifiable with respect to R if and only if there exists a function*

$$F : R(\mathcal{S}) \longrightarrow \mathcal{T}$$

such that

$$T = F \circ R.$$

Proof. Suppose first that such a function F exists. If $R(s_1) = R(s_2)$, then

$$T(s_1) = F(R(s_1)) = F(R(s_2)) = T(s_2).$$

Thus T is observationally identifiable. Conversely, suppose T is observationally identifiable. For each observation $o \in R(\mathcal{S})$, define

$$F(o) = T(s)$$

for any s satisfying $R(s) = o$. The identifiability condition guarantees that the value of $T(s)$ is independent of the representative chosen from the fibre $R^{-1}(\{o\})$. The function F is therefore well-defined, and by construction

$$F(R(s)) = T(s)$$

for every $s \in \mathcal{S}$. □

Corollary 5.1 (Non-identifiability obstruction)**.** *If there exist admissible states s_1 and s_2 satisfying*

$$R(s_1) = R(s_2)$$

but

$$T(s_1) \neq T(s_2),$$

then no function of the observation alone can recover the target correctly for every admissible underlying state.

This obstruction is independent of the computational complexity of the proposed estimator. A larger neural network, a more elaborate inference algorithm, or an increase in available computational resources cannot establish a distinction that the observation map has eliminated. Additional assumptions may restrict the admissible state space, and additional observations may refine the equivalence classes, but neither operation should be confused with the extraction of information already present in the original observation.

5.2 A lower bound on unavoidable estimation error

The preceding obstruction concerns exact recovery. A quantitative version establishes a lower bound on the error of approximate recovery. Suppose the target takes values in a metric space $(\mathcal{T}, d)$ and let

$$\widehat{T} : \mathcal{O} \longrightarrow \mathcal{T}$$

be an arbitrary deterministic estimator.

Proposition 5.2 (Indistinguishable-state error bound)**.** *If $s_1, s_2 \in \mathcal{S}$ satisfy*

$$R(s_1) = R(s_2) = o,$$

then

$$\max\Big\{d\Big(\widehat{T}(o), T(s_1)\Big), d\Big(\widehat{T}(o), T(s_2)\Big)\Big\} \geq \frac{1}{2} d\Big(T(s_1), T(s_2)\Big).$$

Proof. By the triangle inequality,

$$\begin{aligned} d\Big(T(s_1), T(s_2)\Big) \leq\ & d\Big(T(s_1), \widehat{T}(o)\Big) \\ & + d\Big(\widehat{T}(o), T(s_2)\Big). \end{aligned}$$

At least one of the two terms on the right-hand side must therefore be at least half the left-hand side. □

The result establishes a worst-case limitation, which does not imply that every estimator must perform poorly on average under every distribution of underlying states. A probabilistic estimator may achieve small expected error if states producing large discrepancies are sufficiently rare under the relevant distribution. However, such a result depends upon the distributional assumptions and does not remove the observational equivalence responsible for the worst-case limitation. The same reasoning extends to randomized estimators. If the estimator receives the same observation in both states, its output distribution is the same in both cases. Taking expectations in the triangle inequality yields

$$\max_{i\in\{1,2\}} \mathbb{E}\left[d\Big(\widehat{T}(o), T(s_i)\Big)\right] \geq \frac{1}{2} d\Big(T(s_1), T(s_2)\Big).$$

Randomization may change the distribution of estimation errors, but it cannot independently distinguish the underlying states.

5.3 Nutritional estimation from photographs

Consider the increasingly common application of computational vision to the estimation of the nutritional content of meals. The energy content of a meal is a meaningful physical and nutritional quantity, provided the relevant definition of energy content has been specified. Gross chemical energy, digestible energy, and metabolizable energy are not identical quantities. Their determination involves different procedures and, in the latter cases, assumptions concerning biological utilization. For present purposes, let $E(s)$ denote a specified nutritional energy quantity associated with the complete state s of a meal. Let

$$I = R(s)$$

denote a photograph of the meal. The photograph may reveal information about the apparent size, shape, colour, texture, and arrangement of visible components. It need not reveal the complete chemical composition of those components, the mass of ingredients concealed from view, the amount of oil absorbed during preparation, the concentration of dissolved substances, or the identity of visually indistinguishable substitutions.

Consider two otherwise similar beverages, one prepared with a substantial quantity of dissolved sugar and another prepared with a low-calorie sweetener. Under an ordinary imaging procedure, it is possible for their visible appearances to be indistinguishable at the available resolution while their nutritional energy contents differ. The corresponding states satisfy

$$R(s_1) = R(s_2)$$

at the resolution of the recorded observation, while

$$E(s_1) \neq E(s_2).$$

The non-identifiability obstruction therefore applies. An estimator may exploit population-level regularities concerning how beverages of the observed appearance are usually prepared. It may provide a useful conditional expectation under a specified distribution of recipes. Such an expectation is not equivalent to a measurement of the energy content of the particular beverage.

5.4 Conditional expectation and individual composition

Suppose a probability distribution has been introduced over admissible meal states. Under squared-error loss, an optimal estimator based exclusively on the photograph is the conditional expectation

$$\widehat{E}(I) = \mathbb{E}[E(S) \mid R(S) = I].$$

This quantity may be statistically useful, since it represents the expected nutritional energy content of states compatible with the observation, weighted according to the specified conditional distribution. Its interpretation nevertheless depends upon the distribution of underlying meal states. Assume that $E(S)$ is square-integrable, so that the squared-error optimization problem is well-defined, and suppose further that the conditional distribution given I is concentrated upon two states s_1 and s_2 of the observational equivalence class, with energy contents e_1 and e_2. Let

$$p = \mathbb{P}(S = s_1 \mid I).$$

The conditional expectation becomes

$$\widehat{E}(I) = pe_1 + (1 - p)e_2.$$

If $0 < p < 1$ and $e_1 \neq e_2$, then the estimate lies strictly between the two possible energy contents. In particular, it need not equal the energy content of either admissible state. This is no mathematical defect of conditional expectation but a consequence of the objective being optimized, for an estimator that minimizes expected squared error under the specified distribution does not thereby identify the physical composition of the observed meal.

5.5 Refinement through additional observation

The informational limitations of a single photograph can be reduced by acquiring additional observations. A photographic record of the stages of meal preparation may reveal ingredients concealed in the final presentation. Measurements of ingredient masses before and after use may constrain the quantities incorporated into the meal. Ingredient labels may supply compositional information. Records of discarded material, cooking losses, and absorbed substances may further constrain the final composition. Let

$$R_1 : \mathcal{S} \to \mathcal{O}_1$$

denote the original imaging procedure and let

$$R_2 : \mathcal{S} \to \mathcal{O}_2$$

denote an additional observational procedure. The combined observation map is

$$R_{12}(s) = \Big(R_1(s), R_2(s)\Big).$$

Its fibres satisfy

$$R_{12}^{-1}(\{(o_1, o_2)\}) = R_1^{-1}(\{o_1\}) \cap R_2^{-1}(\{o_2\}).$$

The additional observation can therefore refine the original equivalence classes.

Proposition 5.3 (Monotonicity under exact observational refinement)**.** *If a target T is identifiable with respect to R_1, then it is identifiable with respect to the combined observation map R_{12}.*

Proof. Suppose

$$R_{12}(s_1) = R_{12}(s_2).$$

Then, by equality of the first components,

$$R_1(s_1) = R_1(s_2).$$

Identifiability with respect to R_1 implies

$$T(s_1) = T(s_2).$$

Thus T is identifiable with respect to R_{12}. □

The proposition does not establish that every additional observation strictly improves

identifiability. An additional photograph may reveal no information relevant to the target. A second sensor may reproduce the same ambiguity as the first. Several measurements may be affected by a common systematic error. The relevant question is whether the additional observation separates states that were previously indistinguishable with respect to the target quantity. The epistemic improvement consists in the refinement of the admissible state space, not merely in an increase in the volume of recorded data.

5.6 Set-valued estimates and evidentiary bounds

When a target is not identifiable, the available observations may still constrain its possible values. For an observation o, define the identified set

$$\mathcal{T}(o) = \{T(s) : s \in \mathcal{S}_o\}.$$

For a real-valued target, one may construct bounds

$$\underline{T}(o) = \inf_{s \in \mathcal{S}_o} T(s)$$

and

$$\overline{T}(o) = \sup_{s \in \mathcal{S}_o} T(s).$$

Provided the true underlying state belongs to $\mathcal{S}$, the target satisfies

$$\underline{T}(o) \leq T(s) \leq \overline{T}(o).$$

These bounds express the limitations of the observation relative to the admissible state space. They do not require the introduction of a probability distribution over the states within the observational equivalence class. A probabilistic representation can be added where an appropriate distribution is available. However, the identified set and the probability distribution over that set are different epistemic objects. The former describes which target values remain compatible with the observations and assumptions. The latter describes how probability mass has been allocated among those possibilities. A narrow credible interval can be obtained within a broad identified set by assigning most prior probability to a small subset of the compatible states.

Such concentration may be justified by substantive knowledge, but it should not be interpreted as observational resolution supplied by the original measurement. This distinction is central to the literature on partial identification (Manski, 2003).

5.7 Testimony as an additional evidentiary channel

Suppose a person states that the meal was prepared using a particular low-calorie sweetener. The statement may provide relevant information, but it is not identical to a chemical measurement of the meal. The evidentiary relationship between the statement and the underlying physical state depends upon the person's knowledge, recollection, accuracy of expression, and truthfulness. Let Q denote a reported claim concerning the composition of the meal. A probabilistic model may incorporate this testimony through a conditional distribution

$$P(Q \mid S).$$

The posterior distribution then takes the form

$$P(S \mid I, Q) \propto P(I, Q \mid S)P(S).$$

Under an additional conditional independence assumption, one could factor the likelihood as

$$P(I, Q \mid S) = P(I \mid S)P(Q \mid S).$$

The factorization is not automatic. The person's report may depend upon the appearance of the meal, upon information supplied by the computational system, or upon other observations correlated with the photograph. Moreover, unless the testimony is assumed to be infallible, the reported ingredient does not eliminate every underlying state in which the ingredient is absent.

The assertion that the meal contains a particular substance is itself an observation of a communicative event. Its relation to the physical composition of the meal requires a further evidentiary model. This distinction generalizes to scientific reports, institutional records, human annotations, and the output of other computational systems. The existence of a statement about a physical state does not by itself establish that state.

5.8 Statistical identifiability under noisy observations

The deterministic observation-map formulation is useful for exposing the structural limitation of exact recovery. Many scientific measurements, however, involve stochastic observations. Let

$$P_s$$

denote the probability distribution of observations conditional upon the underlying state s. A target T is statistically identifiable within the model if

$$P_{s_1} = P_{s_2} \quad \Longrightarrow \quad T(s_1) = T(s_2).$$

Here equality refers to equality of the observation distributions, rather than equality of individual observed outcomes. The distinction matters because two states may produce overlapping observation distributions while remaining statistically distinguishable through repeated observations. Conversely, if their complete observation distributions coincide, then independent repetitions of the same observational procedure cannot distinguish them. For independent observations,

$$P_{s_1}^{\otimes n} = P_{s_2}^{\otimes n}$$

for every sample size n whenever

$$P_{s_1} = P_{s_2}.$$

Increasing the sample size can reduce sampling uncertainty in an identifiable model. It cannot resolve structural non-identifiability within an unchanged observational model. This distinction is especially important when the proposed remedy for an inadequate measurement procedure consists solely in collecting more observations of the same kind.

More data are not necessarily more informative about the particular distinction under investigation.

5.9 The operational significance of approximation

None of the preceding results implies that nutritional estimation from photographs is necessarily without practical value. An approximate estimate may be adequate for an informal dietary record, particularly when the user understands its dependence upon typical portion sizes, ingredient assumptions, and population-level regularities. The same estimate may be inadequate for an application requiring precise knowledge of chemical composition or nutrient content. The distinction is not exhausted by the numerical confidence assigned to the estimate. It concerns the relationship between the evidentiary resolution of the procedure and the requirements of the intended operation. A heuristic can be appropriate for one purpose and inadmissible for another without undergoing any change in its computational implementation. This observation anticipates the later distinction between statistical reliability and operational authorization.

6 Causal Ambiguity and Observational Equivalence

The nutritional example concerns distinct physical compositions that produce indistinguishable observations. A related problem arises when the same observable property can be generated by different causal mechanisms. A classification procedure may exploit a regular association between an observable feature and a target category. Its reliability can

deteriorate when an alternative mechanism produces the observable feature independently of the target. The difficulty is not adequately described as random measurement noise, because the alternative mechanism may produce a highly structured and entirely accurate observation of a property whose causal significance has changed.

6.1 The spray-painted tree

Consider a computational system trained to identify tree species from photographs of their leaves. Leaf colour may provide useful information because pigmentation, physiological processes, seasonal changes, and environmental conditions exhibit relationships with botanical classification. Suppose, however, that a person applies yellow and orange paint to the leaves of a tree. The resulting photograph may accurately record the visible colour of the leaves while failing to distinguish between natural pigmentation and the artificial intervention. The observation is not thereby incorrect, since the leaves may genuinely appear yellow and orange. What has become unreliable is the inference from their visible colour to the biological processes ordinarily responsible for that colour. Let S denote tree species, B denote biological pigmentation, A denote artificial intervention, and C denote observed colour. A simplified structural model might take the form

$$B = b(S, E, U_B),$$

where E represents environmental conditions and U_B represents additional influences upon biological pigmentation.

Observed colour is then given by

$$C = c(B, A, L, U_C),$$

where L denotes illumination and U_C represents further observational influences. A classifier trained primarily on naturally occurring leaves may implicitly rely upon the restriction

$$A = 0.$$

Within that restricted domain, the relationship between S and C may support useful classification. The introduction of states with

$$A \neq 0$$

changes the set of mechanisms capable of producing the observed colour. If artificial pigmentation can reproduce the appearance ordinarily associated with a different species, the classifier's original association no longer suffices to identify the underlying biological

category.

6.2 Mechanism-dependent inference

The probabilistic relationship

$$P(S \mid C)$$

depends upon the joint distribution of species, pigmentation, environmental conditions, artificial interventions, and observational conditions. By marginalizing over the intervention variable, one obtains

$$P(S \mid C) = \sum_{a} P(S \mid C, A = a) P(A = a \mid C),$$

assuming for simplicity that A takes discrete values. A classifier trained exclusively under the condition $A = 0$ instead approximates a relationship of the form

$$P(S \mid C, A = 0).$$

These two conditional distributions need not coincide. If the deployed classifier encounters observations generated by mechanisms absent from its training domain, the historical relationship between colour and species may no longer describe the relevant inference problem. The important distinction is between the observation of a feature and the identification of the mechanism responsible for that feature. A system can correctly identify the colour of an object while incorrectly identifying its biological origin. The former operation concerns an observable property, whereas the latter concerns a latent classification inferred through assumptions about the causal structure of the domain.

6.3 A mechanism-ambiguity obstruction

The preceding example admits a general formulation. Let $\mathcal{M}$ denote a collection of admissible causal mechanisms. For each $m \in \mathcal{M}$, let

$$R_m : \mathcal{S} \longrightarrow \mathcal{O}$$

denote the observation map associated with that mechanism. Suppose the target is

$$T : \mathcal{S} \longrightarrow \mathcal{T}.$$

The observer receives

$$o = R_m(s)$$

but does not necessarily observe m.

Proposition 6.1 (Cross-mechanism non-identifiability)**.** *Suppose there exist mechanisms $m_1, m_2 \in \mathcal{M}$ and states $s_1, s_2 \in \mathcal{S}$ such that*

$$R_{m_1}(s_1) = R_{m_2}(s_2)$$

while

$$T(s_1) \neq T(s_2).$$

Then no estimator depending exclusively upon the observation can recover the target correctly for every admissible state and mechanism.

Proof. Define the enlarged state space $\widetilde{\mathcal{S}} = \mathcal{S} \times \mathcal{M}$, the observation map $\widetilde{R}(s, m) = R_m(s)$, and the target $\widetilde{T}(s, m) = T(s)$. The hypotheses state that

$$\widetilde{R}(s_1, m_1) = \widetilde{R}(s_2, m_2) \quad \text{while} \quad \widetilde{T}(s_1, m_1) \neq \widetilde{T}(s_2, m_2),$$

and the conclusion follows from Corollary 5.1. □

The proposition is therefore the factorization criterion of Proposition 5.1 applied to an enlarged state space. It emphasizes that identifiability can depend not only upon the state space but also upon the collection of mechanisms admitted by the observational model. A target may be identifiable within each of several restricted mechanisms while failing to be identifiable when the mechanism itself is unknown.

6.4 Intervention and the inadequacy of correlation

The spray-painted tree illustrates why observational association does not automatically establish a stable causal relationship. A classifier may learn that certain colours are strongly associated with particular species under ordinary conditions. An intervention can modify colour without modifying species. In the language of structural causal models, the distribution of an observable feature under intervention need not coincide with its observational distribution (Pearl, 2009; Peters et al., 2017). The distinction is commonly expressed through the difference between

$$P(Y \mid X = x)$$

and

$$P(Y \mid \mathrm{do}(X = x)).$$

The first quantity concerns conditioning upon an observed value, whereas the second concerns the distribution resulting from an intervention that sets the variable to that value. Their equality requires appropriate causal assumptions. In the botanical example, the application of paint is an intervention upon the mechanism producing visible colour. It leaves the biological species unchanged.

The example therefore exposes the limitations of treating a correlated observable feature as an invariant indicator of the target category. The appropriate remedy may involve additional observations, such as leaf morphology, branching structure, bark, chemical composition, or genetic information. It may also involve explicit recognition that the available observation does not establish whether an artificial intervention has occurred. Neither remedy consists simply in assigning a lower confidence score to the original classification.

6.5 Distributional change and transportability

The performance of a classifier is ordinarily evaluated relative to a distribution of observations and target labels. Let P_{train} denote a training distribution and let P_{deploy} denote the distribution encountered during deployment. For a classifier h and loss function ℓ, the corresponding risks are

$$\mathcal{R}_{\text{train}}(h) = \mathbb{E}_{P_{\text{train}}}[\ell(h(X), Y)]$$

and

$$\mathcal{R}_{\text{deploy}}(h) = \mathbb{E}_{P_{\text{deploy}}}[\ell(h(X), Y)].$$

A small value of the first quantity does not establish a small value of the second. The relationship depends upon the nature of the distributional change and the invariance of the features exploited by the classifier. The introduction of an alternative causal mechanism may change the conditional distribution

$$P(Y \mid X),$$

rather than merely changing the frequency with which particular observations occur. In such circumstances, reweighting observations according to their frequency may be insufficient to restore the original classification relationship. The relevant causal or statistical invariances must be established rather than assumed.

The problem of transporting a result from one population or environment to another therefore cannot be reduced to the precision with which the result was estimated in its original domain.

6.6 Biochemical classification and developmental heterogeneity

The same methodological distinctions arise in attempts to infer developmental or psychological characteristics from biochemical measurements. Consider a proposed system for predicting an autism diagnosis from a collection of chemical biomarkers. This example is introduced as a methodological case, rather than as an assessment of a particular published study. Let

$$X = (X_1, \ldots, X_n)$$

denote the measured biochemical variables and let Y denote a specified diagnostic label. A statistical investigation may establish that the distribution of X differs between populations assigned different values of Y. Such an investigation may also establish that a classifier

$$h : X \longrightarrow \{0, 1\}$$

possesses predictive value within a specified validation population. These are potentially legitimate empirical results. They do not, without further evidence, establish that the biochemical measurements identify a unique developmental mechanism, that the classifier is transportable to every population, or that the diagnostic category corresponds to a single measurable biochemical quantity.

Autism is a heterogeneous neurodevelopmental condition identified through clinical and developmental criteria. A biochemical association with an assigned diagnostic label should therefore be distinguished from the identification of the developmental processes responsible for the characteristics of a particular individual. The distinction is especially important when a statistical classifier is described as measuring the condition itself. The classifier may instead be predicting the assignment of a diagnostic label under the particular recruitment, measurement, and assessment procedures used to construct the dataset.

6.7 Prediction, diagnosis, and causal explanation

At least three distinct scientific objectives can arise in the biochemical example. The first is prediction: whether the observed variables provide information about the diagnostic label assigned to an individual. The second is diagnostic validity: whether the measurement procedure provides adequate evidence for a clinical determination under the relevant diagnostic criteria and conditions of use. The third is causal explanation: whether the observed variables participate in, result from, or otherwise illuminate the developmental mechanisms responsible for the characteristics being investigated. Success with respect to one objective does not automatically establish success with respect to the others. A variable may predict a diagnostic label because it is associated with medication use,

dietary patterns, age, co-occurring conditions, environmental exposure, or recruitment procedures. Such an association may be scientifically informative without establishing that the variable causes the developmental condition.

Conversely, a variable may participate in a causal mechanism without possessing sufficient discriminatory power to serve as an individual diagnostic indicator. The predictive utility of a biomarker, its causal significance, and its clinical admissibility are therefore distinct properties.

6.8 Reference standards and construct validity

A classifier evaluated against a diagnostic label inherits limitations associated with the procedure by which that label was assigned. Let Z denote a collection of underlying developmental characteristics and let Q denote an assessment procedure. The observed label may be represented schematically as

$$Y = Q(Z, E, U),$$

where E denotes relevant environmental and contextual conditions and U denotes additional influences upon the assessment. A model trained to predict Y from biochemical measurements X is evaluated against the output of Q, not directly against every aspect of Z. This distinction is familiar from the methodology of construct validity (Cronbach and Meehl, 1955). A high degree of agreement between a predictor and an operational label does not establish that either procedure exhaustively represents the underlying construct. The appropriate validation requirements depend upon the intended interpretation of the predictor.

If the intended purpose is preliminary screening, the relevant requirements may differ from those associated with definitive diagnosis, causal investigation, or individual treatment decisions. The error consists in allowing a numerical performance measure obtained for one of these purposes to acquire authority over the others without additional validation.

6.9 Base rates and predictive interpretation

Even within a well-defined classification problem, the interpretation of a positive prediction depends upon the population in which the procedure is deployed. Let

$$\pi = \mathbb{P}(Y = 1)$$

denote the prevalence of the target label, let

$$\alpha = \mathbb{P}(\widehat{Y} = 1 \mid Y = 1)$$

denote sensitivity, and let

$$\beta = \mathbb{P}(\widehat{Y} = 0 \mid Y = 0)$$

denote specificity. The positive predictive value is

$$\mathbb{P}(Y = 1 \mid \widehat{Y} = 1) = \frac{\alpha\pi}{\alpha\pi + (1-\beta)(1-\pi)},$$

provided the denominator is positive. Consider a hypothetical classifier with

$$\alpha = \beta = 0.95.$$

In a population with prevalence $\pi = 0.01$, its positive predictive value would be

$$\frac{0.95(0.01)}{0.95(0.01) + 0.05(0.99)} \approx 0.161.$$

In a population with prevalence $\pi = 0.20$, the same sensitivity and specificity would yield

$$\frac{0.95(0.20)}{0.95(0.20) + 0.05(0.80)} \approx 0.826.$$

These figures are illustrative and are not estimates of the performance of any particular autism screening procedure. They demonstrate that even an apparently impressive classification performance cannot be interpreted independently of the population and the particular conditional probability under consideration. Moreover, the calculation assumes that sensitivity and specificity remain unchanged between the two populations.

That invariance is itself an empirical assumption, not a consequence of Bayes' theorem.

6.10 Selective ignorance and the excluded mechanism

The botanical and biochemical examples share a methodological structure. An observable feature is selected because it possesses an association with a target category within a specified domain. A computational model is then constructed to exploit that association. The model may perform well under conditions in which the mechanisms responsible for the association remain sufficiently stable. However, alternative mechanisms may produce the same observable feature, and the model may possess no representation of those mechanisms. The resulting ignorance is not necessarily expressed in the model's

confidence estimate. If the possibility of artificial pigmentation is absent from the botanical model, the model may assign a high probability to a species whose natural pigmentation resembles the observed colour. The estimate may be internally coherent relative to the restricted model. Its inadequacy concerns the exclusion of an admissible explanation.

This distinction motivates the following section, in which uncertainty within a probabilistic representation is distinguished from uncertainty concerning the adequacy of the representation itself.

7 Probabilistic Closure and Excluded Possibilities

The preceding examples establish that an observational procedure can suppress distinctions that remain relevant to the target of inference. A related difficulty arises when the probabilistic representation itself excludes possibilities that remain admissible in the represented domain. A probability distribution is normalized relative to a specified space of events or hypotheses. Its normalization does not establish that this space adequately represents the phenomena under investigation. This distinction is elementary from a mathematical perspective, but its methodological consequences are frequently obscured by the interpretation of posterior probabilities as comprehensive measures of epistemic certainty.

7.1 Internal coherence and external adequacy

Let $\mathcal{H}$ denote a finite collection of mutually exclusive hypotheses. A probability distribution over $\mathcal{H}$ satisfies

$$\sum_{h \in \mathcal{H}} P(h) = 1.$$

Following the acquisition of evidence D, Bayesian updating produces

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{\sum_{g \in \mathcal{H}} P(D \mid g)P(g)},$$

provided the denominator is positive. The resulting posterior distribution satisfies

$$\sum_{h \in \mathcal{H}} P(h \mid D) = 1.$$

Both normalization conditions concern the allocation of probability mass within the specified hypothesis space. Neither establishes that the actual mechanism responsible for

the evidence belongs to that space.

Definition 7.1 (Representational closure)**.** A probabilistic representation is closed relative to a hypothesis space $\mathcal{H}$ when its allocation and updating of probability mass are restricted to hypotheses belonging to $\mathcal{H}$.

Representational closure is ordinarily necessary for the implementation of a particular inference procedure. A computational system cannot explicitly evaluate an indefinitely extensible collection of unformulated explanations. The difficulty arises when closure of the representation is mistaken for closure of the domain being represented. The mathematical completeness of a probability distribution is not equivalent to the explanatory completeness of its hypothesis space.

Definition 7.2 (Representational adequacy)**.** A hypothesis space is adequate for a specified inferential task to the extent that it preserves the distinctions and includes the mechanisms necessary to support the intended interpretation and use of the resulting inference.

Adequacy is task-dependent. A simplified model may be adequate for estimating an aggregate quantity while remaining inadequate for identifying the mechanism responsible for an individual observation. The requirement is not that every useful model provide an exhaustive description of reality. Rather, the omitted distinctions must not invalidate the particular conclusions or operations for which the model is employed.

7.2 Posterior concentration under misspecification

Consider a hypothesis space consisting of two candidate mechanisms,

$$\mathcal{H} = \{h_1, h_2\}.$$

Suppose these mechanisms assign probabilities p_1 and p_2 to a binary observation, where

$$p_1 = 0.8, \qquad p_2 = 0.2.$$

Assume, however, that the observations are actually generated by a third mechanism h_*, absent from the hypothesis space, with

$$p_* = 0.7.$$

Let $X_1, \ldots, X_n$ be independent Bernoulli observations generated under h_*. The posterior odds between the represented hypotheses are

$$\frac{P(h_1 \mid X_{1:n})}{P(h_2 \mid X_{1:n})} = \frac{P(h_1)}{P(h_2)} \prod_{i=1}^{n} \frac{P(X_i \mid h_1)}{P(X_i \mid h_2)}.$$

Taking logarithms gives

$$\begin{aligned} \log \frac{P(h_1 \mid X_{1:n})}{P(h_2 \mid X_{1:n})} = \; & \log \frac{P(h_1)}{P(h_2)} \\ & + \sum_{i=1}^{n} \log \frac{P(X_i \mid h_1)}{P(X_i \mid h_2)}. \end{aligned}$$

Under the actual generating mechanism, the expected contribution of each observation to the log posterior odds is

$$\begin{aligned} \mathbb{E}_* \left[\log \frac{P(X_i \mid h_1)}{P(X_i \mid h_2)} \right] &= 0.7 \log \frac{0.8}{0.2} + 0.3 \log \frac{0.2}{0.8} \\ &= 0.4 \log 4 \\ &> 0. \end{aligned}$$

Assuming both represented hypotheses possess positive prior probability, the law of large numbers implies that the posterior probability of h_1 converges almost surely to one. Thus,

$$P(h_1 \mid X_{1:n}) \longrightarrow 1,$$

even though h_1 is not the actual generating mechanism. The posterior concentration is mathematically appropriate relative to the restricted hypothesis space. What it identifies is the better-supported member of that space.

It does not establish that the selected hypothesis is true. This phenomenon is related to classical results concerning Bayesian inference under model misspecification (Berk, 1966; Kleijn and van der Vaart, 2006). The example is deliberately elementary because the relevant distinction does not require an elaborate statistical construction. An inference procedure can become increasingly decisive while remaining structurally incapable of representing the correct explanation.

7.3 Relative support and absolute adequacy

The preceding example distinguishes two questions that are often conflated in the interpretation of statistical evidence. The first asks which member of a specified collection of hypotheses is best supported by the available observations. The second asks whether

any member of that collection adequately represents the phenomenon. A posterior probability can answer the first question without answering the second. Similarly, a likelihood ratio

$$\Lambda(D) = \frac{P(D \mid h_1)}{P(D \mid h_2)}$$

compares the evidentiary support assigned to two hypotheses under their respective likelihood models. A large value of $\Lambda(D)$ does not establish that either hypothesis provides an adequate absolute description of the observations. Both likelihoods may be extremely small, or both models may systematically fail to reproduce features of the data that are relevant to the scientific question. Model comparison must therefore be distinguished from model criticism. The former operates within a collection of specified alternatives. The latter examines whether the collection, its assumptions, and its observational consequences remain adequate for the intended task.

This distinction is especially important when a model-selection procedure is presented as having discovered the correct explanation rather than having selected a preferred member of a restricted candidate set.

7.4 Excluded hypotheses and zero prior probability

The exclusion of a hypothesis can occur either explicitly, through its absence from the hypothesis space, or implicitly, through the assignment of zero prior probability. Suppose a hypothesis h_* is included formally but satisfies

$$P(h_*) = 0.$$

Under ordinary Bayesian conditioning,

$$P(h_* \mid D) = 0$$

for every observation D having positive marginal probability. No accumulation of evidence can restore positive posterior probability to that hypothesis through the specified updating operation alone. The limitation is a consequence of the support of the prior distribution and not a defect of Bayes' theorem. The recovery of an excluded explanation requires a modification of the representation rather than another application of the same updating rule. A distinction must therefore be maintained between updating a distribution over a fixed hypothesis space and revising the hypothesis space itself. The former operation reallocates probability among represented possibilities, whereas the latter introduces or modifies the possibilities among which probability may subsequently be allocated.

These operations have different epistemic significance.

7.5 The distinction between uncertainty and ignorance

The term uncertainty is frequently used to describe several different limitations upon knowledge. One limitation concerns the unknown value of a parameter within an otherwise adequate model. Another concerns the occurrence of a future event governed by a specified stochastic process. A third concerns the inadequacy of the model itself, including the possibility that the relevant variables, mechanisms, or distinctions have not been represented. The first two limitations can often be expressed through ordinary probability distributions. The third may require a revision of the representation before a defensible probability distribution can be constructed.

Definition 7.3 (Internal uncertainty)**.** Internal uncertainty is uncertainty concerning events, parameters, or hypotheses represented within a specified inferential model.

Definition 7.4 (Representational ignorance)**.** Representational ignorance is the absence of distinctions, variables, mechanisms, or possibilities whose inclusion may be necessary for the adequacy of an inferential model.

The distinction does not imply that uncertainty about models can never be represented probabilistically. Hierarchical models, Bayesian model averaging, robust Bayesian procedures, and related methods can represent uncertainty across collections of models. However, these procedures introduce a further space of represented alternatives. They do not eliminate the possibility that relevant alternatives are absent from that space. A distribution over models remains a distribution over specified models, so that the introduction of another probabilistic level therefore does not independently resolve the problem of representational adequacy.

7.6 Sensitivity to the admissible model class

Let $\mathfrak{M}$ denote a collection of admissible probabilistic models. Rather than selecting a single model and reporting its posterior probability, one may examine the range of probabilities obtained across $\mathfrak{M}$. For a hypothesis H and evidence D, define

$$\underline{P}(H \mid D) = \inf_{M \in \mathfrak{M}} P_M(H \mid D)$$

and

$$\overline{P}(H \mid D) = \sup_{M \in \mathfrak{M}} P_M(H \mid D),$$

where the conditional probabilities are assumed to be well-defined. The interval

$$\left[\underline{P}(H \mid D), \overline{P}(H \mid D)\right]$$

expresses sensitivity to the represented variation in model assumptions. A narrow interval indicates agreement across the specified collection of models, whereas a wide interval indicates that the inferential conclusion depends substantially upon the choice of model. This approach can reveal dependencies that would be concealed by reporting a single posterior probability. Nevertheless, its interpretation remains conditional upon the adequacy of $\mathfrak{M}$. The procedure cannot establish robustness against mechanisms that have not been included or otherwise bounded. Sensitivity analysis is therefore valuable precisely because it makes some of the dependence upon representational choices explicit, not because it abolishes that dependence.

7.7 Heuristic closure and practical adequacy

A practical computational system must ordinarily operate with incomplete representations. A journey planner does not require a complete physical model of every vehicle, road surface, pedestrian, and meteorological process that might influence a journey. A nutritional estimator used for informal record-keeping may reasonably employ population-level approximations. A customer-support system may rely upon a restricted collection of documents rather than attempting to reconstruct the entire institutional history of the organization. Such restrictions are not inherently methodological failures, and their legitimacy depends upon the relationship between the distinctions omitted and the consequences of the operations performed. A restricted representation can be adequate for a reversible, inexpensive, or exploratory action while remaining inadequate for an irreversible or high-consequence action. The relevant question is therefore not whether the representation is complete in an absolute sense but whether its limitations have been adequately characterized relative to the intended continuation.

This provides a basis for distinguishing the pragmatic usefulness of heuristic inference from the evidentiary authority sometimes attributed to its outputs.

7.8 The absence of a probability is not a probability

When a relevant mechanism has not been represented, the system may lack a defensible numerical estimate of its occurrence. It does not follow that the probability of that mechanism is zero. Nor does it follow that the appropriate probability is one half, or that an arbitrary small residual probability should be assigned to an unspecified collection of unknown possibilities. A residual category can be useful when its interpretation and observational consequences are sufficiently defined. However, assigning probability mass to a category labelled “other” does not automatically establish the properties of the mechanisms collected within it. In particular, the operational consequences of

an unrepresented mechanism may remain unknown even when a residual probability has been assigned. The absence of a justified numerical estimate should therefore be preserved as a distinct epistemic condition.

A system capable of expressing only numerical probabilities may be forced to conceal this condition through an arbitrary assignment. A more expressive system must be capable of distinguishing a low probability from an unavailable probability, an unresolved model assumption, and an unidentified mechanism. These distinctions become especially important when the resulting representation is used to authorize actions.

8 Calibration Without Certification

The preceding sections concern limitations that arise before a confidence estimate can be interpreted as a meaningful statistical object. The present section returns to the central argument of the podcast and considers the conditions under which a numerical confidence estimate can possess demonstrated empirical validity. The distinction between a verbalized confidence statement and an empirically calibrated estimate should be granted in its entirety. A system that produces a percentage has not thereby demonstrated the statistical properties ordinarily associated with that percentage. However, the mechanism through which the percentage is expressed does not establish that those properties are absent. The question must be resolved through an examination of the estimation procedure and its observed performance.

8.1 The definition of calibration

Let X denote a task presented to a computational system, let A denote the answer produced, and let

$$Y \in \{0, 1\}$$

indicate whether the answer satisfies a specified correctness criterion. Let

$$C \in [0, 1]$$

denote the confidence estimate associated with the answer. A standard definition of probabilistic calibration requires

$$\mathbb{E}[Y \mid C] = C \qquad \text{almost surely.}$$

For discrete confidence values, this condition can be expressed as

$$\mathbb{P}(Y = 1 \mid C = c) = c$$

for every value c occurring with positive probability. Thus, among answers assigned a confidence of 0.8, approximately 80% should satisfy the specified correctness criterion, subject to sampling variability and the stability of the evaluation distribution. The definition is conditional upon several substantive choices. One must specify the task distribution, define the correctness criterion, and establish the relationship between observed evaluation labels and actual correctness. A calibration claim is therefore not a property of a numerical output considered in isolation.

It is a property of a joint distribution involving confidence estimates, answers, tasks, and evaluation outcomes.

8.2 Calibration and discrimination

The podcast correctly emphasizes that calibration alone is insufficient for informative forecasting. Suppose the prevalence of correct answers in a task distribution is

$$\pi = \mathbb{P}(Y = 1).$$

Consider the constant confidence estimator

$$C = \pi.$$

This estimator is calibrated because

$$\mathbb{E}[Y \mid C] = \mathbb{E}[Y] = \pi = C.$$

Nevertheless, it assigns the same confidence to every answer and consequently provides no discrimination between individual cases. In the terminology of probabilistic forecast evaluation, calibration must be distinguished from resolution or discriminatory capacity. The distinction can be expressed through the decomposition of the Brier score (Brier, 1950; Murphy, 1973). Define

$$\mathrm{BS} = \mathbb{E}[(C - Y)^2].$$

Let

$$q(C) = \mathbb{E}[Y \mid C].$$

By conditioning on C,

$$\mathbb{E}[(C - Y)^2 \mid C] = (C - q(C))^2 + \mathrm{Var}(Y \mid C).$$

Taking expectations yields

$$\mathrm{BS} = \mathbb{E}[(C - q(C))^2] + \mathbb{E}[\mathrm{Var}(Y \mid C)].$$

The law of total variance gives

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid C)] + \mathrm{Var}(q(C)).$$

Consequently,

$$\mathrm{BS} = \underbrace{\mathbb{E}[(C - q(C))^2]}_{\text{calibration error}} - \underbrace{\mathrm{Var}(q(C))}_{\text{resolution}} + \underbrace{\mathrm{Var}(Y)}_{\text{outcome uncertainty}}.$$

The constant estimator $C = \pi$ has zero calibration error and zero resolution. An informative calibrated estimator possesses positive resolution, allowing it to distinguish subsets of cases having different observed frequencies of correctness. For a perfectly calibrated estimator,

$$q(C) = C,$$

and therefore

$$\mathrm{BS} = \pi(1 - \pi) - \mathrm{Var}(C).$$

The improvement over the constant base-rate estimator is exactly $\mathrm{Var}(C)$.

This formalizes the podcast's example of a weather forecaster who always predicts a probability of rainfall equal to the historical frequency. The forecaster may possess perfect calibration while providing no case-specific discrimination. The example should not be interpreted as establishing that base-rate forecasts are entirely useless. A base rate can provide valuable information when no more discriminating evidence is available. The limitation is that the forecast does not improve upon that base rate by distinguishing individual cases.

8.3 Calibration and the evidentiary record

Calibration is an empirical relationship between predictions and outcomes, and as such it cannot ordinarily be established by examining a single prediction. Neither a correct answer assigned a confidence of 0.1 nor an incorrect answer assigned a confidence of 0.9 independently establishes that the system is respectively underconfident or overconfident. The interpretation requires a record of comparable predictions and appropriately established outcomes. For a collection of n cases, one may record

$$\mathcal{D} = \{(X_i, A_i, C_i, Y_i)\}_{i=1}^{n}.$$

A simple empirical calibration procedure partitions confidence values into intervals

$$B_1, \ldots, B_K.$$

For each interval, define the index set

$$I_k = \{i : C_i \in B_k\}$$

and the corresponding sample size

$$n_k = |I_k|.$$

For nonempty intervals, define

$$\overline{C}_k = \frac{1}{n_k} \sum_{i \in I_k} C_i$$

and

$$\overline{Y}_k = \frac{1}{n_k} \sum_{i \in I_k} Y_i.$$

A reliability diagram compares $\overline{C}_k$ with $\overline{Y}_k$ across the confidence intervals. A commonly used empirical summary is the expected calibration error,

$$\widehat{\mathrm{ECE}} = \sum_{k : n_k > 0} \frac{n_k}{n} \left| \overline{C}_k - \overline{Y}_k \right|.$$

This statistic can be useful, but its interpretation depends upon the binning procedure, sample size, task distribution, and quality of the outcome labels.

Different binning procedures can produce different empirical calibration errors from the same underlying record. A small aggregate calibration error may also conceal substantial miscalibration within particular subpopulations. The existence of a logged record is therefore necessary for many empirical calibration claims, but the record must itself satisfy appropriate methodological requirements. The importance of such a record

is one of the strongest points made in the podcast.

8.4 Verbalized confidence and token generation

The podcast argues that a language model produces a numerical confidence statement through the same autoregressive mechanism responsible for its other linguistic outputs. For an autoregressive language model with parameters θ, the probability assigned to a token sequence $t_{1:m}$, conditional upon an input x, can be expressed as

$$P_\theta(t_{1:m} \mid x) = \prod_{j=1}^{m} P_\theta(t_j \mid x, t_{1:j-1}).$$

A verbalized confidence estimate is obtained by interpreting some portion of the generated token sequence as a numerical value. Let

$$C = g(t_{1:m})$$

denote this interpretation. The fact that C is generated through an autoregressive procedure establishes neither that it is calibrated nor that calibration is impossible. The same autoregressive mechanism can produce a correct arithmetic result, an accurate transcription, a valid formal expression, or an erroneous assertion. The semantic and epistemic properties of the output cannot be deduced solely from the fact that it is represented as a sequence of tokens.

The relevant question is whether the computational process producing the sequence implements, directly or approximately, an informative estimation procedure. The distinction between the numerical representation of a result and the process responsible for that result must be applied consistently. A number is not rendered epistemically legitimate merely because it was produced by a conventional statistical program. Nor is it rendered epistemically illegitimate merely because it was expressed through a language model. Its legitimacy depends upon the relationship between the generating procedure, the intended interpretation, and the relevant evidence.

8.5 Internal representations and information about correctness

Let Z denote an internal representation available within a computational system during the production of an answer. Let W denote the externally observable features upon which an external estimator relies. Whether Z contains information about the correctness variable Y beyond these features can be expressed through the conditional mutual

information

$$I(Y; Z \mid W).$$

A positive value would indicate that the internal representation contains information about correctness not already exhausted by the observable features. One conditions upon W and not upon the complete pair (X, A) because, when the correctness criterion is a function of the task and answer, Y is determined by (X, A) and $I(Y; Z \mid X, A)$ vanishes identically. The autoregressive factorization of the output distribution does not establish that this quantity must vanish. Whether useful information about correctness is available in internal representations, output probabilities, generated self-evaluations, or other signals is an empirical question. It depends upon the model, the training procedure, the representation being examined, the elicitation method, and the task distribution. Even if useful information is present, it does not follow that an ordinary request for a confidence percentage successfully extracts that information.

The presence of a signal and the validity of a procedure for eliciting that signal are different questions.

8.6 Empirical findings on language-model confidence

Existing research provides grounds for rejecting both the assumption that verbalized confidence is automatically reliable and the categorical claim that language models cannot produce informative estimates of their own correctness. Kadavath et al. (2022) investigate the ability of language models to evaluate the correctness of proposed answers and to predict whether they know the answers to particular questions. Their experiments include probability-based self-evaluation procedures that exhibit useful calibration and discrimination under particular experimental conditions. These results do not establish that an arbitrary language model, prompted in an arbitrary manner, will produce a calibrated confidence percentage. They do establish that the relationship between language-model computation and correctness estimation cannot be dismissed solely by reference to the autoregressive character of generation. Lin et al. (2022) examine the possibility of training language models to express uncertainty through natural language and numerical confidence statements.

Their results demonstrate that verbalized uncertainty can be made informative under specified training and evaluation conditions. The relevant distinction is between an unconstrained linguistic expression of confidence and a linguistic output whose statistical relationship to correctness has been deliberately trained and evaluated. Conversely, Xiong et al. (2024) investigate methods for eliciting confidence from language models and report substantial limitations, including overconfidence in verbalized estimates under a variety of experimental conditions. Their findings reinforce the need to distinguish different

elicitation procedures and to evaluate confidence estimates empirically rather than assuming their reliability. Taken together, these studies support a conditional conclusion. A confidence estimate produced by a language model may be informative, miscalibrated, or largely uninformative depending upon the system and evaluation procedure. Neither universal reliability nor universal unreliability follows from the output format. The podcast's central engineering recommendation therefore remains valid: a verbalized score should not be granted evidentiary authority in the absence of an appropriate validation record.

The speaker qualifies his architectural account, saying that the number usually does not come from any gauge and that in most cases it summarizes no evidence. A stronger reading, according to which the generative mechanism itself precludes correctness estimation, would exceed those qualifications and is not established.

8.7 Token probabilities and answer probabilities

A further distinction concerns the relationship between the probability of generating a particular token sequence and the probability that the proposition expressed by that sequence is correct. Suppose several distinct token sequences express the same answer. Let $\mathcal{A}_p$ denote the collection of sequences expressing a proposition p. The probability of generating an expression of p is then

$$P_\theta(\mathcal{A}_p \mid x) = \sum_{t \in \mathcal{A}_p} P_\theta(t \mid x),$$

assuming the sequences correspond to mutually exclusive generation outcomes. The probability assigned to an individual sequence can be small because there are many equivalent ways to express the answer. Conversely, a highly probable sequence can express an incorrect proposition. The distribution over linguistic expressions is therefore not automatically equivalent to a distribution over the truth of their semantic contents. Nevertheless, information derived from generation probabilities, semantic agreement, or internal representations may contribute to an empirically validated correctness estimator.

The methodological requirement is to establish the relevant relationship rather than to assume either its presence or its absence.

8.8 Stochastic variation and reproducibility

The podcast proposes asking a language model the same question repeatedly and interpreting variation in its confidence estimates as evidence that the estimates are not produced by a genuine calculation. The proposed experiment can reveal sensitivity

to sampling or prompting. However, the inference from variation to the absence of a calculation does not follow. Many legitimate statistical procedures are stochastic. Consider a Monte Carlo estimator

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f(Z_i),$$

where the Z_i are independent draws from a specified distribution. Two executions of the procedure can produce different numerical results despite using the same distribution, the same sample size, and the same mathematical estimation rule. Under appropriate conditions,

$$\mathbb{E}[\widehat{\mu}_n] = \mu$$

and

$$\mathrm{Var}(\widehat{\mu}_n) = \frac{\mathrm{Var}(f(Z))}{n}.$$

The stochastic variation is a property of the estimation procedure and does not establish the absence of a rule. Conversely, a deterministic procedure can produce exactly the same incorrect result on every execution.

Proposition 8.1 (Reproducibility does not imply calibration)**.** *For every nondegenerate binary outcome distribution with*

$$0 < \mathbb{P}(Y = 1) < 1,$$

there exists a deterministic confidence estimator that is perfectly reproducible but not calibrated.

Proof. Let

$$\pi = \mathbb{P}(Y = 1).$$

Choose any constant $c_0 \in [0, 1]$ satisfying

$$c_0 \neq \pi.$$

Define the estimator

$$C(x) = c_0$$

for every input x. The estimator produces the same result on every execution and is therefore perfectly reproducible. However,

$$\mathbb{E}[Y \mid C = c_0] = \pi \neq c_0.$$

It is consequently not calibrated. □

Moreover, a stochastic procedure may be calibrated under an appropriate joint distribution. Exact reproducibility and statistical validity are therefore different properties. Fixing the random seed, using deterministic decoding, or caching a previous confidence estimate may improve reproducibility without improving calibration. The relevant evaluation must examine the distribution of outputs, their relationship to correctness, and their stability under changes that should not alter the intended epistemic task.

8.9 Prompt sensitivity and epistemic invariance

The podcast also proposes rephrasing a request for confidence and observing whether the numerical estimate changes. This experiment addresses a more substantive property: the invariance of an estimation procedure under changes that preserve the relevant information. Let p denote a prompt and let $\tau(p)$ denote a transformation of that prompt intended to preserve the epistemic task. If the transformation does not change the evidence, the target, the correctness criterion, or the reference class, then substantial changes in the confidence estimate may indicate undesirable sensitivity to linguistic framing. For a stochastic estimator, the appropriate comparison concerns the conditional distributions

$$P(C \mid p)$$

and

$$P(C \mid \tau(p)),$$

rather than the exact equality of individual sampled values. One may quantify the difference through a suitable divergence or distance between these distributions. However, the interpretation of such an experiment depends upon whether the prompt transformation actually preserves the epistemic task.

A request to reconsider an answer may induce additional computation, the examination of previously neglected assumptions, or the identification of an error. A resulting change in confidence need not be attributable solely to social pressure or linguistic imitation. The transformation may have changed the computational procedure even if it introduced no new external evidence. Prompt sensitivity is therefore a legitimate object of empirical investigation, but its interpretation requires a distinction between changes in evidence, changes in inference, and changes in linguistic presentation.

8.10 Confidence conditioned upon the produced answer

The podcast further observes that a model which has already produced a strongly worded answer may subsequently generate a correspondingly high confidence estimate. This raises a legitimate concern about the dependence of self-evaluation upon the linguistic presentation of the answer. Let A denote the semantic content of an answer and let W denote features of its wording, such as assertiveness or rhetorical certainty. An idealized correctness estimator would not change merely because the same proposition was expressed in a more confident tone, provided that the wording supplies no additional information relevant to correctness. One may therefore investigate whether

$$P(C \mid X, A, W = w_1)$$

differs systematically from

$$P(C \mid X, A, W = w_2)$$

under controlled transformations of the wording. A difference would provide evidence of sensitivity to the presentation of the answer. It would not, by itself, establish the complete absence of information about correctness.

An estimator can combine informative signals with undesirable dependencies. The engineering problem is to determine which signals contribute to predictive validity, which introduce systematic distortions, and which remain stable under relevant changes in deployment conditions.

8.11 Tool-computed estimates and numerical provenance

The podcast explicitly distinguishes a confidence estimate generated without an established estimation procedure from a numerical result computed by an external tool and communicated through a language model. This distinction is important. Suppose a computational tool implements a specified estimator

$$C = f(D; \theta),$$

where D denotes the input data and θ denotes the relevant model parameters. A language model may subsequently communicate the result of this calculation. In such an architecture, the statistical validity of the numerical estimate depends primarily upon the estimator, its assumptions, its inputs, and its validation record. The linguistic interface introduces a separate requirement: the communicated number must faithfully represent the result returned by the tool. This separation can improve evidentiary provenance by making the numerical calculation independently inspectable. Nevertheless, the external

location of the calculation does not establish its validity.

An external tool can implement an inadequate model, receive defective data, use an inappropriate reference class, or produce a miscalibrated estimate. The architectural separation is useful because it permits the estimation procedure to be examined independently of the linguistic output. It is not a substitute for that examination.

8.12 Marginal calibration and case-specific evidence

Even a perfectly calibrated confidence estimator may fail to preserve distinctions that are relevant to particular cases. Let G denote a subpopulation or task characteristic. Marginal calibration requires

$$\mathbb{E}[Y \mid C] = C.$$

A stronger conditional requirement would be

$$\mathbb{E}[Y \mid C, G] = C.$$

The first condition does not imply the second. Consider two equally frequent groups, $G = 0$ and $G = 1$, for which every answer receives confidence

$$C = 0.8.$$

Suppose

$$\mathbb{P}(Y = 1 \mid G = 0) = 0.6$$

and

$$\mathbb{P}(Y = 1 \mid G = 1) = 1.$$

Then

$$\mathbb{P}(Y = 1 \mid C = 0.8) = \frac{0.6 + 1}{2} = 0.8.$$

The estimator is marginally calibrated. However, it is overconfident in the first group and underconfident in the second. A calibration claim can therefore conceal systematic differences among task categories, populations, mechanisms, or operating conditions. The choice of conditioning variables is a substantive component of the evaluation. Methods for evaluating and improving calibration across multiple subpopulations address related concerns (Hébert-Johnson et al., 2018). However, no finite evaluation can automatically establish calibration across every possible distinction relevant to future deployment.

The preservation of task-specific evidence remains important even when aggregate statistical reliability has been demonstrated.

8.13 From calibration to certification

Calibration describes a statistical relationship between estimates and outcomes. Certification concerns whether a specified object, claim, or proposed action satisfies particular requirements. Although the two operations can interact, they are not interchangeable. A calibrated estimate of 0.95 establishes, under its validated interpretation, an expected frequency of correctness among comparable cases. It does not establish which particular cases belong to the remaining 5%. Nor does it establish that the correctness criterion used in the calibration procedure exhausts the requirements governing a particular action. The distinction becomes operationally significant when confidence is used as a routing mechanism. A high score may justify reducing the expected amount of additional inspection under specified conditions without thereby justifying eliminating every independent check. The adequacy of the routing policy must itself be evaluated, including its effects upon the evidentiary record from which future confidence estimates are constructed.

9 The Externalization of Verification

The constructive proposal developed in the podcast consists in separating language generation from the estimation of reliability. Rather than asking a language model to assign an unvalidated confidence score to its own answer, the proposed architecture constructs a statistical estimation procedure from observable features, historical outcomes, and a continuously maintained record of performance. This proposal deserves serious consideration. It addresses several genuine defects of confidence-based autonomous systems and provides a more defensible starting point than the unrestricted use of verbalized self-evaluation. However, externalizing the confidence calculation does not automatically externalize the epistemic limitations of the generative system. The proposed architecture remains dependent upon the validity of its observable indicators, the adequacy of its correctness labels, the representativeness of its verification record, and the stability of the relationships learned from that record.

The following analysis examines these dependencies using the customer-support example developed in the episode.

9.1 Reconstruction of the proposed procedure

The podcast considers a customer-support agent that retrieves relevant help-centre documents and generates answers to customer questions. The objective is to construct a confidence estimate that determines whether a generated answer can be communicated automatically or must first be examined by a human reviewer. The proposed procedure

begins by identifying observable features of each answer. These features include whether the retrieval system found a relevant document, how many documents were retrieved, whether similar questions have previously been answered, and whether repeated executions of the language model produce agreeing answers. The procedure then introduces a historical base rate. For example, if 150 of 200 previously examined answers were judged correct, the initial estimate of reliability is

$$\widehat{\pi} = \frac{150}{200} = 0.75.$$

Subsequent cases are recorded together with their observable features, assigned confidence estimates, and observed outcomes.

As the record grows, cases possessing similar features are grouped, and their observed frequencies of correctness are used to estimate the reliability of future answers possessing those features. The podcast suggests, for example, that answers supported by a matching document and agreement between repeated generations might receive a confidence estimate of 0.95 if previous answers satisfying those conditions were correct in approximately 95% of examined cases. Finally, the relationship between predicted confidence and observed correctness is periodically reassessed. The proposal therefore contains several methodologically valuable components: observable predictors, historical base rates, recorded outcomes, empirical estimation, and continuing calibration assessment. Its principal advantage is that the numerical estimate can be examined independently of the language model's verbal expression of certainty. The analysis below, far from rejecting this architecture, identifies the additional conditions required for its evidentiary interpretation.

9.2 Observable indicators and latent correctness

Let

$$X = (X_1, \ldots, X_m)$$

denote the observable features associated with a generated answer. These might include retrieval scores, document counts, answer agreement, historical task frequencies, and other properties available to the surrounding computational system. Let Y denote actual correctness under a specified evaluation criterion. The proposed confidence estimator seeks to approximate

$$q(x) = \mathbb{P}(Y = 1 \mid X = x).$$

The estimator may be constructed using empirical frequency tables, logistic regression, probabilistic networks, nonparametric methods, or other statistical procedures. Its success depends upon the extent to which the selected observable features provide stable

information about correctness. The external calculation does not establish that this relationship exists. For example, a retrieval system may return several documents containing the same obsolete instruction. A language model may generate two agreeing answers because both are derived from the same erroneous document. A question may resemble many previously answered questions while differing in a single condition that changes the correct answer.

The observable features can therefore be accurate descriptions of the computational process without constituting sufficient evidence for the correctness of its result. The distinction is analogous to the spray-painted tree discussed earlier. An observation can accurately represent a feature while failing to establish the causal or semantic relationship through which that feature acquires evidentiary significance.

9.3 Document retrieval and semantic support

The existence of a retrieved document must be distinguished from the evidentiary relationship between that document and the generated answer. Let D denote a retrieved document and let a denote a proposition expressed by the answer. Several distinct conditions may be relevant. The document may concern the same general topic as a, or contain a statement that appears to support a, or support a only under conditions omitted from the generated answer. It may also have been superseded by a more recent document, or its authority may be restricted to a particular product, jurisdiction, or customer category. The retrieval event

$$\text{Retrieved}(D) = 1$$

does not establish the entailment

$$D \models a.$$

Even the latter expression, where an appropriate formal interpretation is available, does not establish that the document is authoritative, current, or empirically correct. A more expressive verification procedure must preserve distinctions among topical relevance, semantic support, document provenance, temporal validity, and institutional authority.

The number of retrieved documents is particularly susceptible to misleading interpretation. Several documents may be derivative copies of a single source, in which case their agreement need not constitute independent corroboration. Conversely, a single authoritative and current document may be sufficient to establish a particular institutional fact. The evidentiary value of retrieval cannot therefore be inferred from document count alone.

9.4 Agreement and common-mode error

The podcast proposes repeated generation as an observable indicator of reliability. Agreement between repeated answers can provide useful information about the stability of a generative procedure. However, stability must be distinguished from correctness. Suppose two executions of a language model produce answers A_1 and A_2. Define their agreement indicator

$$G = \mathbf{1}\{A_1 = A_2\}.$$

For simplicity, suppose the task possesses a unique correct answer a^*. Let

$$Y_i = \mathbf{1}\{A_i = a^*\}, \qquad i \in \{1, 2\}.$$

Agreement can occur when both answers are correct, but it can also occur when both answers express the same error. Consequently,

$$\begin{aligned}\mathbb{P}(G = 1) = \mathbb{P}(A_1 = A_2 = a^*) \\ + \mathbb{P}(A_1 = A_2 \neq a^*).\end{aligned}$$

The evidentiary significance of agreement depends upon the relative magnitudes of these two terms. The second term can be substantial when the executions share training data, model parameters, retrieved documents, prompt assumptions, or systematic reasoning defects. Agreement between repeated executions of the same model is therefore not equivalent to agreement between independent evidentiary procedures.

9.5 An explicit agreement model

Consider a simplified binary-answer task in which a system produces the correct answer with probability p. Suppose two executions are independent conditional upon the task and possess the same probability of correctness. Since the answer space contains exactly two alternatives, the two executions agree precisely when both are correct or both are incorrect. Thus,

$$\mathbb{P}(G = 1) = p^2 + (1 - p)^2.$$

The conditional probability of correctness given agreement is

$$\mathbb{P}(Y_1 = 1 \mid G = 1) = \frac{p^2}{p^2 + (1 - p)^2}.$$

When $p > 1/2$, agreement improves the probability of correctness relative to the individual execution. When $p < 1/2$, agreement instead reinforces the more frequently produced

incorrect answer. The improvement therefore depends upon assumptions about the underlying reliability of the generative procedure. Now suppose the two executions possess correlated correctness indicators with

$$\mathbb{P}(Y_i = 1) = p$$

and

$$\mathrm{Corr}(Y_1, Y_2) = \rho,$$

where $0 < p < 1$ and ρ is within the feasible range for the specified Bernoulli marginals, with $\mathbb{P}(G = 1) > 0$. Then

$$\mathbb{P}(Y_1 = 1, Y_2 = 1) = p^2 + \rho p(1 - p)$$

and

$$\mathbb{P}(Y_1 = 0, Y_2 = 0) = (1 - p)^2 + \rho p(1 - p).$$

The probability of correctness conditional upon agreement becomes

$$\mathbb{P}(Y_1 = 1 \mid G = 1) = \frac{p^2 + \rho p(1 - p)}{p^2 + (1 - p)^2 + 2\rho p(1 - p)}.$$

For $\rho = 0$, this reduces to the independent-execution formula.

For $\rho = 1$, it reduces to

$$\mathbb{P}(Y_1 = 1 \mid G = 1) = p.$$

Perfectly correlated executions provide no improvement over the original reliability. The calculation concerns a deliberately restricted binary-answer model. More general answer spaces require explicit treatment of the distribution of incorrect answers and the probability that distinct errors produce semantically equivalent outputs. Nevertheless, the example establishes the relevant engineering principle: agreement does not possess a fixed evidentiary value independently of the dependence structure responsible for it.

9.6 The observational equivalence of agreement records

The limitations of agreement can be expressed without assuming a particular probabilistic model. Suppose a collection of computational systems returns the same answer to a particular question. An observer who possesses only their answers and agreement relations may be unable to distinguish a case of independent successful inference from a case of shared systematic error.

Proposition 9.1 (Agreement-only obstruction)**.** *Let O denote an observational procedure*

that records generated answers and their agreement relations but does not independently observe the relevant truth conditions.

If two admissible underlying states s_1 and s_2 satisfy

$$O(s_1) = O(s_2)$$

while the common answer is correct in s_1 and incorrect in s_2, then correctness is not identifiable from the agreement record alone.

Proof. Let $Y(s)$ indicate whether the common answer is correct in state s. Since $O(s_1) = O(s_2)$ while $Y(s_1) \neq Y(s_2)$, the claim is the instance of Corollary 5.1 with observation map O and target Y. □

The proposition does not imply that agreement is empirically uninformative. What it establishes is that agreement cannot constitute an unconditional certificate of correctness. Its value must be established relative to the mechanisms generating the answers and the independent evidence available for evaluating them.

9.7 Outcome labels and the substitution of proxies

The proposed customer-support architecture requires a record of whether each generated answer was correct. The podcast suggests that relevant outcomes might include approval by a human reviewer or the absence of a subsequent question from the customer. These observations can be useful operational indicators, although they do not necessarily measure the same property as correctness. Let Y denote actual correctness, let R denote reviewer approval, and let N denote the absence of a subsequent customer inquiry. In general,

$$R \neq Y$$

and

$$N \neq Y.$$

A reviewer may approve an incorrect answer because the error is subtle, because the relevant document is obsolete, or because the reviewer lacks the information required to detect the discrepancy. A customer may not ask another question because the answer was satisfactory. The customer may also have abandoned the inquiry, accepted an incorrect answer, sought assistance elsewhere, or concluded that further communication would be unproductive.

The absence of a subsequent inquiry is therefore an observation about customer behaviour. Its relationship to answer correctness requires additional validation. The

substitution of approval or silence for correctness can produce an apparently well-calibrated system whose calibration concerns the wrong target.

9.8 Calibration against imperfect labels

Let $\widetilde{Y}$ denote the recorded evaluation label, which may differ from actual correctness Y. Suppose a confidence estimator satisfies

$$\mathbb{E}[\widetilde{Y} \mid C] = C.$$

This establishes calibration with respect to the recorded label. It does not establish

$$\mathbb{E}[Y \mid C] = C.$$

To examine the relationship, define the label sensitivity

$$s(c) = \mathbb{P}(\widetilde{Y} = 1 \mid Y = 1, C = c)$$

and the label specificity

$$t(c) = \mathbb{P}(\widetilde{Y} = 0 \mid Y = 0, C = c).$$

Let

$$q(c) = \mathbb{P}(Y = 1 \mid C = c)$$

denote the actual calibration function. The observed positive-label frequency is

$$\begin{aligned}\widetilde{q}(c) &= \mathbb{P}(\widetilde{Y} = 1 \mid C = c) \\ &= s(c)q(c) + (1 - t(c))(1 - q(c)).\end{aligned}$$

Equivalently,

$$\widetilde{q}(c) = \Big(s(c) + t(c) - 1\Big)q(c) + 1 - t(c).$$

If the label sensitivity and specificity are known and satisfy

$$s(c) + t(c) \neq 1,$$

then one may recover

$$q(c) = \frac{\widetilde{q}(c) + t(c) - 1}{s(c) + t(c) - 1}.$$

However, if the label error properties are unknown, the observed positive-label frequency generally does not identify the actual correctness frequency. A system can therefore

become accurately calibrated against reviewer approval while remaining miscalibrated against actual correctness. The same concern applies to customer satisfaction, absence of complaints, successful execution, and other operational proxies. These outcomes may be valuable in their own right.

The error consists in silently substituting them for the distinct property that the confidence estimate purports to represent.

9.9 Non-identifiability under unknown label error

The preceding observation admits a direct identifiability result.

Proposition 9.2 (Proxy-calibration obstruction)**.** *Suppose the only observed evaluation quantity at confidence level c is*

$$\widetilde{q}(c) = \mathbb{P}(\widetilde{Y} = 1 \mid C = c),$$

and no restrictions are imposed upon the conditional relationship between $\widetilde{Y}$ and Y. Then the actual correctness frequency

$$q(c) = \mathbb{P}(Y = 1 \mid C = c)$$

is not identified by $\widetilde{q}(c)$.

Proof. Fix any observed frequency

$$\widetilde{q}(c) = r$$

with $r \in [0, 1]$. For any candidate actual correctness frequency $q \in [0, 1]$, construct a conditional joint distribution under which

$$Y \sim \text{Bernoulli}(q)$$

and

$$\widetilde{Y} \sim \text{Bernoulli}(r),$$

with Y and $\widetilde{Y}$ conditionally independent given $C = c$. Every such distribution produces the same observed label frequency r, while the actual correctness frequency can take any value $q \in [0, 1]$.

Consequently, the observed proxy frequency does not identify the target frequency without additional assumptions. □

The proposition uses an unrestricted label-error model to expose the structural limitation. In practical applications, substantive knowledge about the evaluation procedure may exclude many of the constructed possibilities. The resulting restrictions must nevertheless be established rather than silently assumed. A larger record of proxy outcomes can improve estimation of the proxy frequency without independently improving knowledge of its relationship to actual correctness.

9.10 Selective verification

A further difficulty arises from the interaction between confidence estimation and the routing policy that the estimate is intended to control. Suppose low-confidence answers are sent to human reviewers, while high-confidence answers are communicated automatically. Let

$$V \in \{0, 1\}$$

indicate whether an answer receives independent verification. The resulting record may contain reliable correctness labels only for cases satisfying $V = 1$. The empirical quantity available from that record is

$$\mathbb{P}(Y = 1 \mid C = c, V = 1).$$

The quantity required for ordinary calibration is

$$\mathbb{P}(Y = 1 \mid C = c).$$

These quantities need not coincide. The difference is a form of selective verification or verification bias, related to the selective labels problem in statistical decision systems (Begg and Greenes, 1983; Lakkaraju et al., 2017). The difficulty is particularly important because the verification policy is not external to the confidence system. The confidence estimate influences whether verification occurs, and the resulting verification record is subsequently used to update the confidence estimator.

The estimator therefore participates in determining which evidence will become available for its own future evaluation.

9.11 A decomposition of selective verification bias

Define

$$e(c) = \mathbb{P}(V = 1 \mid C = c)$$

as the probability of verification at confidence level c. Let

$$q_1(c) = \mathbb{P}(Y = 1 \mid C = c, V = 1)$$

and

$$q_0(c) = \mathbb{P}(Y = 1 \mid C = c, V = 0).$$

The law of total probability gives

$$q(c) = e(c)q_1(c) + (1 - e(c))q_0(c),$$

where

$$q(c) = \mathbb{P}(Y = 1 \mid C = c).$$

When both conditional frequencies are defined,

$$q_1(c) - q(c) = (1 - e(c))\Big(q_1(c) - q_0(c)\Big).$$

The observed verification frequency equals the target frequency if all cases are verified, or if verified and unverified cases have the same conditional correctness frequency. Neither condition is guaranteed by the existence of a verification record. If reviewers preferentially examine unusual or difficult cases within a confidence interval, the verified cases may be less reliable than the unverified cases. Conversely, if verification is concentrated upon cases that are easy to adjudicate, the observed record may overstate reliability. The direction of the bias thus depends upon the selection mechanism and cannot be determined from the existence of missing verification labels alone.

9.12 Deterministic thresholds and the absence of overlap

Consider the routing policy

$$V = \mathbf{1}\{C < \tau\},$$

where τ is a specified confidence threshold. Every answer below the threshold is verified, while no answer at or above the threshold is verified. For confidence levels below τ,

$$e(c) = 1.$$

For confidence levels at or above τ,

$$e(c) = 0.$$

This example exposes a limitation stronger than ordinary selection bias. Above the threshold, the system possesses no independent verification outcomes at all.

Proposition 9.3 (Unverified-region obstruction)**.** *Suppose a verification policy satisfies*

$$\mathbb{P}(V = 1 \mid C \in B) = 0$$

for a measurable confidence region B having positive probability. Assume that correctness labels are observed only when $V = 1$, and that no additional restrictions determine the correctness distribution within B.

Then the calibration function on B is not identified by the verification record.

Proof. Consider two joint distributions that agree on the distribution of C, the verification policy, and the conditional distribution of Y outside B.

Within B, assign different conditional correctness distributions to the two models. Because no case in B is verified, the differing correctness labels are never observed. The two models therefore induce the same distribution of observable verification records while possessing different calibration functions within B.

The calibration function on B is consequently not identifiable from the observed record. □

The result holds even if the system maintains a complete record of every confidence estimate and every routing decision. Logging an unverified answer does not establish its correctness. The obstruction concerns the absence of the relevant outcome observation, not the absence of a database entry.

9.13 Partial identification under incomplete verification

Even when unverified correctness outcomes are unavailable, the observed verification record may supply bounds upon the actual correctness frequency. For confidence levels satisfying $e(c) > 0$, the decomposition

$$q(c) = e(c)q_1(c) + (1 - e(c))q_0(c)$$

and the restriction

$$0 \leq q_0(c) \leq 1$$

imply

$$e(c)q_1(c) \leq q(c) \leq e(c)q_1(c) + 1 - e(c).$$

The width of this identified interval is

$$1 - e(c).$$

If the verification probability is small, the interval can remain wide even when the correctness frequency among verified cases is estimated precisely. For example, suppose that only 10% of answers in a confidence interval receive verification, and that 95% of the verified answers are correct. Without further assumptions concerning selection, the actual correctness frequency satisfies

$$0.095 \leq q(c) \leq 0.995.$$

These bounds are deliberately conservative. They do not imply that the actual correctness frequency is likely to be near the lower bound. They establish that the verification record alone does not exclude such a possibility. A precise estimate of the observed success rate cannot compensate for the absence of information about the unverified population.

When $e(c) = 0$, the corresponding unrestricted identified set is simply

$$q(c) \in [0, 1].$$

The confidence estimate may still be supported by other evidence, but the verification record supplies no direct calibration evidence for that region.

9.14 Randomized audits and the positivity requirement

One remedy for selective verification is to introduce randomized audits across the full range of confidence estimates. Let

$$e(x, c) = \mathbb{P}(V = 1 \mid X = x, C = c)$$

denote the verification propensity. Suppose the audit design ensures

$$e(x, c) > 0$$

throughout the relevant deployment domain. Suppose further that verification is assigned independently of the unobserved correctness outcome conditional upon the recorded variables:

$$V \perp Y \mid X, C.$$

Then the verified cases can be used to estimate population-level correctness through appropriate weighting.

Proposition 9.4 (Identification through randomized verification)**.** *Assume that Y is observed accurately whenever $V = 1$, that*

$$V \perp Y \mid X, C,$$

and that

$$e(X, C) > 0$$

almost surely. Then

$$\mathbb{E}\left[\frac{VY}{e(X,C)} \,\middle|\, C\right] = \mathbb{E}[Y \mid C].$$

Proof. By iterated conditional expectation,

$$\mathbb{E}\left[\frac{VY}{e(X,C)} \,\middle|\, X, C, Y\right] = \frac{Y\mathbb{E}[V \mid X, C, Y]}{e(X,C)}.$$

Conditional independence implies

$$\mathbb{E}[V \mid X, C, Y] = e(X, C).$$

Consequently,

$$\mathbb{E}\left[\frac{VY}{e(X,C)} \,\middle|\, X, C, Y\right] = Y.$$

Conditioning again upon C establishes

$$\mathbb{E}\left[\frac{VY}{e(X,C)} \,\middle|\, C\right] = \mathbb{E}[Y \mid C].$$

□

This result is an application of inverse-probability weighting, related to classical sampling estimators (Horvitz and Thompson, 1952). For a confidence interval B, an empirical ratio estimator of its correctness frequency is

$$\widehat{q}_B = \frac{\sum_{i=1}^{n} \frac{V_i Y_i}{e_i} \mathbf{1}\{C_i \in B\}}{\sum_{i=1}^{n} \frac{V_i}{e_i} \mathbf{1}\{C_i \in B\}},$$

provided the denominator is positive, where

$$e_i = e(X_i, C_i).$$

The ratio estimator is not generally exactly unbiased in finite samples, although it can be consistent under appropriate sampling and regularity conditions. The verification propensities must be recorded or otherwise known. The positivity condition is essential, since no weighting procedure can recover outcomes from a region that possesses zero probability of verification without introducing additional identifying assumptions.

9.15 A practical mixed verification policy

A customer-support system may retain mandatory verification for low-confidence answers while introducing random audits for answers that would otherwise be communicated automatically. For example, consider

$$e(c) = \begin{cases} 1, & c < \tau, \\ \varepsilon(c), & c \geq \tau, \end{cases}$$

where

$$0 < \varepsilon(c) \leq 1.$$

The audit probability can vary across confidence intervals, task categories, and operating conditions. It may be increased for recently modified documents, unfamiliar questions, new model versions, or domains in which the consequences of an incorrect answer are particularly serious. However, the audit design must preserve a nonzero probability of examining cases that the system otherwise regards as routine. Otherwise, the verification record can become incapable of detecting systematic errors in precisely those regions where the system has become most confident. Randomized auditing is not a substitute for mandatory pre-execution verification when the consequences of an incorrect action require such verification. Its purpose in the present architecture is to maintain an informative evaluation record for operations that are already permitted to proceed without individual certification.

9.16 The cost of rare-event detection

The positivity requirement establishes identifiability under the specified sampling assumptions. It does not establish that the available audit sample is sufficiently large to detect rare but consequential failures. Suppose independent audits examine n cases from a specified population with actual failure probability r. The probability of observing no

failures is

$$(1-r)^n.$$

To obtain a probability of at least $1-\delta$ of observing at least one failure, one requires

$$(1-r)^n \leq \delta.$$

Equivalently,

$$n \geq \frac{\log \delta}{\log(1-r)}.$$

For small r,

$$\log(1-r) \approx -r,$$

giving the approximation

$$n \gtrsim \frac{\log(1/\delta)}{r}.$$

For example, detecting at least one failure with probability 0.95 when the underlying failure rate is 0.001 requires approximately three thousand independent audits. This calculation does not establish the severity or nature of the failures. It illustrates the sample requirements associated with detecting infrequent events. A small audit record containing no observed failures does not establish that the failure probability is zero.

Moreover, the independence assumption can be violated when failures occur in clusters associated with shared documents, software versions, or recurring task structures. The effective information supplied by an audit record may then be substantially smaller than its nominal sample size.

9.17 Small groups and estimation uncertainty

The podcast proposes grouping previous answers according to observable features and assigning future answers the historical success frequency of the corresponding group. This procedure can provide a reasonable initial estimator. However, the empirical frequency must be distinguished from the underlying success probability. Suppose a group contains n independently evaluated cases, of which k are correct. The empirical frequency is

$$\widehat{q} = \frac{k}{n}.$$

Under a homogeneous Bernoulli model with success probability q,

$$\operatorname{Var}(\widehat{q}) = \frac{q(1-q)}{n}.$$

The same empirical frequency can therefore possess substantially different statistical precision depending upon the number of observations. For example,

$$\frac{19}{20} = \frac{1900}{2000} = 0.95.$$

The two records produce the same point estimate, but they do not provide equivalent evidence concerning the underlying success probability. Using the Wilson interval (Wilson, 1927), an approximate 95% confidence interval for the first record is

$$[0.764, 0.991],$$

whereas the corresponding interval for the second record is approximately

$$[0.940, 0.959].$$

These intervals are conditional upon the specified sampling model.

They do not account for systematic label error, distributional change, or omitted causal mechanisms. Nevertheless, they illustrate why a confidence estimator constructed from empirical groups should retain information about the sample sizes and uncertainty associated with its estimated success frequencies. The reported probability of answer correctness and the uncertainty concerning the estimation of that probability are distinct statistical objects.

9.18 Smoothing and the role of prior assumptions

When groups contain few observations, one may regularize the empirical frequency through a prior distribution. Suppose

$$q \sim \mathrm{Beta}(\alpha, \beta)$$

and

$$K \mid q \sim \mathrm{Binomial}(n, q).$$

The posterior distribution is

$$q \mid K = k \sim \mathrm{Beta}(\alpha + k, \beta + n - k).$$

The posterior mean is

$$\mathbb{E}[q \mid K = k] = \frac{\alpha + k}{\alpha + \beta + n}.$$

This procedure can reduce the instability associated with very small empirical groups. However, the improvement depends upon the suitability of the prior distribution and the assumption that the group possesses a sufficiently stable success probability. The prior does not constitute additional independent observations of the particular group, but introduces information or regularization through the specified probabilistic model. The interpretation of the result should preserve this distinction. Hierarchical models can similarly borrow information across related groups. Their usefulness depends upon the adequacy of the relationships assumed among those groups. An estimator should not treat statistical smoothing as evidence that distinctions between groups are irrelevant to correctness.

9.19 Changes in the help centre and the generative system

The podcast suggests periodically updating the confidence rule from newly recorded outcomes. This recommendation is appropriate, but the interpretation of historical success frequencies requires attention to changes in the system being evaluated. A customer-support environment can change through revisions to help-centre documents, modifications to institutional policies, changes in customer populations, updates to retrieval algorithms, and replacement or modification of the language model. Let t denote the time or operating epoch associated with a case. The relevant conditional correctness function may then be written

$$q_t(x) = \mathbb{P}_t(Y = 1 \mid X = x).$$

A historical estimator constructed from cases at epoch t_0 approximates

$$q_{t_0}(x).$$

Its applicability at epoch t_1 depends upon the relationship between

$$q_{t_0}(x)$$

and

$$q_{t_1}(x).$$

The equality

$$q_{t_0}(x) = q_{t_1}(x)$$

is a stationarity assumption and not a consequence of the law of large numbers. An arbitrarily large historical record can estimate an obsolete success frequency with arbitrarily high precision.

The statistical precision of that historical estimate does not establish its relevance to the current system.

9.20 The law of large numbers and historical relevance

The podcast appeals to the law of large numbers when describing the improvement of the confidence rule as the record grows. The appeal is appropriate under suitable conditions. However, the law of large numbers does not establish that an arbitrary accumulation of historical cases converges to the current correctness probability. Consider independent binary outcomes

$$Y_i \sim \text{Bernoulli}(q_i),$$

where the success probability q_i may vary between cases. The empirical mean is

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Its expectation is

$$\mathbb{E}[\overline{Y}_n] = \frac{1}{n}\sum_{i=1}^{n} q_i.$$

The empirical mean estimates the average of the historical success probabilities. It does not necessarily estimate the success probability q_{n+1} of the next case. For example, suppose a system possesses success probability 0.95 before a software update and success probability 0.60 after the update. A large historical record dominated by pre-update cases can continue to report a success frequency near 0.95 even though the current system performs substantially worse.

The addition of further historical observations does not repair this mismatch unless the evaluation procedure accounts for the change in the generating process. Methods for detecting and adapting to concept drift address related problems (Gama et al., 2014). For the present architecture, the practical implication is that calibration records should preserve the versions and operating conditions under which their observations were generated.

9.21 Versioned evidentiary records

A useful verification record should preserve more than a confidence score and a binary outcome. A case record may be represented schematically as

$$r_i = \Big(x_i, a_i, c_i, v_i, \widetilde{y}_i, e_i, m_i, d_i, t_i\Big),$$

where x_i denotes the task, a_i the answer, c_i the assigned confidence, v_i the verification indicator, $\widetilde{y}_i$ the recorded evaluation outcome when available, e_i the verification propensity, m_i the model and inference configuration, d_i the relevant document versions, and t_i the operating epoch. The record should also preserve the provenance of the evaluation outcome, including the procedure used to establish correctness and any unresolved disagreements among reviewers. Such a record permits the system to distinguish changes in the generative procedure from changes in the evaluation procedure. It also allows historical calibration claims to be associated with the specific configuration under which they were established. Without this information, a calibration record can silently aggregate outcomes from materially different systems.

The resulting success frequency may not describe any single deployed configuration.

9.22 Endogenous evidence and the feedback problem

The interaction between confidence estimation, routing, and verification can be represented as a feedback process. Let θ_t denote the parameters of the confidence estimator at epoch t. For a case x, the estimator produces

$$C_t = f_{\theta_t}(x).$$

A routing policy then determines verification according to

$$V_t \sim \text{Bernoulli}\Big(e_t(x, C_t)\Big).$$

The observed evaluation record $\mathcal{D}_t$ depends upon these verification decisions. The confidence estimator is subsequently updated through an operation

$$\theta_{t+1} = U(\theta_t, \mathcal{D}_t).$$

Consequently, the data used to evaluate and update the estimator are partly generated by the estimator's previous outputs. This is an endogenous evidentiary process. A confidence estimator that assigns high scores to a systematically incorrect class of answers may reduce the probability that those answers receive independent examination. The absence of detected errors can then be interpreted as evidence of continued reliability, even though it partly results from the verification policy.

The architecture can thereby develop a self-protecting region of epistemic invisibility. This phenomenon does not require deliberate deception, strategic behaviour, or an explicit objective of avoiding inspection. It can arise from the ordinary interaction of a confidence threshold with an incomplete evaluation record.

9.23 An illustrative self-confirming routing failure

Consider two task categories, $G = A$ and $G = B$. Suppose the confidence estimator assigns

$$C(A) = 0.75$$

and

$$C(B) = 0.98.$$

Assume the actual correctness frequencies are

$$q(A) = 0.75$$

and

$$q(B) = 0.60.$$

The system uses the verification threshold

$$\tau = 0.9.$$

All cases in category A receive independent verification, while cases in category B proceed automatically. The observed verification record correctly indicates that the confidence estimate for category A is calibrated. However, the record contains no independent correctness outcomes for category B. If the system evaluates its calibration only on verified cases, it may report excellent calibration. The reported result is compatible with a substantial systematic error in the automatically processed category. Increasing the number of verified cases in category A does not resolve the uncertainty concerning category B. The system requires either additional independent evidence concerning category B or a verification policy that samples it directly.

The example illustrates why calibration evaluation must include the population whose automatic processing the confidence estimator is intended to justify.

9.24 Evaluation independence and institutional incentives

The verification problem is not exhausted by the statistical properties of sampling. The organization operating the system may possess incentives to reduce the amount of human review, increase the proportion of automatically processed cases, or demonstrate improvements in measured reliability. These objectives can influence the selection of correctness proxies, the design of confidence thresholds, the treatment of ambiguous outcomes, and the allocation of auditing resources. A system evaluated primarily through reviewer approval may improve its measured performance by generating answers that

reviewers find plausible. A system evaluated through the absence of subsequent customer questions may improve its measured performance without improving the correctness of its answers. A system evaluated only through the outcomes of escalated cases may improve its reported calibration by changing which cases are escalated. These are different instances of the general problem of proxy optimization.

The operational indicator becomes a target of organizational or computational optimization, while its relationship to the underlying objective can deteriorate (Campbell, 1979). The externalization of measurement should therefore be accompanied by an appropriate degree of independence in the evaluation procedure. This independence need not require that every evaluation be conducted by an unrelated institution. It requires that the procedures responsible for detecting errors not be entirely controlled by the same signals and incentives whose reliability they are intended to establish.

9.25 Human review as a fallible evidentiary operation

The podcast assigns humans the responsibility for judging cases in which the consequences of an incorrect answer are particularly serious. This is a reasonable architectural distinction. However, the presence of a human reviewer does not automatically establish the correctness of the reviewed answer. Human reviewers can share the same defective documents, interpretive assumptions, or institutional incentives as the generative system. They can also be influenced by the confidence estimate displayed alongside the answer. For example, a reviewer informed that a system has assigned a confidence of 0.99 may approach the answer differently from a reviewer informed that the confidence is 0.40. If the confidence estimate influences the evaluation label, the record can contain a dependence between the predicted quantity and the procedure used to measure its supposed target. Where practicable, calibration audits should therefore consider whether reviewers ought to be blinded to the confidence score, the routing decision, or other information that could bias the evaluation.

Independent source consultation, explicit correctness criteria, documented disagreements, and additional adjudication procedures may be necessary for cases whose correctness cannot be established through ordinary review. The relevant distinction is not simply between human and machine judgment. It is between procedures that supply independent evidence and procedures that reproduce the assumptions or errors of the original answer.

9.26 Original answers and repaired answers

Verification can alter the object being evaluated, which is a difficulty distinct from the absence of labels. A reviewer who detects an error will ordinarily repair the answer before it reaches the customer. Let Y_0 indicate the correctness of the answer as generated, and let Y_1 indicate the correctness of the answer finally delivered. For binary indicators,

$$\begin{aligned}\mathbb{E}[Y_1 \mid C] - \mathbb{E}[Y_0 \mid C] = \mathbb{P}(Y_0 = 0, Y_1 = 1 \mid C) \\ - \mathbb{P}(Y_0 = 1, Y_1 = 0 \mid C).\end{aligned}$$

The first term records successful repairs and the second records damage introduced by review. The confidence estimate C was assigned to the generated answer, so that its calibration concerns $\mathbb{E}[Y_0 \mid C]$. An evaluation conducted against Y_1 credits the estimator with corrections made by the reviewer, and it does so even when every case is verified. The identification result for randomized verification must accordingly be applied to Y_0. An audit record should preserve the original answer, its original confidence estimate, the adjudication of its correctness, any repaired answer, and the separate evaluation of that repair.

9.27 Requirements for an external confidence architecture

The preceding analysis suggests a strengthened interpretation of the podcast's constructive proposal. The confidence estimator should possess a specified target whose interpretation is not silently replaced by a convenient operational proxy. Its observable features should be evaluated for their relationship to that target, including their dependence structure and susceptibility to common-mode error. Its evaluation record should contain appropriately established outcomes, verification propensities, and the versions of the systems and documents responsible for the recorded cases. Its verification policy should maintain adequate coverage of the deployment population, including randomized audits of cases that would otherwise escape independent examination. Its statistical estimates should retain information about finite-sample uncertainty, subpopulation differences, and the possibility of changes in the generating process. Finally, the confidence estimate should remain distinguishable from the operational decision that uses it.

Among these requirements, *the preservation of observability is an operational obligation*: a system should not be permitted to eliminate the observations needed to evaluate its reliability merely because its current estimates suggest that they are unnecessary.

These requirements preserve the useful division of labour proposed in the podcast while preventing the external estimator from acquiring an evidentiary authority that has not been established.

9.28 The question that follows calibration

The podcast concludes by proposing that a numerical confidence statement be interrogated through two questions: what the percentage refers to and against what outcomes it has been checked. These questions provide an appropriate starting point for the evaluation of numerical confidence. A satisfactory answer requires a specified target, a defensible evaluation procedure, a representative record, and an adequate statistical interpretation. The preceding sections have identified further questions concerning what remains unobservable, which possibilities have been excluded from the representation, and how the verification policy determines which errors become visible. Even if all of these questions receive satisfactory answers, one further question remains. What does the resulting evidence authorize the system to do? The transition from a validated statistical estimate to an authorized operation requires an additional architectural principle. A confidence estimate may contribute to the selection of an appropriate verification procedure.

It cannot independently establish that every obligation governing the proposed action has been satisfied. The following section develops this distinction through an obligation-based architecture for autonomous continuation.

10 Authorization Without Scalar Reduction

The preceding analysis has established that numerical confidence estimates can possess legitimate statistical interpretations without constituting comprehensive certificates of correctness, evidentiary sufficiency, or operational admissibility. This distinction becomes especially important when computational systems cease to function exclusively as sources of information and acquire the ability to modify external states. An informational error may become an operational error when an answer is used to authorize a financial transaction, modify a software repository, alter an institutional record, execute a physical process, or communicate a consequential instruction. The transition from assertion to action introduces requirements that cannot ordinarily be derived from the probability of assertion correctness alone. An autonomous system must distinguish what it believes to be true from what it has established, what it is permitted to do, and what consequences its proposed continuation may produce.

These distinctions motivate an obligation-based architecture in which confidence estimates contribute to the organization of verification but do not automatically determine the admissibility of action.

10.1 Correctness and permission

Consider a computational system that correctly identifies the command required to delete a file. The correctness of the command does not establish that the file should be deleted. The system may have correctly identified the relevant file, accurately described the effects of the command, and demonstrated complete technical competence concerning its execution. Nevertheless, the deletion may be unauthorized. Similarly, a system may correctly identify the procedure for transferring funds between accounts without possessing permission to initiate the transfer. The epistemic status of the proposed operation and its normative or institutional status are different properties. Let $Y(a, s)$ denote the correctness of an assertion associated with action a in state s. Let

$$A(a, s) \in \{0, 1\}$$

denote whether the action is authorized under the applicable institutional and operational rules. There is no general implication

$$Y(a, s) = 1 \quad \Longrightarrow \quad A(a, s) = 1.$$

Nor does authorization imply correctness.

An authorized individual can make an incorrect decision, just as an unauthorized individual can correctly identify an appropriate procedure. The two conditions must therefore be established through distinguishable operations. A confidence estimate concerning Y cannot independently establish A unless the applicable authorization rule explicitly makes the relevant statistical condition sufficient. Such a rule would itself require justification.

10.2 The multidimensional status of a proposed action

Let $\mathcal{S}$ denote a space of operational states and let $\mathcal{A}$ denote a space of candidate actions. For each pair (s, a), consider a status representation of the form

$$\sigma(s, a) = \Big(r(s, a), v(s, a), e(s, a), p(s, a), k(s, a), b(s, a)\Big),$$

where r denotes an empirically evaluated reliability estimate, v denotes logical or computational validity, e denotes evidentiary sufficiency, p denotes permission or authorization, k denotes compatibility with other obligations and operations, and b denotes the availability of recovery or reversal procedures. These components need not possess the same mathematical structure. The reliability component may be represented by a probability estimate accompanied by information about its validation domain. The validity compo-

nent may be represented by a proof object, a verification result, or an unresolved formal obligation. The evidentiary component may consist of records, observations, source dependencies, and an assessment of their adequacy. The authorization component may depend upon an institutional rule, a capability, an explicit instruction, or a documented grant of permission.

The compatibility component may require an examination of the proposed action in relation to concurrent operations and system-wide invariants. The recovery component may describe the conditions under which the effects of an action can be reversed, compensated for, or otherwise contained. The status representation is therefore not necessarily a vector in a common Euclidean space. It is more appropriately understood as an element of a heterogeneous product,

$$\Sigma = \Sigma_r \times \Sigma_v \times \Sigma_e \times \Sigma_p \times \Sigma_k \times \Sigma_b.$$

The purpose of the representation is to preserve distinctions relevant to the admissibility of continuation. Its components are not introduced for the purpose of constructing a weighted average.

10.3 Obligations and their satisfaction

An obligation is a condition that must be established before a specified continuation becomes admissible. Let

$$\mathcal{O}(s, a) = \{O_1, \ldots, O_m\}$$

denote the obligations applicable to a proposed action a in state s. The obligations may concern the identity of the affected object, the validity of a computation, the provenance of a document, the existence of permission, the preservation of an invariant, or the availability of a recovery procedure. Their satisfaction is assessed relative to the available evidence E. For each obligation O_i, introduce an evaluation function

$$\nu_i(s, a, E) \in \{\mathsf{pass}, \mathsf{fail}, \mathsf{unresolved}\}.$$

The distinction between failure and unresolved status is important. An obligation fails when the available evidence establishes that its condition is not satisfied, whereas it remains unresolved when the available evidence does not establish either its satisfaction or its violation. These conditions can require different responses. A failed authorization condition may require refusal.

An unresolved evidentiary condition may require additional observation. An unresolved computational obligation may require the execution of a verifier. An unresolved

compatibility condition may require coordination with another operation. Compressing these conditions into a single numerical penalty can conceal the specific remedial operation required.

Definition 10.1 (Obligation-based admissibility)**.** A proposed action a in state s is admissible relative to evidence E when every mandatory obligation applicable to the action has been established as satisfied:

$$\mathrm{Adm}(s, a, E) = 1$$

if and only if

$$\nu_i(s, a, E) = \mathsf{pass}$$

for every mandatory obligation $O_i \in \mathcal{O}(s, a)$.

This definition is deliberately conditional upon the adequacy of the obligation set and the soundness of its evaluation procedures. It does not establish that an arbitrary collection of formal checks exhausts every consideration relevant to an action. The specification of obligations remains a substantive architectural and institutional responsibility.

10.4 The noncompensatory character of mandatory conditions

A familiar decision architecture assigns numerical scores to several properties of an action and combines them through a weighted sum. Suppose, for example, that a system assigns scores

$$x = (x_1, \ldots, x_n) \in [0, 1]^n$$

and constructs

$$U(x) = \sum_{i=1}^{n} w_i x_i,$$

where

$$w_i > 0, \qquad \sum_{i=1}^{n} w_i = 1.$$

The system authorizes an action whenever

$$U(x) \geq \tau.$$

This architecture permits compensation between components, so that a reduction in one component can be offset by an increase in another, provided the aggregate score remains above the threshold. Such compensation may be appropriate for ordinary preferences. A traveller may accept a longer journey in exchange for a lower transportation cost. A

researcher may accept a slower algorithm in exchange for reduced memory consumption. Mandatory obligations possess a different structure. An absence of permission is not ordinarily compensated for by greater computational accuracy, nor is a violation of a conservation law repaired by an increase in the confidence assigned to a prediction.

An invalid proof is not made valid by the rhetorical clarity of its presentation. The relevant architectural distinction is between compensatory preferences and noncompensatory requirements.

Proposition 10.1 (Obstruction to compensatory authorization)**.** *Let $n \geq 2$ and suppose admissibility requires*

$$x_i \geq t_i \qquad \text{for every } i,$$

where

$$0 < t_i < 1.$$

There is no strictly positive weighted sum

$$U(x) = \sum_{i=1}^{n} w_i x_i$$

and threshold τ such that

$$U(x) \geq \tau$$

is equivalent, throughout $[0,1]^n$, to the satisfaction of every individual threshold.

Proof. Let

$$t = (t_1, \ldots, t_n).$$

Since t satisfies every individual threshold, equivalence requires

$$U(t) \geq \tau.$$

Choose distinct indices i and j. Since $t_i > 0$ and $t_j < 1$, one may choose $\varepsilon > 0$ sufficiently small that the vector x defined by

$$x_i = t_i - \varepsilon,$$

$$x_j = t_j + \frac{w_i}{w_j}\varepsilon,$$

and

$$x_k = t_k \qquad \text{for } k \notin \{i, j\}$$

belongs to $[0,1]^n$. By construction,

$$U(x) = U(t) \geq \tau.$$

However,

$$x_i < t_i,$$

so x violates a mandatory condition.

This contradicts the proposed equivalence. □

The proposition does not establish that admissibility cannot be encoded by any scalar-valued function. For example, the indicator of the admissible set is itself scalar-valued. A minimum-margin function can also represent a conjunction of numerical thresholds. The obstruction concerns compensatory weighted aggregation. The broader architectural concern is that a scalar representation of the final decision need not preserve the evidentiary structure responsible for that decision. A binary authorization flag may be sufficient for controlling execution while remaining insufficient for explaining why execution was authorized, which obligations were examined, or what must be repaired following a refusal. The decision and its supporting certificate must therefore remain distinguishable.

10.5 Confidence as a routing signal

A legitimate role for numerical confidence estimates remains available within an obligation-based architecture. Confidence can contribute to the selection of an appropriate verification procedure. Let

$$\mathcal{V} = \{v_1, \ldots, v_k\}$$

denote a collection of available verification procedures. These may include document retrieval, symbolic checking, independent computation, empirical measurement, human examination, or randomized auditing. A routing policy can be represented as

$$\rho : \mathcal{S} \times \mathcal{A} \times \Sigma \longrightarrow \mathcal{V}.$$

The policy selects a verification procedure using the current state, the proposed action, and the available status information. A validated confidence estimate may help determine which verification procedure is likely to be efficient. A low-confidence answer may warrant additional retrieval or independent calculation. A high-confidence answer may warrant a less expensive verification procedure when the applicable obligations permit such treatment. However, the selection of a verification procedure does not establish the result of that procedure.

The distinction can be expressed through two separate mappings:

$$\rho(s, a, \sigma) = v,$$

and

$$\text{Verify}(v, s, a, E) = \sigma'.$$

The first mapping selects an evidentiary operation, while the second produces an updated status representation from the execution of that operation. The routing decision must not be mistaken for the verification result. This distinction permits statistical optimization of verification costs without making the confidence estimator the final authority over the obligations being examined.

10.6 The decision-theoretic role of probability

There are circumstances in which a probability threshold is an appropriate component of an operational decision. Consider a simplified choice between executing an action and abstaining. Suppose execution produces a benefit $B > 0$ when successful and a loss $L > 0$ when unsuccessful. Let p denote the probability of success under a specified model. If abstention has utility zero, the expected utility of execution is

$$\mathbb{E}[U(a)] = pB - (1 - p)L.$$

Execution is preferred under this model when

$$pB - (1 - p)L \geq 0,$$

or equivalently,

$$p \geq \frac{L}{B + L}.$$

The probability threshold thus follows from the specified consequences and utility model and is not an intrinsic property of confidence, since changing the relative magnitude of the benefit and loss changes the threshold. Moreover, the calculation assumes that the relevant consequences have been represented, that the probability estimate is appropriate to the particular action, and that the decision-maker possesses authority to accept the associated risks.

The expected utility calculation does not establish these assumptions. A more general constrained decision problem takes the form

$$a^* \in \arg\max_{a \in \mathcal{A}_{\text{adm}}} \mathbb{E}[U(a, S)],$$

where

$$\mathcal{A}_{\text{adm}} = \{a \in \mathcal{A} : \text{Adm}(s, a, E) = 1\}.$$

Here probabilistic optimization operates within an independently established admissible set. The distinction preserves the usefulness of decision theory without assuming that every institutional, physical, or evidentiary constraint can be replaced by a finite utility penalty. Some obligations may themselves be probabilistic. For example, an engineering specification may require that the estimated probability of a particular failure remain below a prescribed threshold. Such an obligation is entirely compatible with the proposed architecture. Its satisfaction nevertheless requires evidence concerning the validity of the probability estimate, the definition of failure, the operating domain, and the authority responsible for establishing the threshold. The objection is therefore directed, not at probabilistic requirements, but at the unrestricted substitution of a generic confidence score for the specific requirements governing an action.

10.7 Admissibility under observational ambiguity

The non-identifiability results developed in Section 5 can be extended from inference to operational decision-making. Let

$$R : \mathcal{S} \longrightarrow \mathcal{O}$$

denote an observation map. For each underlying state s, let

$$\mathcal{A}_{\text{adm}}(s) \subseteq \mathcal{A}$$

denote the actions that are admissible in that state, as determined by the complete state and not, as in the preceding evidence-relative definition, by the available evidence E. An observer receiving $o = R(s)$ does not necessarily know which underlying state produced the observation. The observer instead possesses an observational equivalence class

$$\mathcal{S}_o = R^{-1}(\{o\}).$$

If the system requires an action that remains admissible in every state compatible with the observation, the appropriate action set is

$$\mathcal{A}_{\text{rob}}(o) = \bigcap_{s \in \mathcal{S}_o} \mathcal{A}_{\text{adm}}(s).$$

Definition 10.2 (Robust observational admissibility)**.** An action is robustly admissible relative to observation o when it belongs to the admissible action set of every underlying state compatible with o.

This criterion concerns a worst-case interpretation of the available evidence and may be more restrictive than a probabilistic risk criterion, so that its appropriateness depends upon the obligations governing the operation.

Proposition 10.2 (Safe-continuation obstruction)**.** *Suppose an observation o satisfies*

$$\bigcap_{s \in \mathcal{S}_o} \mathcal{A}_{\mathrm{adm}}(s) = \varnothing.$$

Then no deterministic policy depending exclusively upon o can select an action that is admissible in every state compatible with the observation.

Proof. Let

$$\pi : \mathcal{O} \longrightarrow \mathcal{A}$$

be an arbitrary deterministic policy. The action $\pi(o)$ is the same for every underlying state belonging to $\mathcal{S}_o$. For this action to be admissible in every such state, it would have to satisfy

$$\pi(o) \in \bigcap_{s \in \mathcal{S}_o} \mathcal{A}_{\mathrm{adm}}(s).$$

The intersection is empty by assumption.

Consequently, no such action exists. □

The proposition establishes an informational limitation upon autonomous continuation. The difficulty is not necessarily that the system has failed to identify the most probable state, but that the available observation does not support a uniformly admissible action. A randomized policy cannot satisfy the same robust requirement unless its distribution assigns probability one to the admissible action set of every compatible state. For finite or countable action spaces, an empty intersection rules out such a policy directly. More general action spaces require appropriate measurability and intersection conditions. The distinction matters because stochastic action selection can distribute risk without eliminating the underlying ambiguity.

10.8 Refinement as an alternative to premature action

The safe-continuation obstruction does not imply that the system must terminate permanently. It may be possible to acquire additional evidence that refines the observational equivalence class. Suppose an additional observation o' restricts the admissible underlying states to

$$\mathcal{S}_{o,o'} \subseteq \mathcal{S}_o.$$

The corresponding robust action set is

$$\mathcal{A}_{\mathrm{rob}}(o, o') = \bigcap_{s \in \mathcal{S}_{o,o'}} \mathcal{A}_{\mathrm{adm}}(s).$$

Since the refined state space is a subset of the original state space,

$$\mathcal{A}_{\mathrm{rob}}(o) \subseteq \mathcal{A}_{\mathrm{rob}}(o, o').$$

Additional exact information can therefore enlarge the collection of actions that are certifiably admissible. An action previously excluded by unresolved ambiguity may become admissible after the relevant distinction has been established. This suggests an alternative interpretation of autonomous competence. A competent system is not merely one that selects actions from the available information. It is also one that identifies which additional information would make an otherwise inadmissible continuation possible. The acquisition of information must itself remain subject to applicable obligations.

An additional observation may require permission, consume resources, or alter the state being observed. Nevertheless, the architectural principle remains that an unresolved distinction should motivate an appropriate refinement operation rather than an arbitrary increase or decrease in a confidence score.

10.9 Predictive sufficiency and operational sufficiency

A representation can be adequate for prediction while remaining inadequate for action selection. Let

$$\phi : \mathcal{S} \longrightarrow \mathcal{Z}$$

denote an abstraction of the underlying operational state. Suppose ϕ preserves the information required to predict a particular outcome Y. This predictive property does not imply that ϕ preserves the distinctions required to determine authorization or admissibility. Two states may satisfy

$$\phi(s_1) = \phi(s_2)$$

while possessing different authorization conditions. For example, two otherwise identical files may be associated with different ownership or access permissions. A representation that preserves their contents but suppresses their permission metadata may be sufficient for predicting the result of a text-processing operation. The same representation may nevertheless be insufficient for determining whether the operation is authorized.

Definition 10.3 (Admissibility-preserving abstraction)**.** An abstraction ϕ preserves

action admissibility when

$$\phi(s_1) = \phi(s_2) \quad \Longrightarrow \quad \mathcal{A}_{\mathrm{adm}}(s_1) = \mathcal{A}_{\mathrm{adm}}(s_2).$$

This is a strong condition. It requires the abstraction to preserve every distinction relevant to the admissible action set. A weaker conservative condition permits the abstraction to retain only actions that remain admissible throughout each fibre:

$$\mathcal{A}_{\phi}(z) \subseteq \bigcap_{\phi(s)=z} \mathcal{A}_{\mathrm{adm}}(s).$$

Such an abstraction may exclude some actions that would be admissible if the complete state were known. However, it does not authorize an action merely because the distinctions responsible for its inadmissibility have been suppressed. The distinction between predictive and operational sufficiency is especially important for systems that use compressed representations, learned embeddings, or summary descriptions as the basis for consequential decisions. Predictive accuracy does not establish that the representation preserves every distinction relevant to permission, compatibility, or recovery.

10.10 Local verification and global compatibility

A further limitation of scalar confidence arises when several individually evaluated actions must be composed. Suppose two proposed operations, a_1 and a_2, each satisfy their local obligations when considered independently. It does not follow that their composition satisfies the obligations governing the complete system. Consider a resource with available capacity B. Suppose operation a_1 requires resource quantity b_1, while operation a_2 requires quantity b_2. The individual conditions

$$b_1 \leq B$$

and

$$b_2 \leq B$$

do not imply

$$b_1 + b_2 \leq B.$$

Both operations may be locally feasible while their simultaneous execution violates a global resource constraint. Analogous failures arise in software modification, concurrent database operations, financial allocation, and distributed control. Two changes may each pass local tests while their composition violates an interface assumption. Two transactions may each be authorized relative to the same initial balance while their

combined execution exceeds the available funds.

The difficulty concerns the compatibility of operations, not merely their individual reliability.

Proposition 10.3 (Local admissibility does not imply compositional admissibility)**.** *There exist operational systems and actions a_1, a_2 such that*

$$\mathrm{Adm}(s, a_1, E) = 1$$

and

$$\mathrm{Adm}(s, a_2, E) = 1,$$

while their proposed joint execution is inadmissible.

Proof. Consider a resource with capacity $B = 10$.

Let each action require six units of that resource during the same execution interval. Each action satisfies the individual capacity constraint

$$6 \leq 10.$$

Their simultaneous execution requires

$$6 + 6 = 12 > 10.$$

The joint execution therefore violates the global capacity constraint. □

The proposition is elementary, but its architectural consequences are substantial. An autonomous system cannot establish global admissibility merely by accumulating local confidence scores or local verification results. It must examine the conditions under which those results can be composed. The compatibility obligations may depend upon shared resources, temporal overlap, mutable state, authorization scope, and dependencies among operations. These conditions must remain available to the system responsible for committing the combined continuation.

10.11 Temporal validity and stale evidence

Evidence establishing the admissibility of an action at one time need not establish its admissibility at a later time. Let s_t denote the operational state at time t. Suppose an action a is examined relative to s_{t_0} and receives a successful verification result. Before execution at time t_1, the operational state may change. The relevant authorization may

expire, a file may be modified, a resource may be consumed by another operation, or an institutional document may be superseded. Consequently,

$$\mathrm{Adm}(s_{t_0}, a, E) = 1$$

does not generally imply

$$\mathrm{Adm}(s_{t_1}, a, E) = 1.$$

The verification result must be bound to the state, version, or conditions under which it was established. This is related to familiar time-of-check and time-of-use problems in computational systems. The appropriate remedy may involve version checks, transactional execution, resource reservation, or reverification immediately before commitment. A high historical confidence estimate does not resolve the discrepancy between the state that was examined and the state that will actually be modified.

The temporal provenance of evidence is therefore part of its operational significance.

10.12 Reversibility and the consequences of error

The admissibility of an action also depends upon the consequences of an incorrect decision and the availability of recovery procedures. Let

$$F : \mathcal{S} \times \mathcal{A} \longrightarrow \mathcal{S}$$

denote a state transition operation. An action a is exactly reversible on a specified domain if there exists a recovery operation r such that

$$r(F(s, a)) = s$$

for every relevant underlying state s. Many practical operations are not exactly reversible. A compensating transaction may restore an account balance without erasing the original transaction. A corrected message may communicate the appropriate information without undoing the effects of the original message. A restored software file may recover its contents without reversing external processes triggered by its temporary absence. Recovery should therefore be evaluated relative to the properties that must be restored. Let

$$Q : \mathcal{S} \longrightarrow \mathcal{Q}$$

denote a representation of the relevant recovery properties. A weaker recovery condition is

$$Q(r(F(s, a))) = Q(s).$$

This condition permits differences in the complete state while requiring the restoration of specified properties.

Its adequacy depends upon whether Q preserves every distinction relevant to the recovery obligation. The existence of a nominal undo operation does not establish that the consequences of an incorrect action can be contained. Consequently, two actions with identical estimated probabilities of success may require different verification procedures because their failure consequences and recovery conditions differ.

10.13 Evidence-bearing authorization

The preceding analysis suggests that authorization should be associated with an explicit evidentiary object. Let a proposed continuation be represented by

$$P = (s, a, \mathcal{O}, E),$$

where s identifies the relevant state, a identifies the proposed action, $\mathcal{O}$ specifies the applicable obligations, and E contains the available evidence. A verification procedure produces a certificate

$$K = \mathrm{Certify}(P).$$

The certificate may contain the results of individual checks, source references, authorization records, version identifiers, and unresolved conditions. A separate commitment operation examines whether the certificate is sufficient for the proposed transition. Schematically,

$$\mathrm{Commit}(P, K) = \begin{cases} s', & \text{if the applicable conditions hold,} \\ \mathsf{refuse}, & \text{otherwise.} \end{cases}$$

The certificate need not be a mathematical proof in every application, and it may instead be a structured record of empirical, institutional, or computational evidence. Its defining property is that the grounds of authorization remain available for examination. A scalar confidence estimate may appear within the certificate where relevant.

It does not replace the remaining evidentiary objects.

10.14 The distinction between certificate existence and certificate validity

The existence of a certificate does not establish its validity. A computational system can generate a document asserting that every obligation has been satisfied without

actually performing the required checks. A certificate can also be outdated, associated with the wrong object, or constructed under an inadequate specification. The proposed architecture therefore requires a distinction between certificate production and certificate verification. Let

$$\mathrm{Check}(P, K) \in \{0, 1\}$$

denote the procedure responsible for examining the certificate. A soundness condition would require

$$\mathrm{Check}(P, K) = 1 \quad \Longrightarrow \quad \mathrm{Adm}(s, a, E) = 1.$$

The establishment of this condition depends upon the semantics of the certificate, the correctness of the checker, and the adequacy of the underlying obligation specification. This is soundness relative to a specification and a body of evidence, and it must be distinguished from the validity of the resulting transition in the world. Let $F(s, a)$ denote the state produced by executing a in s, and let $\mathcal{G} \subseteq \mathcal{S}$ denote the states satisfying the requirements actually intended. External validity of a certified action is the condition

$$\mathrm{Check}(P, K) = 1 \quad \Longrightarrow \quad F(s, a) \in \mathcal{G}.$$

It follows from internal soundness only under the additional bridging assumption

$$\mathrm{Adm}(s, a, E) = 1 \quad \Longrightarrow \quad F(s, a) \in \mathcal{G},$$

which asserts that the obligations are adequate to the intended requirements and that the evidence accepted as satisfying them is correct. No checker confined to P and K can establish this assumption merely by being internally sound.

Where the obligations concern formal properties, soundness may be established through mathematical proof. Where they concern empirical or institutional conditions, soundness may instead be conditional upon the reliability of specified observations, records, and authorities.

The distinction between formal and empirical certification must therefore remain explicit. Without promising infallibility, the architecture establishes a division of responsibility in which the production of an assertion does not automatically constitute the verification of that assertion.

10.15 A worked customer-support example

The proposed architecture can be applied to the customer-support system discussed in the podcast. Suppose a customer asks whether a particular purchase is eligible for a refund. The generative system retrieves a document and produces an affirmative answer.

An external confidence estimator assigns the answer a score of 0.96. Under a simple threshold architecture, the answer might be communicated automatically, and the refund might subsequently be initiated without additional examination. The obligation-based architecture distinguishes these operations. Communicating a general description of the refund policy, establishing that the particular purchase satisfies the policy, and initiating the financial transaction are three separate operations. Each operation possesses different evidentiary and authorization requirements. The general description may require a current and authoritative policy document. The eligibility determination may additionally require the purchase date, product category, transaction status, and applicable exceptions.

The financial transaction may require authentication, permission, verification of the recipient, and confirmation that the refund has not already been issued. The confidence score of 0.96 does not establish any missing transaction record or authorization condition. It may indicate that answers possessing similar observable features have historically been correct at approximately the reported frequency. That information can contribute to the selection of an appropriate review procedure. The actual refund remains subject to its specific operational obligations. Suppose the policy document is current and the generated description is accurate, but the purchase date is unavailable. The system can communicate the general policy while leaving the eligibility determination unresolved. It can request the missing information or retrieve the relevant transaction record if authorized to do so. It need not assign an arbitrary lower confidence to the entire interaction.

The unresolved condition can be represented directly. Suppose instead that the purchase is established as eligible, but the system lacks permission to initiate the refund. The eligibility claim can remain verified while the transaction is refused or routed to an authorized agent. The absence of permission does not make the eligibility claim less true. The verified claim and the refused action possess different statuses. This example illustrates the practical advantage of preserving the structure of the obligations rather than compressing them into a single confidence estimate.

10.16 Audit and certification as complementary operations

The randomized audit procedures developed in Section 9 remain important within an obligation-based architecture. Certification addresses the requirements of a particular continuation, whereas auditing addresses the empirical performance of the procedures responsible for producing and evaluating such continuations. A verification procedure can be formally specified while still exhibiting implementation defects, inadequate empirical assumptions, or systematic failures outside its validated domain. Audits can reveal such defects. Conversely, a favourable audit record does not establish that every particular

continuation satisfies its mandatory obligations. The two operations therefore supply different forms of evidence. An architecture that relies exclusively upon case-specific checks may fail to detect systematic weaknesses in its verification procedures. An architecture that relies exclusively upon aggregate audit results may authorize particular actions without examining their relevant conditions. A defensible system requires both forms of evaluation where their respective purposes apply.

The audit record should inform the revision of estimators, verification procedures, and obligation specifications. It should not silently replace those specifications.

10.17 Refusal as an informative computational result

A system that cannot establish the admissibility of a proposed continuation must be capable of representing that condition without fabricating a substitute result. The refusal should preserve the distinction between an established violation and an unresolved requirement. For example, a proposed action may be refused because the available authorization explicitly excludes it. Alternatively, execution may be deferred because a required observation has not yet been obtained. The first condition may require a change in the proposed action or a legitimate change in authorization. The second may require additional measurement or verification. An informative refusal can be represented schematically as

$$R = (a, \mathcal{O}_f, \mathcal{O}_u, E),$$

where $\mathcal{O}_f$ denotes failed obligations and $\mathcal{O}_u$ denotes unresolved obligations. The representation can additionally identify permissible operations that would resolve particular unresolved conditions. Such a refusal is not merely an absence of output.

It is an evidentiary result describing the boundary of currently admissible continuation. A system capable of producing this result possesses a form of methodological competence that is not captured by the frequency with which it supplies answers. Its competence includes the preservation of unresolved distinctions and the identification of conditions under which further activity would become legitimate.

10.18 The limits of obligation-based architectures

An obligation-based architecture does not eliminate every difficulty associated with autonomous systems. The obligations may themselves be incomplete, incorrect, mutually incompatible, or inadequately specified, the evidence used to evaluate them may be defective, and the authority responsible for establishing them may be mistaken or institutionally illegitimate. A verification procedure may establish a formal property that fails to correspond to the intended empirical requirement. These possibilities are not

resolved merely by introducing additional formal structure. The proposed architecture instead makes the dependencies more explicit. It permits a failure to be associated with a particular obligation, observation, verification procedure, or authorization rule. It also permits the obligation set itself to become an object of examination and revision. The architecture should therefore be understood as a means of preserving epistemic and operational distinctions, rather than as a universal guarantee of correct behaviour.

Its central advantage is that it does not require every unresolved scientific, institutional, or computational question to be translated into a single numerical quantity. It retains the possibility of identifying what remains unknown, what has been established, what is permitted, and what must occur before further continuation becomes admissible.

11 Conclusion: The Right Not to Quantify

The argument began with a comparatively narrow engineering criticism, namely that a numerical confidence estimate generated by a language model is not, by virtue of its numerical form, evidence of statistical calibration. The criticism is correct when understood as a demand for empirical validation, although it does not show that the underlying computation is incapable of carrying information about correctness. Its significance also extends beyond any particular architecture. A probability may be assigned without a specified reference class, alternatives may be ranked without a justified comparison relation, a quantity may be estimated from observations that do not identify it, and an estimator may be calibrated against a proxy while remaining unvalidated against its intended target. These are different failures in the relationships among representation, observation, inference, evaluation, and action, and their common feature is the substitution of a formally available result for a substantively established relationship.

11.1 What the distinctions preserve

None of this counsels against mathematical representation or heuristic inference. Approximate estimators, sparse probabilistic networks, informal rankings, and compressed abstractions remain valuable when their limitations are characterized relative to the use made of them. The requirement is that an approximation remain recognizable as such and that a representation adequate for suggesting an action not be presumed adequate for authorizing it. Where an observation fails to separate states that differ in the quantity of interest, the remedy lies in further evidence, in justified restrictions upon the state space, or in reporting an identified set. Where the correct explanation is absent from a hypothesis space, posterior concentration within that space offers no remedy, and an

unavailable probability should not be confused with a small one.

11.2 Observability and authority

The selective verification results show that a confidence estimator cannot be evaluated apart from the policy determining which outcomes become observable, since a precise record may remain structurally incomplete. The preservation of observability is therefore an operational obligation, to be discharged through audit designs, recorded verification propensities, independent adjudication of original answers, and versioned records. Even a validated estimate does not by itself establish permission, compositional compatibility, temporal validity, or the availability of recovery. Confidence may organize verification and may guide optimization within an admissible set, but it does not substitute for the results of the checks that define that set, and the soundness of those checks remains relative to the adequacy of their specification.

11.3 The right not to quantify

The phrase “the right not to quantify” names a methodological principle and not an unrestricted legal or institutional entitlement. A scientific or computational procedure should be capable of declining a requested numerical representation when the conditions for its interpretation have not been established, provided that it states which condition is missing. Refusal of this kind may be accompanied by a weaker but defensible result, so that an exact estimate gives way to an identified set, a total ordering to a partial ordering, and an unsupported probability to an account of the available evidence, while an action that is not yet admissible gives way to a request for the information or authorization required to proceed.

The closing question posed in the podcast asks what a reported percentage refers to and against what it has been checked. The present study adds four further questions: what remains unobservable, which possibilities have been excluded, how the evaluation policy determines which errors become visible, and what operation the resulting evidence authorizes. Their answers need not share a mathematical form, and a universal confidence score would suppress that diversity of structure. Confidence can inform the judgment of what may legitimately occur next, but confidence is not certification.

References

Begg, C. B., and Greenes, R. A. (1983). Assessment of Diagnostic Tests When Disease Verification Is Subject to Selection Bias. *Biometrics*, 39(1), 207–215.

Berk, R. H. (1966). Limiting Behavior of Posterior Distributions When the Model Is Incorrect. *The Annals of Mathematical Statistics*, 37(1), 51–58.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1), 1–3.

Campbell, D. T. (1979). Assessing the Impact of Planned Social Change. *Evaluation and Program Planning*, 2(1), 67–90.

Cronbach, L. J., and Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, 52(4), 281–302.

Data Science at Home (2026, November 13). *Your LLM's Confidence Score Is Made Up*. YouTube. `https://www.youtube.com/watch?v=UJ3f4anDQNY`.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A Survey on Concept Drift Adaptation. *ACM Computing Surveys*, 46(4), Article 44.

Hébert-Johnson, U., Kim, M. P., Reingold, O., and Rothblum, G. N. (2018). Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 80, 1939–1948.

Horvitz, D. G., and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47(260), 663–685.

Kadavath, S., et al. (2022). Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*. `https://arxiv.org/abs/2207.05221`.

Kleijn, B. J. K., and van der Vaart, A. W. (2006). Misspecification in Infinite-Dimensional Bayesian Statistics. *The Annals of Statistics*, 34(2), 837–877.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). *Foundations of Measurement. Volume I: Additive and Polynomial Representations*. Academic Press.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Lin, S., Hilton, J., and Evans, O. (2022). Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*. https://arxiv.org/abs/2205.14334.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer.

Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, 12(4), 595–600.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Second edition. Cambridge University Press.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103(2684), 677–680.

Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158), 209–212.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. (2024). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *Proceedings of the International Conference on Learning Representations*. https://arxiv.org/abs/2306.13063.