# Machine Understanding

## Representation, Explanation, and Transfer

**Komorebi P.**
8 October 2026

*First consolidated edition*
Version 1.0.0

### Abstract

A correct answer does not identify a reusable method, a faithful account of a computation, or an explanation another reasoner can use. This monograph develops a contract-relative account of computational understanding through representations that support prediction, intervention, transfer, and correction. Finite behavioral equivalence characterizes exact reusable state and the distinctions required by new operations. Sender access and recipient background determine what an explanation must convey. State-changing experiment planning makes acquisition costs and physical effects explicit. For noisy evidence, a likelihood factorization characterizes preservation of every Bayesian update, while separate bounds distinguish average information loss from rare-event posterior error. The first-edition synthesis fixes the recipient procedures as well as the information channel. A covering criterion identifies usable artifacts, and a finite resource construction jointly accounts for online work and the entire compiled observation tree. Task-dependent interpreter changes are charged rather than hidden in background knowledge. A three-register instrument connects these results in one executable account of acquisition, correction, and explanation. The contribution is an integrated realization-and-preservation framework with conditional proofs and explicit failures. It is not a universal test of understanding, an empirical claim about a deployed model, or a criterion for subjective experience.

**First-edition scope.** A theoretical account with explicit finite contracts, self-contained prose proofs, and executable witnesses. Mathematical validity, implementation evidence, and empirical attribution remain distinct. Version 1.0 freezes this scope rather than claiming a universal test of understanding.

# Contents

**Reading the first edition** ii

**I Representations and implemented use** 1
**1 What a correct answer leaves open** 1
**2 A small instrument with a hidden register** 2
**3 A contract for reusable understanding** 4
**4 Behavioral compression and the price of new questions** 5
**5 Available information is not implemented understanding** 7
**6 Transporting a method rather than matching an answer** 9

**II Explanation, inquiry, and correction** 13
**7 An explanation belongs to a recipient** 13
**8 One explanation for recipients with different backgrounds** 14
**9 Learning where the current representation fails** 16
**10 Acquisition as an experiment planning problem** 18
**11 When translation commutes with correction** 20
**12 Noisy evidence and Bayesian correction** 22
**13 Approximation, rare evidence, and decision stability** 25
**14 Sequential inquiry through a retained evidence channel** 27

**III Common witnesses and charged resources** 29
**15 One witness for prediction, transfer, and revision** 29
**16 Joint acquisition and explanatory communication** 30
**17 What the resource account must charge** 33
**18 Explanations for fixed recipient procedures** 34
**19 Joint acquisition with a compiled resource budget** 35
**20 A costed first-edition instrument certificate** 37

**IV Interpretation and scope** 40
**21 Limits, objections, and the next mathematical questions** 40
**22 Conclusion** 41
**A A complete trace of the instrument** 43
**B Companion verification and evidence boundaries** 44
**C Claim map** 45
**References** 47

## Reading the first edition

The paper asks what a computational system can carry from one question to the next, from one representation to another, and from an accepted model to a corrected one. Its central object is a usable representation together with its operations and limits. It is not a score assigned to a model after one successful answer.

A conceptual reading can follow Sections 1–3, the interpretations after the formal results, and Section 21. A mathematical reading follows four parts: reusable representations and implemented use, explanatory communication and correction, common resource witnesses, and interpretive boundaries. The final synthesis freezes the recipient procedures and charges both online work and compiled controller storage. Appendix A gives one complete numerical trace. The source package includes exact reference programs, a statement index, a claim ledger, and a review record.

**What is inherited and what is proposed.** State minimization, causal abstraction, hypothesis testing, counting bounds, and error propagation have substantial established literatures. The proofs below make the particular assumptions used here inspectable. Their presence is not a claim of priority for those mathematical ingredients. The proposed integration is to require a claim about understanding to identify its target, its realized use, its admissible changes of representation, its recipient, its correction procedure, and its total resource boundary.

**Relation to Operation Komorebi.** *Machine Societies* studies persistent plural institutions and the preservation of their operations and commitments [1]. This companion changes the object of study. It concerns methods and representations that one computational reasoner can use, explain, transfer, and revise. It uses the same realization perspective without treating institutional continuity as evidence of individual understanding.

**Conventions.** The main systems are finite and fully specified. Deterministic acquisition and stochastic correction are distinguished. Each probabilistic result states its own likelihood and selection assumptions. A metric transport result also applies outside finite spaces, with explicit uniform bounds. A finite check supports its implemented construction, not the general theorem. Mathematical validity inside a model is distinct from evidence that an external system realizes that model.

| Symbol | Meaning |
|---|---|
| $C$ | Declared family of states, operations, and future questions. |
| $F_u$, $F_w$ | One implemented operation, or operations in the order specified by a word $w$. |
| $r : \mathcal{X} \to \mathcal{Z}$ | Retained representation. Its acquisition procedure is a separate obligation. |
| $\bar{F}_u$, $\bar{q}$ | An update and a query decoder operating on the retained state. |
| $T$, $\tau$ | Translation of representations and the corresponding operation names. |
| $V \subseteq \mathcal{M}$ | Current set of candidate models, not a probability distribution. |
| $C_{e,o}(V)$ | Candidates retained after experiment $e$ returns observation $o$. |
| $\phi$, $\eta$ | Translation of candidate models and translation of observations. |
| $d_{\mathrm{TV}}(P, Q)$ | $\frac{1}{2}\sum_x \lvert P(x) - Q(x)\rvert$ on a finite outcome space. |

# Part I

# Representations and implemented use

## 1 What a correct answer leaves open

A reasoner is asked what a device will display after two operations. It gives the right answer. That success is a fact worth preserving. It does not yet tell us whether the reasoner has a method that will work after three operations, whether the order of the operations matters, or whether the same method could be expressed in another notation. It also does not tell us whether the explanation offered afterward describes the computation that produced the answer.

These are not objections to correctness. They are different questions about what correctness rests on and what it makes possible. A proof can establish a conclusion from premises without teaching a recipient how to select the relevant premises in a new case. A useful model can guide new predictions without being a faithful description of the particular implementation that supplied it. A fluent explanation can be understandable while failing to identify the mechanism it claims to explain. The distinction between persuasive and faithful explanations is explicit in interpretability research [2].

This paper investigates a narrower and more tractable question than whether a machine understands in every human sense of the word. It asks when a computational realization possesses a representation that remains useful across a declared family of operations, questions, translations, and corrections. We call this *reusable computational understanding* as an operational research proposal. It is relative to a domain and a contract. It is not a metaphysical property inferred from a single formal condition.

The central thesis is that a representation earns its explanatory role through what a system can do with it. It must retain distinctions needed for future questions, support the operations it claims to summarize, and expose the conditions under which its use should change. When it is communicated, a recipient must be able to recover something useful from it. When it is attributed to a particular mechanism, the attribution requires evidence about that mechanism. A representation that meets one of these obligations need not meet the others.

### 1.1 Three separations organize the argument

The first separation is between an answer and a reusable state. The answer to today's question can discard distinctions that tomorrow's permitted operation makes important. Understanding in the present sense therefore concerns a family of possible continuations, not only a completed input and output pair.

The second is between a correct account of a subject and a faithful account of a solver. An algebraic derivation can explain a numerical result even when the original solver used a table. It would be wrong to infer that the solver executed the derivation merely because the derivation is valid. Subject understanding and process faithfulness require different witnesses.

The third is between successful transport and preserved correction. Two languages may translate every currently expressible observation while one language merges distinctions needed to reject a mistaken model. Agreement in the translated language is then weaker than agreement about what has been learned. The correction theorem in Section 11 makes that loss exact.

### 1.2 Position among existing theories

Several neighboring theories already supply essential pieces. Computational mechanics studies minimal predictive state representations [3]. Controlled state aggregation studies equivalences that preserve transition

structure [4]. Information bottleneck methods formalize compression relative to a specified relevance variable [5]. These are not interchangeable claims, but they establish that useful compression must be assessed against what it promises to preserve.

Causal abstraction gives a stronger account of agreement under interventions. Beckers and Halpern distinguish increasingly restrictive abstraction conditions [6]. Geiger and colleagues develop this perspective as a foundation for mechanistic interpretability, including transformations of mechanisms [7]. Our finite commuting equations occupy a deliberately restricted part of that subject. They do not replace a general theory of causal models.

There are also reasons to constrain the maps used in an interpretation. Hewitt and Liang show why a trained probe's success must be assessed against the work the probe itself can perform [8]. Sutter and colleagues examine how sufficiently expressive alignment maps can make some causal-abstraction claims uninformative [9]. The constructive theory and this criticism address different aspects of the problem. We retain both lessons: intervention compatibility matters, and the translator must not be allowed to hide the entire explanation inside an unrestricted computation.

Explanation has a recipient as well as a referent. Miller's synthesis emphasizes the contextual and social features of explanation [10]. Gentner's structure-mapping account places relational organization, rather than surface resemblance alone, at the center of analogy [11]. Here we formalize only a limited recipient-relative property: an artifact enables specified future operations in a recipient with stated background resources. That property does not exhaust the reasons a human explanation can succeed or fail.

Finally, learning through informative queries and counterexamples has a substantial theoretical history, including Angluin's work on regular languages [12]. Our correction model is much simpler. It is a finite filter on candidate models, chosen so that the interaction between evidence and translation can be proved directly.

### 1.3 The intended contribution

The paper does not propose that every useful representation is new, that a symbolic method is inherently superior to a learned one, or that a large table can never support understanding. It proposes a common account of the obligations that such representations must satisfy. The exact finite results locate information requirements and compatibility conditions. The counterexamples show why isolated successes cannot silently substitute for a combined claim.

This is also a limit on the theory itself. A finite contract can be satisfied by many implementations, including tables. To distinguish them, one must add relevant constraints on acquisition, causal organization, storage, computation, or transfer. Those constraints cannot be recovered from a definition after the results are known.

## 2 A small instrument with a hidden register

Consider a mathematical instrument with three registers

$$x = (a, b, c) \in \mathbb{F}_5^3.$$

Each register contains one of $0, 1, 2, 3, 4$. All arithmetic in this example is modulo five. The display reports

$$q(a, b, c) = a + b.$$

There are 125 possible complete states, but only five possible readings. The registers and their operations are fully specified analytical objects. They are not measurements from an agent experiment.

Initially, two operations are available:

$$H(a, b, c) = (a + b, b, c), \qquad W(a, b, c) = (b, a, c). \tag{1}$$

The first adds the second register into the first. The second exchanges the first two registers. The third register is preserved by both and does not affect the display. The operation order is public. A prediction concerns what the display would show after a named finite sequence of these operations.

### 2.1 Knowing the current display is insufficient

A record containing only $s = a + b$ answers the current question exactly. It cannot, in general, predict the display after $H$. The two complete states

$$x = (0, 0, 0), \qquad y = (1, 4, 0)$$

both display zero. After $H$, their displays are zero and four. Any predictor receiving only the common old display must return the same answer for both, so it cannot be exact in both cases.

A sufficient retained state for the original operation family is

$$z = (s, b) = (a + b, b).$$

Its updates are

$$\bar{H}(s, b) = (s + b, b), \qquad \bar{W}(s, b) = (s, s - b), \tag{2}$$

and its displayed output is its first coordinate. These equations retain the information needed for every finite $H/W$ continuation without retaining $c$.

The representation is also minimal in its number of states. The present reading $s$ and the reading $s_H$ after a hypothetical $H$ determine

$$b = s_H - s, \qquad a = 2s - s_H.$$

Two different pairs $(a, b)$ therefore differ on at least one of these questions. Every exact reusable representation for this contract needs at least 25 distinguishable values. Section 4 gives the general argument.

### 2.2 A new operation changes what must be known

Now admit

$$J(a, b, c) = (a + c, b, c). \tag{3}$$

The old representation remains correct for $H$ and $W$. It is no longer sufficient for the enlarged contract. States with the same $(a, b)$ but different $c$ now produce different readings after $J$.

The additional reading $s_J$ yields

$$c = s_J - s.$$

Together with the earlier distinctions for $a$ and $b$, the enlarged question family distinguishes all 125 complete states. The useful state counts are therefore five for the current display, 25 for reusable $H/W$ prediction, and 125 after $J$ is admitted.

This is not a case of an old theorem becoming false. The old theorem had a smaller operation family. What failed was an unlicensed extension of its conclusion. The distinction matters whenever a learned method is reused under a change in task, environment, or intervention.

### 2.3 Correction requires evidence, not a renamed variable

A predictor that has retained only $(s, b)$ cannot recover $c$ merely by adding a third field to its data structure. Conditional on a fixed $(a, b)$ and a uniformly distributed $c$, every possible display after $J$ is equally likely. Any guess based only on the old record succeeds with probability $1/5$. Randomizing the guess does not improve that probability.

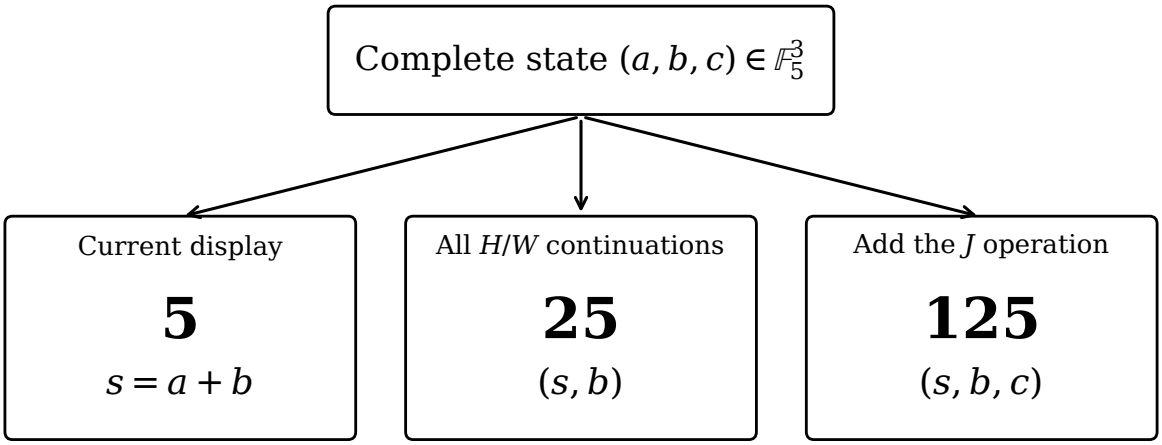



Figure 1: The same instrument supports different minimal representations under different contracts. The numbers count distinguishable states, not scores of intelligence. A new permitted operation makes a previously irrelevant register relevant.

An actual execution of $J$, followed by a new reading, supplies the missing distinction. Because $J$ changes the instrument, the recovered state is the state *after* that operation:

$$(a', b', c') = (s_J - b,\ b,\ s_J - s). \tag{4}$$

The old state can then be reconstructed because this particular operation is invertible. Neither reversibility nor a free observation should be assumed for other instruments.

The example separates four achievements: reading an answer, retaining a method, recognizing a contract change, and acquiring evidence that repairs the resulting information gap. No amount of confidence in the original method supplies the fourth.

## 3 A contract for reusable understanding

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite nonempty sets. Let $\mathcal{U}$ be a finite operation alphabet, possibly empty, and let $\mathcal{Q}$ be a finite nonempty family of queries $q : \mathcal{X} \to \mathcal{Y}$. Each $u \in \mathcal{U}$ has an implemented total transition $F_u : \mathcal{X} \to \mathcal{X}$. The deterministic mathematical contract is

$$C = (\mathcal{X}, \mathcal{U}, (F_u)_{u \in \mathcal{U}}, \mathcal{Q}, \mathcal{Y}).$$

For a word $w = u_1 \cdots u_n$, define $F_w = F_{u_n} \circ \cdots \circ F_{u_1}$. The empty word acts as the identity. Thus $q(F_w(x))$ asks a question after a stated continuation.

Totality is a simplifying assumption. If an operation is sometimes unavailable, its admissibility must be included in the observable contract or the model must use partial transitions with explicit domain preservation. An unmodeled refusal cannot be treated as the same event as a permitted update.

**Definition 3.1** (Exact reusable representation)**.** An exact reusable representation for $C$ is a surjection $r : \mathcal{X} \to \mathcal{Z}$, updates $\bar{F}_u : \mathcal{Z} \to \mathcal{Z}$, and decoders $\bar{q} : \mathcal{Z} \to \mathcal{Y}$ such that, for all $x, u, q$,

$$r(F_u(x)) = \bar{F}_u(r(x)), \qquad q(x) = \bar{q}(r(x)). \tag{5}$$

Induction immediately gives $q(F_w(x)) = \bar{q}(\bar{F}_w(r(x)))$ for every finite word. The same initial record supports many future questions. It is not allowed to inspect the selected future query and replace itself with that query's answer before the challenge begins.

### 3.1 A representation of the problem is not yet a representation in the solver

The map $r$ specifies what information would suffice. It does not establish that a particular agent can acquire $r(x)$ from its sensors or that the agent uses this state. Those claims require an acquisition procedure, a realized computational path, and admissible interventions on that path.

For the instrument, an omniscient mathematical encoder can read $(a, b, c)$. A device restricted to the display cannot do so. It can obtain $b$ through a controlled $H$ observation, but that procedure costs an operation and changes the state. Any account of its understanding must include those effects rather than identify mathematical existence with operational access.

A practical contract therefore adds an information boundary, a family of recipient or decoder algorithms, resource limits, an intervention catalogue, and a revision policy. The finite theorems below solve particular obligations within that larger contract. They do not fill in unspecified sensors, access rights, or learning procedures.

### 3.2 A profile rather than a single threshold

We propose reporting a profile with separate entries for predictive adequacy, implemented use, intervention compatibility, transfer, recipient reconstruction, correction, and resource cost. Unknown entries remain unknown. A high prediction score must not silently fill the mechanism or correction entries.

The profile can describe a traditional program, a learned model, a hybrid system, or a collective procedure. It does not require language-like internal states. Nor does it establish an ordering of all systems from least to most understanding. Two systems can preserve different families of questions and be incomparable under the declared criteria.

**Counterexample 3.2** (A separate answer for each question is not one reusable record)**.** Let $x = (a, b) \in \{0, 1\}^2$ and let the two queries return $a$ and $b$. If the encoder knows which query will be asked, one bit suffices: retain that answer. If one record must be fixed before the query is selected and must answer both exactly, it needs four distinct values. A one-bit record has only two values and must identify two different answer profiles.

The order of quantifiers is decisive. A different adequate representation for every question is not one adequate representation for their union. This is why a claimed method should be assessed with its encoder, updater, and decoder fixed across the family it promises to support.

## 4 Behavioral compression and the price of new questions

Define

$$x \equiv_C y \quad \Longleftrightarrow \quad q(F_w(x)) = q(F_w(y)) \text{ for all } q \in \mathcal{Q},\ w \in \mathcal{U}^*. \tag{6}$$

This is equality of all declared future answer profiles. It is an equivalence relation. It describes behavioral sufficiency under the contract, not equality of hidden mechanisms or identity outside that contract.

**Theorem 4.1** (The minimal exact reusable state)**.** *The quotient $\mathcal{X}/\equiv_C$ carries an exact reusable representation for $C$. Every other exact reusable representation $r : \mathcal{X} \to \mathcal{Z}$ refines this quotient: if $r(x) = r(y)$, then $x \equiv_C y$. There is a unique map $h : \mathcal{Z} \to \mathcal{X}/\equiv_C$ with $h(r(x)) = [x]$. In particular,*

$$|\mathcal{Z}| \geq |\mathcal{X}/\equiv_C|.$$

*A fixed-length binary state needs at least* $\lceil \log_2 |\mathcal{X}/\equiv_C| \rceil$ *bits. The quotient can be computed by finite partition refinement.*

*Proof.* If $x \equiv_C y$, the empty word gives $q(x) = q(y)$ for every query. Moreover, $F_u(x) \equiv_C F_u(y)$ because any continuation after $u$ is a continuation in the original contract. Thus $[x] \mapsto [F_u(x)]$ and $[x] \mapsto q(x)$ are well-defined updates and decoders.

For any exact representation, the equations in (5) imply equality of all future decoded answers whenever $r(x) = r(y)$. This proves refinement. Surjectivity of $r$ gives existence and uniqueness of $h$ and the cardinality bound. The binary bound follows by counting available codewords.

For computation, let $P_0$ group states with identical current query vectors. Given a partition $P_k$, refine each block by the block of $F_u(x)$ for every $u$. Induction shows that $P_k$ identifies exactly the states agreeing on words of length at most $k$. Every strict refinement increases the block count. At most $|\mathcal{X}| - 1$ strict refinements are possible. At a fixed point the partition is closed under all operations, hence under every finite word, and is exactly (6). □

This is a deterministic state-minimization result, presented in the language of reusable questions. It belongs to the established state-aggregation and predictive-state lineage [4, 3]. Its significance here is not the discovery of quotients. It is the explicit identification of which future distinctions an asserted understanding is required to preserve.

### 4.1 Minimal information is not minimal effort

The theorem minimizes the number of distinct retained states. It does not minimize the work required to acquire the state, compute an update, verify a query, or discover the quotient. A smaller table can have harder entries to compute. A larger representation can make a frequently used transformation local and cheap.

Even uniqueness of the behavioral quotient does not supply a uniquely intelligible notation. The coordinates $(a, b)$ and $(s, b)$ describe the same 25 behavioral classes in the original instrument. One makes the physical register update obvious. The other makes the readout immediate. Their difference is computational and explanatory rather than a difference in information content.

### 4.2 The cost of expanding the contract

Let $C'$ have the same complete state space and contain every operation and query of $C$, while possibly adding more. Write $\equiv_{C'}$ for its finer behavioral equivalence. Take any representation $r$ adequate for $C$. For each $z$, let $k_z$ be the number of $\equiv_{C'}$ classes that meet the fibre $r^{-1}(z)$.

**Proposition 4.2** (Contract expansion and supplementary information)**.** *The relation* $\equiv_{C'}$ *refines* $\equiv_C$*. Any supplementary code* $s : \mathcal{X} \to \mathcal{S}$ *for which* $(r, s)$ *supports exact reusable behavior under* $C'$ *must satisfy*

$$|\mathcal{S}| \geq \max_z k_z.$$

*If every* $\equiv_{C'}$ *class lies within a single fibre of* $r$*, this bound is attained by an exact reusable pair* $(r, s)$*. This condition holds when* $r$ *is the old minimal behavioral quotient. A supplementary value computed solely from* $r(x)$ *cannot repair a fibre meeting more than one expanded class.*

*Proof.* Agreement on the larger continuation family implies agreement on the smaller one. Within a fixed fibre of $r$, an adequate pair $(r, s)$ must separate the $k_z$ new classes, giving the lower bound. For attainment under the additional condition, assign distinct labels $1, \ldots, k_z$ to the expanded classes within each old fibre and let $s(x)$ be its class label. The pair $(r(x), s(x))$ then identifies exactly the expanded behavioral class. The

quotient update and query decoder supply its exact dynamics. Finally, any value computed only from $r(x)$ is constant on that fibre and separates none of its states. □

The attainment condition matters. An old representation can retain irrelevant extra coordinates that cease to be dynamically closed under new operations. For example, start with two bits $(a, b)$, no operations, a constant-zero query, and the unnecessarily detailed record $r(a, b) = a$. Add the swap operation while keeping the query constant. There is still only one expanded behavioral class, so $k_z = 1$ in each old fibre. Yet updating the retained coordinate $a$ after the swap requires $b$. The information lower bound for answers is not sufficient to preserve arbitrary redundant coordinates. The minimal old quotient has no such problem.

In the instrument, each old $(a, b)$ class contains five expanded classes. An additional five-valued symbol is sufficient and necessary. A separately stored fixed-length symbol therefore needs three bits. This should not be confused with subtracting the rounded bounds for a joint 25-state and 125-state encoding, which are five and seven bits respectively. Separate buffers and joint recoding are different coding contracts.

**Counterexample 4.3** (A finite continuation test leaves an unbounded alternative)**.** For any fixed $h \geq 0$, a one-operation machine can return zero through the first $h$ steps and return one at step $h + 1$. A constant-zero machine agrees with it on all those earlier observations. Without a state bound or another restriction on the hypothesis family, success through horizon $h$ cannot establish agreement at every future horizon.

This does not invalidate finite certification of a supplied finite model. It distinguishes exhaustive checking of a known transition system from identifying an unrestricted unknown system by a finite record.

## 5 Available information is not implemented understanding

The finite quotient answers an information question. A claim about a particular computational reasoner must answer an additional question: where is the useful representation realized, and how does it affect the computation?

A representation can contain a recoverable fact without the system using that fact. Conversely, a system can use information through a distributed or redundant encoding that a chosen simple probe fails to recover. The comparison of information structures in statistical decision theory provides a useful background for asking what a record could support [13]. That question is different from identifying the route by which one implemented decision was made.

### 5.1 A carrier can be readable and irrelevant

Consider a solver state $m = (a, b, c, z)$, with $z = a + b$. Suppose its output circuit reads $a$ and $b$ directly and ignores $z$. A decoder applied to $z$ predicts the output perfectly. Replacing $z$ while keeping the actual circuit inputs unchanged has no effect. The code $z$ is available to an observer but is not the operative route used by this solver.

Now alter the implementation so that the output reads $z$, while a repair circuit recomputes $z$ from $a, b$ before every output. An attempted replacement may again have no observable effect, but for a different reason: the affected carrier is restored. A null intervention cannot identify which of these implementations is present unless its timing and the remaining information routes are controlled.

These examples justify neither a universal ablation test nor a demand that every useful representation be uniquely necessary. They show why an intervention must specify what changes, what remains accessible, and what the hypothesis predicts afterward. A proof about the instrument's state does not substitute for an intervention on the solver's implementation.

**Proposition 5.1** (The identification limit of a transcript)**.** *Suppose two proposed implementations induce complete transcript laws $P_0$ and $P_1$ on a finite transcript alphabet under a specified permitted experiment,*

*including any adaptive choices made by its evaluator. With equal prior probability, the greatest identification accuracy available from that transcript is*

$$\frac{1}{2}\big(1 + d_{\mathrm{TV}}(P_0, P_1)\big).$$

*In particular, identical transcript laws permit no accuracy above* $1/2$.

*Proof.* Let $d(o) \in [0, 1]$ be the probability of guessing implementation 0 after transcript $o$. Accuracy is

$$\frac{1}{2}\sum_o \big(P_0(o)d(o) + P_1(o)(1 - d(o))\big) = \frac{1}{2} + \frac{1}{2}\sum_o (P_0(o) - P_1(o))d(o).$$

The sum is maximized by setting $d(o) = 1$ on positive differences and zero on negative ones. The positive differences sum to total variation. This choice attains the bound. □

The experiment may include interventions. The proposition says only that the *permitted complete transcript* bounds identification. Adding a new sensor or an admissible internal intervention changes that transcript law and can change the limit.

### 5.2 The interpreter must not secretly do all the work

Take any bijection $p : \mathcal{X} \to \mathcal{Z}$ of a finite state space. Define

$$\bar{F}_u = p \circ F_u \circ p^{-1}, \qquad \bar{q} = q \circ p^{-1}.$$

All exact representation equations hold. The labels in $\mathcal{Z}$ may nevertheless be arbitrary, and the translated updates may require a complete lookup table. Exact representability alone does not explain why the representation is easier to acquire, use, or communicate.

This does not make the construction false. It makes the claim narrow. A table can be a valid finite model. Its interpretive value depends on the allowed questions and costs. The mistake would be to infer that any such recoding identifies a compact or causally privileged concept inside a solver.

We therefore require an *interface account*. The encoder, decoder, intervention implementation, and translator are part of the evaluated system. Their information access is declared, their computation is charged, and any training or target-specific background they receive is recorded. A decoder that is free to solve the whole task is not evidence that the state it reads supplied the method. Control tasks for probes and the critique of unrestricted alignment maps are especially relevant here [8, 9].

Restrictions can concern circuit size, memory, communication rounds, a finite program language, or an explicit family of allowed maps. There is no universally correct restriction. A linear map is not automatically the right representation of a nonlinear concept, and a nonlinear map is not automatically uninformative. The appropriate restriction must be selected with the scientific question, not retrofitted to distinguish preferred results.

### 5.3 A bounded stochastic extension

The distinction between a current answer and a reusable state also survives stochastic dynamics. Let $K_u(x, x')$ be finite transition kernels and let $r : \mathcal{X} \to \mathcal{Z}$ be surjective. Assume the current query values are constant on fibres of $r$.

**Proposition 5.2** (Uniform controlled stochastic reuse)**.** *A single reduced kernel* $\bar{K}_u$ *reproduces every projected one-step law from every complete state if and only if, whenever* $r(x) = r(y)$,

$$\sum_{v : r(v) = z'} K_u(x, v) = \sum_{v : r(v) = z'} K_u(y, v) \quad \textit{for every } u, z'.$$

*With matched initial projected laws, the reduced model preserves the law of every finite sequence of reduced states and queries under any policy depending only on the reduced history.*

*Proof.* A common projected row is necessary by considering point-mass initial states. If the rows agree, use their shared value to define $\bar{K}_u(r(x), z')$. Conditional on any reduced history, the complete state has a distribution supported on the current fibre. Averaging equal projected rows gives the same reduced transition law. Induction yields equality of the finite history laws, including the deterministic query values. A common randomized history policy preserves the same argument after conditioning on its selected action. □

This is the controlled lumpability condition underlying stochastic state aggregation [4]. The policy qualification is essential. A controller that secretly observes a discarded variable can choose different actions in states that the reduced description identifies. Small prediction error on one observed trajectory distribution is not a substitute for the uniform condition.

## 6 Transporting a method rather than matching an answer

An analogy is useful when relations and operations survive the comparison, not simply when two descriptions use similar words. Structure-mapping theory makes relational correspondence central to analogy [11]. We study a more limited mathematical question: when can an already specified method be translated without changing the answers to its declared continuation questions?

Let a source representation have state space $Z$, updates $f_u$, and query decoders $q$. Let a target have state space $W$, updates $g_v$, and decoders $q'$. A map $\tau$ translates operation names. Query pairs $(q, q')$ and a representation map $T : Z \to W$ are fixed in advance. The operation names alone do not establish that the target implements the intended source operation.

**Theorem 6.1** (Exact transport of reusable operations)**.** *Suppose, for every source operation and matched query,*

$$Tf_u = g_{\tau(u)}T, \qquad q'T = q. \tag{7}$$

*Then, for every source state and every finite source word $w$,*

$$Tf_w = g_{\tau(w)}T, \qquad q'(g_{\tau(w)}(Tz)) = q(f_w(z)).$$

*Two such transports compose, with their operation and query correspondences composed as well. No injectivity assumption on $T$ is needed for this forward guarantee.*

*Proof.* The empty word is immediate. If the identity holds for $w$, then

$$Tf_uf_w = g_{\tau(u)}Tf_w = g_{\tau(u)}g_{\tau(w)}T.$$

Induction proves the word identity, and applying the matched decoder proves the answer identity. For a second transport $S$ with operation map $\sigma$, substitution gives $STf_u = h_{\sigma(\tau(u))}ST$ and the corresponding decoder identity. This is the same condition for the composite. □

This is an intertwining argument. Related composition questions for interventional causal models are studied in greater generality by Rischel and Weichwald [14]. Here the theorem guarantees forward answers only from states in the image of $T$ and for translated source operations. It does not guarantee coverage of all target states or all target operations. Surjectivity, reverse translation, and recovery of discarded source distinctions are separate requirements.

### 6.1 The instrument in display coordinates

For the original instrument quotient, let the source state be $(a, b)$ and set

$$T(a, b) = (a + b, b).$$

The target updates in (2) satisfy (7). The source decoder $a + b$ becomes the target's first coordinate. Thus a method can be carried between a register description and a display-centered description without relearning the meaning of every operation sequence.

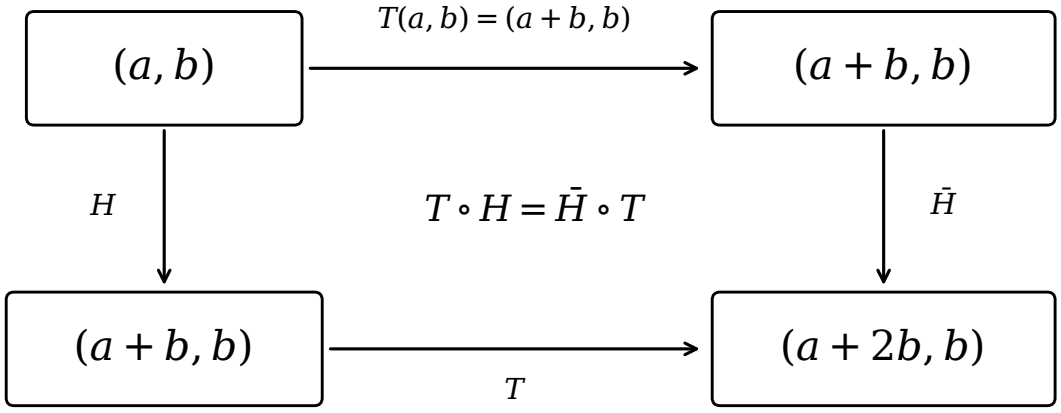



Figure 2: One verified operation square for the instrument. Translating and then applying the target update produces the same state as applying the source update and then translating. The equality concerns the declared $H$ operation, not arbitrary interventions.

**Counterexample 6.2** (Correct current answers with a wrong transported operation)**.** Retain the correct map $T(a, b) = (a + b, b)$ and the correct display decoder. Now incorrectly define the target swap by $\widetilde{W}(s, b) = (s, b)$. The current display and the display immediately after a swap are still correct. Starting from $(a, b) = (1, 3)$, however, the source sequence $W$ then $H$ ends with display zero. The incorrect target sequence ends with display two. Matching a restricted collection of outputs did not preserve the method.

The error becomes visible only when the altered internal state is used by a later operation. That is precisely why reusable understanding requires continuations rather than a list of isolated answers.

### 6.2 Approximate transfer accumulates error

Exact equalities are not always available. For this result let the target state space carry a metric $d$. Suppose each target update is $L$-Lipschitz for a common $L \geq 0$, and suppose

$$d(Tf_u(z), g_{\tau(u)}(Tz)) \leq \epsilon$$

uniformly on every state and operation that the contract permits. Let a target query be $M$-Lipschitz with real-valued output and satisfy $|q'(Tz) - q(z)| \leq \delta$.

**Proposition 6.3** (Finite-horizon transfer budget)**.** *For matched source and target operation sequences, start from $z_0, w_0$ with $d(Tz_0, w_0) \le e_0$. After $n$ operations,*

$$d(Tz_n, w_n) \le L^n e_0 + \epsilon \sum_{j=0}^{n-1} L^j,$$

$$|q(z_n) - q'(w_n)| \le \delta + M\left(L^n e_0 + \epsilon \sum_{j=0}^{n-1} L^j\right).$$

*The empty sum is zero and $L^0 = 1$, including when $L = 0$. At $L = 1$ the state bound is $e_0 + n\epsilon$.*

*Proof.* Writing $e_n = d(Tz_n, w_n)$, insert $g_{\tau(u)}(Tz_n)$ and use the triangle inequality and the Lipschitz bound to obtain $e_{n+1} \le \epsilon + Le_n$. Iterating gives the stated expression. Insert $q'(Tz_n)$ between the two outputs and use the decoder mismatch and its Lipschitz constant. □

The bound concerns the matched operation sequence. It does not automatically cover two controllers that select different operations after slightly different observations. Adaptive choices require a matching policy contract or a separate stability argument for the controller.

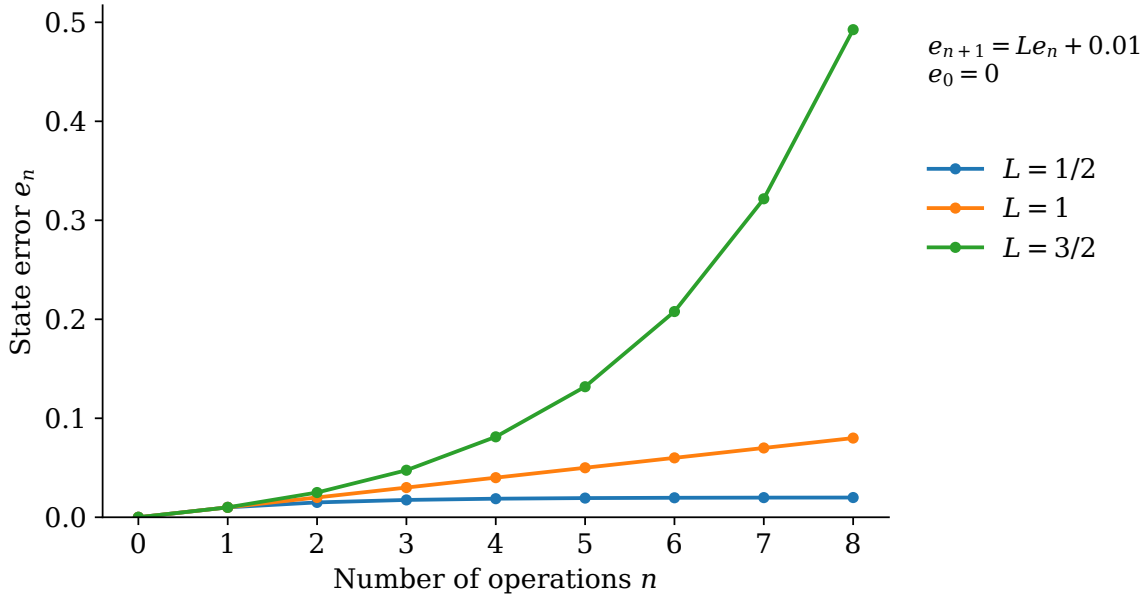



Figure 3: Exact values of the recurrence $e_{n+1} = Le_n + 1/100$, with $e_0 = 0$, illustrate the transfer bound. The curves are analytical constructions, not measured model errors. A small local defect can remain bounded, accumulate linearly, or be amplified.

A map can therefore be accurate for a short continuation and unsuitable for a longer one. The required tolerance belongs to the task, not to a universal declaration that the two representations are close. An approximate relation with one tolerance also need not be transitive. A chain of individually accepted translations may exceed the original error budget.

### 6.3 Local representations form an atlas only with transition rules

One representation may simplify prediction while another simplifies correction or proof. A useful system need not select one notation forever. It can maintain a family of representations with declared domains and translators. We call this an atlas only in the modest sense of a set of local descriptions connected by checked transition maps.

The domain guards matter. A method valid on reversible operations may fail after an irreversible reset. A chart switch may require evidence absent from the current chart. A common output name does not establish a common unit or query semantics. The system must know which transitions are supported and must retain the unresolved discrepancies when a transition is partial.

There is no general theorem here that more representations improve reasoning. Additional charts consume memory and create more opportunities for a mistaken translation. The research target is a family whose additional coverage or lower repeated cost justifies its construction and maintenance. Exact intertwining supplies one part of that analysis, not its economic conclusion.

# Part II

# Explanation, inquiry, and correction

## 7 An explanation belongs to a recipient

An explanation can be a correct description and still fail to help a particular recipient. It may presume unfamiliar notation, omit a necessary background result, or encode all the difficult work in a phrase the recipient cannot execute. Conversely, a deliberately simplified explanation can be useful for a limited question family without being a complete description of the source mechanism.

We therefore distinguish two contracts. A *subject explanation* enables the recipient to reason correctly about specified aspects of the subject. A *process explanation* additionally makes a faithful claim about how a particular solver reached its result. A process explanation must satisfy an implementation attribution as well as a subject-level or transcript-level account. Neither fluent presentation nor a valid derivation alone establishes the attribution [2].

**Definition 7.1** (Recipient-relative explanatory adequacy)**.** Fix a recipient procedure $L$, background information $B$, resource allowance, and continuation family. An artifact $E$ is adequate for a target model $m$ when $L(B, E)$ yields a representation and operations that answer the required continuations for $m$ within the declared error and resource limits. The artifact's inputs, the recipient's prior access, and any target-specific background are included in the contract.

This is an operational property of an artifact and a recipient. It does not equate instruction with human comprehension, and it does not require natural language. A diagram, proof, worked example, executable transformation, or combination can play the role of $E$.

### 7.1 An exact communication lower bound

Let $\Theta$ be a finite family of target models. Relative to the required continuation questions, divide these models into $N$ classes with distinct answer profiles. Use one fixed recipient background and one fixed deterministic recipient. The teacher can select a message but cannot change the recipient's background separately for each target without that change being counted as communication.

**Proposition 7.2** (The information an exact explanation must carry)**.** *If a fixed-length $b$-bit artifact lets the recipient answer the declared questions exactly for every target in $\Theta$, then $2^b \geq N$. With target-dependent background $B_\theta$, the same argument gives*

$$2^b \geq \max_B \; N_B,$$

*where $N_B$ counts distinct answer profiles among targets sharing that exact background value.*

*Proof.* Two different required answer profiles cannot produce the same background and message. The deterministic recipient would then make the same responses for both and fail on a distinguishing question. There are at most $2^b$ fixed-length messages. Apply this counting argument to the whole family, or separately within each background fibre. □

The bound is about information, not explanatory elegance. It does not establish that any shortest message is easy to understand or that the recipient can decode it efficiently. A compressed index into an enormous shared table can meet the bound while transferring little new method. The table's construction and availability must be charged to the background account.

### 7.2 What the instrument explanation transfers

For the original instrument, an adequate artifact can specify arithmetic modulo five, the representation $(s, b)$, the two update rules in (2), the output decoder, and the domain restriction excluding $J$. A recipient supplied with an initial $(s, b)$ can then answer every $H/W$ continuation by applying these operations.

That establishes a subject explanation for the specified recipient. It does not establish that a neural system that produced the artifact used $(s, b)$ internally. A separate process explanation might show that the system's actual intermediate states realize these coordinates and respond to corresponding interventions. These two certificates can support one another, but they answer different questions.

An especially useful explanation includes the failure boundary. The line "the third register is irrelevant" is too broad. The line "the third register does not affect any display reachable through the current $H/W$ operation family" is exact. The latter statement tells the recipient what should be reconsidered when a new operation arrives.

### 7.3 Explanations can preserve methods without reproducing discovery

A proof may give a shorter route than the route by which the result was found. A teacher may introduce coordinates that the discoverer never used. These are legitimate explanatory choices when their subject-level claims are correct. They become misleading only when presented as faithful records of the original reasoning without evidence.

This distinction avoids an unnecessary demand that every helpful explanation be an autobiography of a computation. It also prevents the opposite mistake of treating a helpful reconstruction as proof of the original mechanism. Human context and communicative purpose remain important beyond the formal recipient criterion [10].

## 8 One explanation for recipients with different backgrounds

An explanation can be adequate for each recipient separately without there being one short artifact that works for all of them. Conversely, a shared artifact can be shorter than a complete description because recipients already hold complementary information. The relevant object is the combination of sender access, communicated record, and recipient background. Counting only the words in the artifact misses two thirds of that object.

The finite problem below belongs to zero-error communication with side information. Witsenhausen related side-information coding to graph coloring [15], while broadcasting research studies a common message used by several informed receivers [16]. We adapt that established construction to explicit sender access and recipient-specific demands. The result is a communication criterion within an explanation contract, not a new definition of comprehension.

### 8.1 An accessible sender and a family of recipients

Let $\mathcal{M}$ be a finite nonempty set of possible cases. The sender observes only $\alpha(m)$, where $\alpha : \mathcal{M} \to S$ is a declared access map. Recipient $i \in \{1, \dots, k\}$ receives background $\beta_i(m)$ and must recover $d_i(m)$. A demand can be a single answer, a reusable state, or the complete finite answer profile for a declared task. The maps and recipients are fixed before the case is selected.

One encoder $E : S \to W$ produces a common artifact. A decoder $D_i$ for each recipient must satisfy

$$D_i(E(\alpha(m)), \beta_i(m)) = d_i(m) \qquad \text{for every } m, i. \tag{8}$$

The encoder is not allowed to consult a recipient-specific hidden state. Recipients do not receive an uncharged transcript or a separate description of the selected case. Here decoders may be designed once for the declared

finite contract and then remain fixed. If an existing interpreter is prescribed, the criterion below is only a necessary information condition unless that interpreter can implement the constructed decoder.

Build a graph $G_\alpha$ on $\alpha(\mathcal{M})$. Join $s$ and $t$ when some cases $m, n$ and recipient $i$ satisfy

$$\alpha(m) = s, \quad \alpha(n) = t, \quad \beta_i(m) = \beta_i(n), \quad d_i(m) \neq d_i(n).$$

An edge can be a loop when $s = t$. It then records a distinction that neither the sender's access nor that recipient's background can resolve.

**Theorem 8.1** (A common explanatory artifact under sender and recipient constraints)**.** *An exact artifact satisfying* (8) *with at most $K \geq 1$ symbols exists if and only if $G_\alpha$ has no loop and admits a proper coloring with at most $K$ colors. In the loop-free case, the minimum alphabet size is $\chi(G_\alpha)$, and the minimum fixed binary length with designable decoders is $\lceil \log_2 \chi(G_\alpha) \rceil$. A loop makes the exact contract infeasible for every alphabet size.*

*Proof.* An encoder assigns a symbol to every accessible sender value. If adjacent vertices receive the same symbol, their witnessing recipient receives the same artifact and background in two cases requiring different outputs. No deterministic decoder can be correct in both. Thus every valid encoder is a proper coloring, and a loop is impossible.

Conversely, use a proper coloring as the encoder. For any recipient and any artifact/background pair that actually occurs, assign the demanded output from one compatible case. Another compatible case cannot demand a different output, since it would witness an edge whose endpoints have the same color. Thus the decoder is well defined. Define it arbitrarily on unattainable pairs. Counting binary strings proves the final bound. □

The theorem provides a smallest *alphabet*, not the cheapest explanation. Its decoders can be large tables. Their construction, runtime, background acquisition, and semantic interpretation remain part of the resource account. The graph is useful precisely because it does not pretend that a color label is intrinsically meaningful.

### 8.2 Complementary knowledge can reduce communication

For two bits $(a, b)$, let the first recipient know $a$ and need $b$, while the second knows $b$ and needs $a$. A sender with access only to $a \oplus b$ can broadcast that one bit. Both recipients recover their missing bit. The sender need not reconstruct the complete pair. An inference from successful explanation to possession of the recipient's full state would therefore be invalid.

This does not make the recipients' knowledge free. Their prior bits are explicitly supplied. Removing them changes the graph and the communication requirement. Nor does the example show that the sender has a useful general theory of the two-bit system. It establishes a specific jointly realizable communication contract.

**Counterexample 8.2** (Individually binary explanations can require three common symbols)**.** Take five cases labeled $0, \ldots, 4$ and five recipients. Recipient $i$ cannot distinguish cases $i$ and $i + 1$ modulo five from its background, but distinguishes every other case individually. It demands zero on case $i$ and one on case $i + 1$, with arbitrary fixed demands elsewhere. The sender knows the case. Each recipient separately needs only two artifact symbols. The common confusability graph is a five-cycle. It is not two-colorable, but the sequence of colors $A, B, A, B, C$ is proper. One common artifact therefore needs three symbols, or two fixed binary bits.

The obstruction is global: the largest ambiguity faced by any one recipient does not determine the size of a shared explanation. The same issue appears when one artifact is expected to support a novice, a specialist, and a verifier who each possess different background information. The finite theorem tells us what must be distinguished. It does not prescribe how a human-facing explanation should express those distinctions.

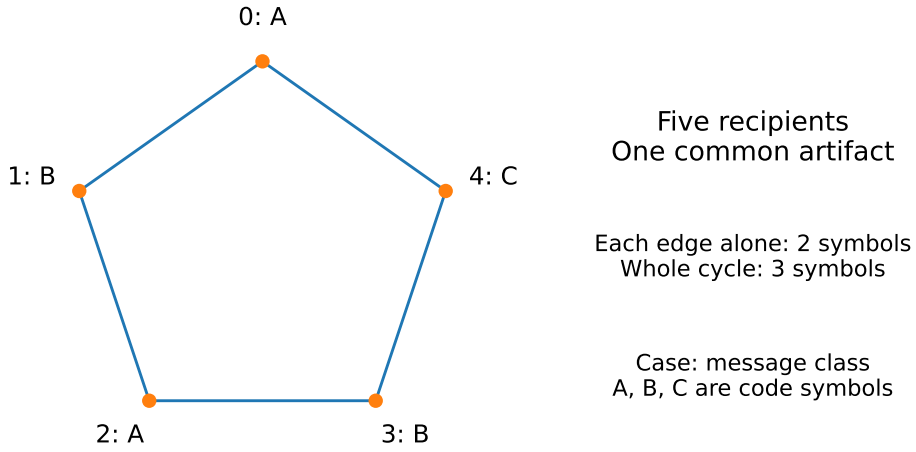



Figure 4: Each edge records one recipient's unresolved pair of cases. Each recipient alone can use a binary message. Their combined graph needs three symbols. Letters denote message classes, not empirical clusters or intrinsic semantic categories.

### 8.3 Source restrictions are substantive

If two cases have the same sender observation and the same recipient background but require different outputs, adding more fluent text cannot fix the information deficit. The graph has a loop. The options are to acquire evidence, change the recipient background, relax the demand, or report that the current contract cannot be met. Calling a more powerful decoder changes none of those facts if its permitted inputs remain identical.

## 9 Learning where the current representation fails

A reusable representation should support correction without treating every new observation as a reason to discard everything. A valid method can have a restricted domain. A recorded premise can be false. A previously omitted variable can become relevant. These failures call for different revisions.

We distinguish three cases. In a *state correction*, the operations remain valid but the current state estimate changes. In a *model correction*, an assumed transition or observation law changes. In a *contract expansion*, the requested operations or questions change. A representation can be entirely adequate before the third kind of change and inadequate afterward, as the instrument shows.

### 9.1 A finite model of informative correction

Let $\mathcal{M}$ be a finite nonempty set of deterministic candidate models. Let $\mathcal{E}$ be a finite set of permitted experiments. Each experiment $e$ has an outcome map $f_e : \mathcal{M} \to O_e$. The current information is a candidate set $V \subseteq \mathcal{M}$. On observing $o$ from $e$, update it by

$$C_{e,o}(V) = \{m \in V : f_e(m) = o\}. \tag{9}$$

This is consistency filtering, not Bayesian inference and not a general model of scientific reasoning. An empty set is meaningful: the observation contradicts every currently retained candidate. It must not be replaced by an invented best answer.

An experiment is informative for $V$ when its outcome map is not constant on $V$. Models indistinguishable under every permitted experiment form an observational equivalence class. The aim can be to identify such a class rather than a uniquely named model.

**Proposition 9.1** (Finite correction under separating evidence)**.** *Suppose the true model $m_*$ belongs to the initial set $V_0$, every observation is exact, and the allowed experiment family separates any two distinct remaining observational classes. Repeatedly selecting an informative experiment preserves $m_*$ and reaches a single observational class after at most $|V_0| - 1$ informative observations. If there are $N$ distinguishable target classes, every adaptive procedure identifying all of them using experiments with at most $b \geq 2$ outcomes and at most $h$ observations must satisfy $b^h \geq N$.*

*Proof.* The true model predicts each exact observation, so filtering never removes it. An informative experiment splits the current set into at least two nonempty outcome classes. Whatever true outcome occurs, at least one candidate is removed. The nonempty set can strictly shrink at most $|V_0| - 1$ times. If two observational classes remain, the separating assumption supplies another informative experiment. For the second claim, an adaptive decision tree of depth at most $h$ and branching at most $b$ has at most $b^h$ leaves. Distinct required classes must reach different leaves. □

This elementary result shares the query-and-counterexample perspective of classical active learning [12]. It does not give an efficient method for discovering the candidate family, finding an inexpensive separating experiment, or dealing with inaccurate observations. Its upper bound is a count of informative observations, not a bound on wall-clock work or physical cost.

### 9.2 A question can be more useful than another answer

If all surviving candidates give the same answer to the next task, the system can act without identifying which candidate is true. If they disagree on a consequential operation, an experiment that distinguishes that disagreement may be more valuable than repeating familiar tasks on which all candidates agree.

The criterion is task-relative. Two models can differ on an inaccessible quantity and still support all permitted decisions. Conversely, a rarely asked question can distinguish models whose disagreement controls whether a future intervention is safe. The system should retain which distinctions have been settled and which have merely been irrelevant so far.

A request for evidence is also an operation with a realization. It may consume resources, change the system, or reveal only a biased part of the state. The finite filtering model presumes that the stated outcome maps already include those effects. Section 10 makes current-state changes and their cost explicit. For a changing instrument, one must use a controlled reset, a model of the full observation history, or a state update that tracks what the experiment changed. Independent repeated access to the same initial state is not free.

### 9.3 Repairing the running example

Before $J$ is admitted, all five possible values of $c$ are equivalent under the base contract. Treating them as one class is correct, not ignorance that the old task required the predictor to resolve. After $J$ is admitted, these classes split. The system can report that its old state no longer determines the answer.

The current record supplies $(s, b)$. Executing $J$ and observing $s_J$ determines $c = s_J - s$, while (4) updates the affected first register. The corrected extended coordinates are $(s_J, b, c)$. Their operations are

$$\bar{H}(s, b, c) = (s + b, b, c), \quad \bar{W}(s, b, c) = (s, s - b, c), \quad \bar{J}(s, b, c) = (s + c, b, c).$$

The base method survives as the restriction of the enlarged method. Only the newly relevant distinction is acquired.

A different kind of failure would occur if the instrument secretly implemented $H(a, b, c) = (a + 2b, b, c)$. That would contradict the old transition model rather than merely extend the operation catalogue. A correction record should say which assumption changed. Otherwise a sequence of patched answers can conceal a permanently incorrect account of the mechanism.

### 9.4 Revision must preserve its own uncertainty

The true-model assumption in Proposition 9.1 is a substantive premise. When it is not justified, the system must allow the possibility that every candidate is wrong. When observations are noisy, deleting a model after one conflicting record can remove the best explanation. The statistical treatment in Sections 12–14 supplies exact and conditional results once a likelihood model and selection process have been declared. It does not establish their empirical calibration.

The finite framework nevertheless gives a useful discipline. A correction should identify the new evidence, the rejected candidates, the surviving claims, and the assumption linking the observation to the model. It should not rewrite an old uncertainty as though the corrected answer had always been established.

## 10 Acquisition as an experiment planning problem

A lower bound on the information a representation must hold is not yet a plan for acquiring it. Even a perfectly informative experiment can be unavailable, expensive, or destructive. We now make that acquisition problem exact for a finite deterministic model.

Belief-state planning has a substantial foundation in partially observed control [17]. The construction here uses sets of possible current states rather than probability distributions, a worst-case cost rather than expected reward, and an explicitly bounded horizon. It is a finite dynamic program for a stated contract, not a general solution to scientific discovery.

### 10.1 Experiments change both information and state

Let $X$ be a finite nonempty state space. Each experiment $e$ has an admissibility set $A_e \subseteq X$, a transition $T_e : X \to X$, an observation $o_e : X \to O_e$, and a known cost $c_e \geq 0$. The observation is indexed by the state before the experiment, but can describe a reading taken afterward: for example $o_e = q \circ T_e$. At an information state $V \subseteq X$, the experiment is permitted only when $V \subseteq A_e$. After outcome $o$, the new support is

$$V_{e,o} = \{T_e(x) : x \in V,\ o_e(x) = o\}. \tag{10}$$

Only nonempty branches are possible. This update tracks the world after the intervention, not just a narrowed list of old states.

Let $r : X \to Z$ be the target record. A support is terminal when $r$ is constant on it. Let $h$ be the number of experiments still available. Write $\mathcal{V}_h(V)$ for the least worst-case sum of experiment costs among policies that terminate with the target known within $h$ steps. Infeasibility is denoted $+\infty$. Then define

$$\begin{aligned} \mathcal{V}_h(V) &= 0 && \text{if } |r(V)| = 1, \\ \mathcal{V}_0(V) &= +\infty && \text{otherwise,} \\ \mathcal{V}_h(V) &= \min_{e : V \subseteq A_e} \left( c_e + \max_{o : V_{e,o} \neq \varnothing} \mathcal{V}_{h-1}(V_{e,o}) \right) && \text{otherwise,}\ h > 0. \end{aligned} \tag{11}$$

An empty minimum is $+\infty$. The terminal case permits stopping immediately, so an unnecessary additional experiment is never forced.

**Theorem 10.1** (Exact finite acquisition with state-changing observations)**.** *For every nonempty support $V$ and finite horizon $h$, recurrence* (11) *equals the minimum worst-case acquisition cost among deterministic*

*adaptive policies with that horizon, access boundary, and current-state target. Every finite value has an attaining policy obtained by choosing a minimizing experiment and its recursively attaining branches. The value is infinite exactly when no such policy exists.*

*Proof.* At horizon zero, a target can be returned without further evidence precisely when it is constant over the current support. This gives the base case. At a nonterminal support with $h > 0$, the first experiment must be admissible at every possible state. After its observed outcome the possible current states are exactly (10). By induction, the best possible remaining worst-case cost on that branch is $\mathcal{V}_{h-1}$. Every nonempty branch can occur, giving the maximum over branches. Minimizing over the finite experiment set gives a lower bound for every adaptive policy. Choosing a minimizing experiment and attaining branch policies achieves that bound whenever it is finite. If every choice has an impossible branch, no guaranteed policy exists. □

The complete information state must include every variable needed for future transitions, admissibility, costs, and the target. A permission depending on an earlier event must be retained as state. Otherwise two histories with the same physical support could admit different experiments, invalidating the recurrence. Nonnegative costs and the finite horizon make zero-cost actions harmless to the definition. No unbounded termination claim is being made.

The state search can be large: there are up to $2^{|X|} - 1$ nonempty supports, before accounting for additional memory variables. The theorem is an exact finite characterization, not a polynomial-time learning guarantee. The delivered reference implementation memoizes supports and remaining horizons. Its companion check compares the result with a separately implemented enumeration of complete observation-history policies on small systems.

### 10.2 The diagnostic operation is a paid transformation

In the instrument, take a known base record $(s, b)$ and let the five possible values of $c$ define the initial support. Set the cost of $J$ to two units, and the cost of $H$ and $W$ to one unit each. Each experiment returns the exact display after its operation. The target is the enlarged current record $(s, b, c)$.

A $J$ observation resolves all five possibilities and costs two. An $H/W$ sequence cannot reveal $c$, so no policy costing less than two can meet the target. The optimum is therefore two, attained by executing $J$ and retaining the post-operation record. This example fixes its sensor and operation catalogue. It does not claim that every physical diagnostic with five outcomes has the same cost.

### 10.3 Knowing the current state is not recovering its history

**Counterexample 10.2** (A reset supplies certainty without recovering the past)**.** Let the current state be an unknown bit. The only experiment sets the bit to zero and returns a constant symbol. One experiment makes the current bit known with certainty. It supplies no information about whether the old bit was zero or one. Thus an acquisition policy for the current-state target can succeed by changing the world, not by identifying its original condition.

To require recovery of a past distinction, replace the state by $(x_0, x_t)$, let every transition preserve $x_0$, and make the terminal target include the required part of $x_0$. The same recurrence then treats historical reconstruction honestly. In the reset example, the two origins remain indistinguishable at every horizon. If both current knowledge and preserved history matter, use their joint target rather than claim that either implies the other.

For the instrument, the diagnostic $J$ is invertible once $c$ is known, so the exact observation can support both records. That property comes from the specified operation. It cannot be transferred to an arbitrary experiment.

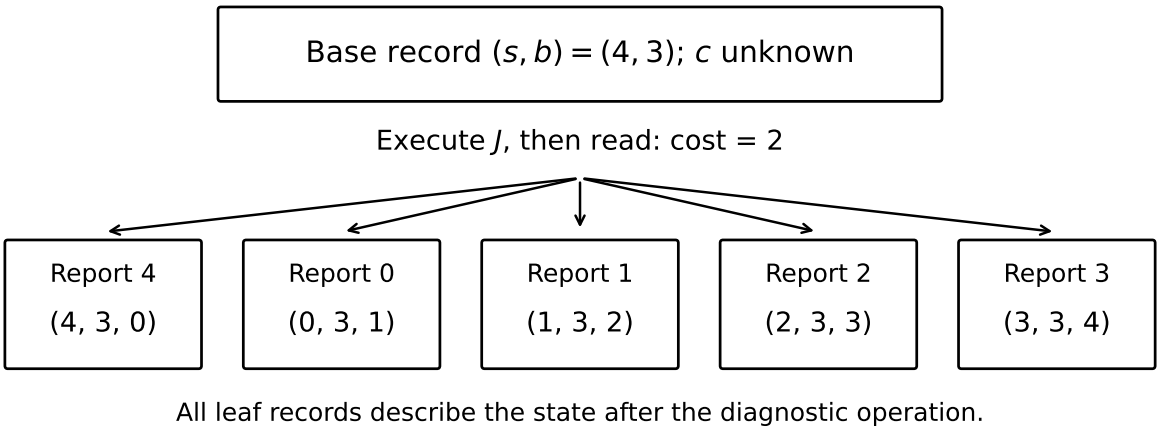



Figure 5: A single $J$ diagnostic from base record $(s, b) = (4, 3)$ has five possible exact reports. Each report determines a different *post-operation* record. The cost is charged to the operation, and the hidden value is acquired rather than supplied to the decoder.

Likewise, a planner that chooses an informative experiment is not thereby credited with discovering its transition law or building its sensor.

### 10.4 Experiment value is relative to the intended continuation

An experiment need not distinguish every candidate model to be useful. It need only resolve distinctions that control the promised future behavior. Conversely, acquiring a value that changes none of the relevant decisions can waste resources even when it reduces uncertainty. Information-based experimental design has a long history [18]; the present cost criterion deliberately asks a different, exact question about a specified target and worst-case completion.

The horizon is a contractual limit, not an arbitrary truncation of a failed computation. A negative result at horizon $h$ says that the declared task cannot be guaranteed within that many permitted experiments. Increasing the horizon changes the problem. It does not retroactively validate a promise made at the smaller horizon.

## 11 When translation commutes with correction

A method can travel through a language that preserves its current answers while losing the evidence needed to improve it. This is a different obstruction from a wrong update equation. It concerns the interaction between two otherwise sensible operations: translating a model and filtering it against evidence.

Let $\mathcal{M}$ and $\mathcal{N}$ be finite model families, with model translation $\phi : \mathcal{M} \to \mathcal{N}$. For one source experiment $e$, let $f : \mathcal{M} \to O$ give its outcomes. A corresponding target experiment has outcome map $g : \mathcal{N} \to O'$. Let $\eta : O \to O'$ translate evidence. Assume the basic observation square commutes:

$$g(\phi(m)) = \eta(f(m)) \quad \text{for all } m \in \mathcal{M}. \tag{12}$$

This condition says that a model's predicted evidence translates correctly. It permits $\eta$ to merge source outcomes.

Write $\phi[V] = \{\phi(m) : m \in V\}$. Let $C_o(V) = \{m \in V : f(m) = o\}$ and $C'_{o'}(W) = \{n \in W : g(n) = o'\}$. Both are ordinary consistency filters.

**Theorem 11.1** (Correction transport and evidence reflection)**.** *Under* (12)*, for every $V \subseteq \mathcal{M}$ and every attainable source outcome $o \in f(\mathcal{M})$,*

$$\phi[C_o(V)] \subseteq C'_{\eta(o)}(\phi[V]). \tag{13}$$

*Equality holds for all such $V$ and $o$ if and only if $\eta$ is injective on the attainable outcome set $f(\mathcal{M})$. The claim includes candidate sets contradicted by the observation, so that preservation of an empty update is part of the contract.*

*Proof.* If $n \in \phi[C_o(V)]$, then $n = \phi(m)$ for some $m \in V$ with $f(m) = o$. Equation (12) gives $g(n) = \eta(o)$, proving inclusion.

If $\eta$ is injective on $f(\mathcal{M})$ and $n \in C'_{\eta(o)}(\phi[V])$, choose $m \in V$ with $\phi(m) = n$. Then $\eta(f(m)) = g(n) = \eta(o)$, so $f(m) = o$. Thus $n \in \phi[C_o(V)]$, proving equality.

Conversely, suppose two distinct attainable outcomes $a, b$ have $\eta(a) = \eta(b)$. Choose $m$ with $f(m) = b$, let $V = \{m\}$, and consider source observation $a$. The source update is empty, whereas the target update retains $\phi(m)$ by (12). Uniform equality therefore fails. □

This is an exact set-theoretic result about direct images and filtering. The institutional or cognitive interpretation does not make it a new general theorem of set theory. Its value is to expose a condition that output translation alone omits: the translated evidence must preserve the distinctions by which a candidate is excluded.

The necessity statement deliberately tests observations that contradict the current candidate set. If those cases are excluded by a standing assumption that the current model is already correct, weaker conditions can suffice. That is a different and less demanding correction contract. A reasoner meant to discover that its model is wrong should not assume away the contradictory case.

**Counterexample 11.2** (A translated observation preserves both mistaken candidates)**.** Take $\mathcal{M} = \{m_0, m_1\}$ with $f(m_0) = 0$ and $f(m_1) = 1$. Translate models bijectively to $\{n_0, n_1\}$, but translate both observations to a single symbol $*$. Let $g(n_0) = g(n_1) = *$. The observation square commutes. Starting from both candidates, source observation 1 leaves only $m_1$, whose image is $\{n_1\}$. Translating first and then filtering by $*$ leaves both $n_0$ and $n_1$.

The failure is not an inaccurate target answer in its coarse language. Both models really do predict $*$ there. The failure is the loss of a previously meaningful ability to discriminate. This is why a representation that can communicate conclusions may still be an inadequate carrier of scientific revision.

### 11.1 The instrument explains the boundary

For base experiments built only from $H$, $W$, and the display, the translation from $(a, b, c)$ to $(a, b)$ can preserve evidence without collapsing outcome symbols. Different values of $c$ produce the same permitted observations, so their identification loses no base correction power.

For $J$, that translation cannot preserve the full outcome map at all: two states with identical $(a, b)$ can have different readings after $J$. One could force a commuting square by translating all new readings into one coarse symbol. That would preserve the trivial coarse prediction, but it would destroy the ability to discover $c$. Enlarging the representation is a different response from weakening the question until the old representation appears adequate.

This distinction is important for evaluation. A system should not be rewarded for transfer by silently replacing a difficult target question with an easier translated one. The source and target question contracts, including evidence reflection when correction matters, must be fixed before the comparison.

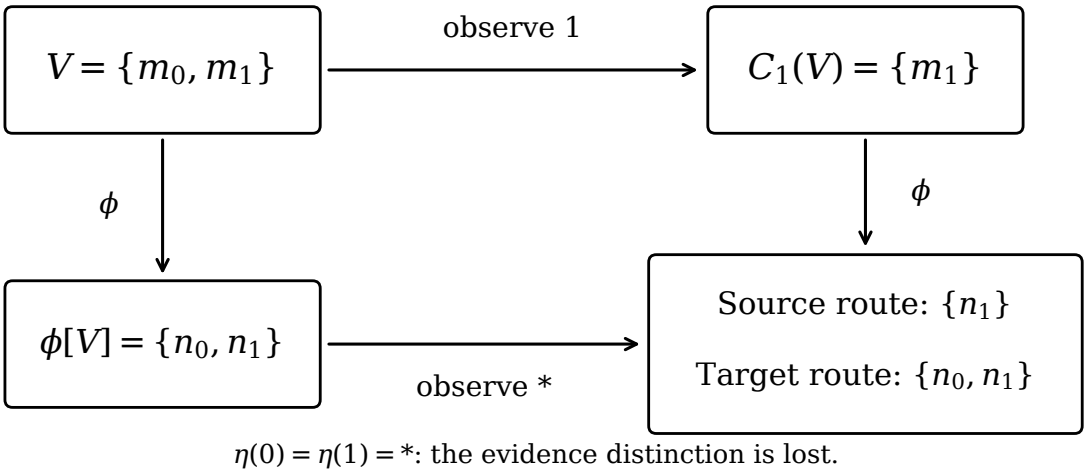



Figure 6: Correct translation of each predicted observation need not preserve correction. The lower evidence channel merges zero and one, so the target cannot remove the candidate excluded at the source. The theorem specifies when the inclusion becomes equality.

### 11.2 Composing corrective explanations

Suppose an explanatory artifact carries a recipient's candidate set, translated experiments, and evidence interpretation. If each translation preserves the observation square and reflects the relevant outcomes, successive filters and translations commute by repeated application of Theorem 11.1. The recipient can then discard the same translated candidates as the source after the specified observation sequence.

This conclusion is only about filtering the represented model families. It does not transfer an ability to invent a missing model, establish that the observations are trustworthy, or guarantee efficient implementation. An explanation can teach how to correct within a family while leaving the enlargement of that family as a separate problem.

## 12 Noisy evidence and Bayesian correction

Exact filtering removes a candidate as soon as an observation contradicts it. That is appropriate for the deterministic evidence contract in Section 11. A noisy observation can instead be unlikely under a model without being impossible. Reusing the exact filter in that setting can discard a well-supported model because a sensor behaved exactly as its noise model allowed.

Let $M \in \mathcal{M}$ be a fixed unknown model in a finite family, with prior $p(m)$. An experiment produces $O \in \mathcal{O}$ with likelihood $L_m(o)$. A common reporting channel $B(o, z)$ produces $Z \in \mathcal{Z}$. Each likelihood row and each channel row is a probability distribution. The complete joint model is

$$\Pr(M = m, O = o, Z = z) = p(m)L_m(o)B(o, z). \tag{14}$$

In particular, the channel does not secretly inspect $M$. Define

$$Q_m(z) = \sum_o L_m(o)B(o, z), \qquad p^o(m) = \frac{p(m)L_m(o)}{\sum_n p(n)L_n(o)}, \qquad p^z(m) = \frac{p(m)Q_m(z)}{\sum_n p(n)Q_n(z)}.$$

A posterior is defined only when its denominator is positive. An event assigned probability zero by the entire current model has no posterior under that model. It is a model-failure signal, not permission to invent a probability distribution.

This section retains the same model identities at source and target so that it isolates evidence loss. Merging model identities is a separate operation. A target likelihood on a merged class is independent of the allocation of prior mass within that class only when the relevant likelihood is constant there, or when the within-class prior itself is retained as additional state. That assumption must not be concealed in a notation change.

### 12.1 When a coarser report preserves every update

The reporting model implies $M \to O \to Z$. Therefore, on every positive-probability pair $(o, z)$, the full posterior after observing both records equals $p^o$. The coarse posterior is a mixture of these fine posteriors:

$$p^z = \sum_o \Pr(O = o \mid Z = z)\, p^o. \tag{15}$$

Averaging is not generally the same as retaining each possible update. The distinction is classical statistical sufficiency, whose factorization foundations include Halmos and Savage [19]. We state and prove the finite channel criterion used here.

**Theorem 12.1** (Exact Bayesian correction through a reporting channel)**.** *For fixed finite likelihoods $L$ and reporting channel $B$, the following are equivalent:*

*(i) For every prior $p$, $p^o = p^z$ on every positive-probability pair $(o, z)$.*

*(ii) This equality holds for one prior assigning positive mass to every model.*

*(iii) There exists a row-stochastic kernel $\rho(o \mid z)$ such that, for all $m, o, z$,*

$$L_m(o)B(o, z) = Q_m(z)\rho(o \mid z). \tag{16}$$

*Thus a report preserves every posterior update exactly when the discarded observation can be conditionally reconstructed from that report without knowing the model, in the actual joint experiment.*

*Proof.* The first condition implies the second. Suppose the second holds under a fully supported prior. For each $z$ of positive probability set $\rho(o \mid z) = \Pr(O = o \mid Z = z)$. If $(o, z)$ has positive probability, posterior equality and Bayes' formula give

$$p(m)L_m(o)B(o, z) = \Pr(O = o, Z = z)p^o(m) = \rho(o \mid z)\Pr(Z = z)p^z(m) = p(m)Q_m(z)\rho(o \mid z).$$

Cancel the positive $p(m)$. If the pair has probability zero, full prior support makes $L_m(o)B(o, z) = 0$ for every model, and $\rho(o \mid z) = 0$. If $z$ is impossible under every model, choose any probability row $\rho(\cdot \mid z)$ since both sides are zero. This proves the third condition.

Conversely, the factorization makes the likelihood of $(o, z)$ a model-independent multiple of the likelihood of $z$. Normalization under any prior with positive pair probability therefore gives $p(M \mid o, z) = p^z$. The common reporting channel gives $p(M \mid o, z) = p^o$. This proves the first condition. □

For deterministic reporting $z = \eta(o)$, the criterion says that the nonzero vectors $(L_m(o))_{m \in \mathcal{M}}$ within each report fibre must be proportional. An ancillary symbol may be discarded even when the report is not injective. In the deterministic evidence model, distinct attainable outcomes have nonzero likelihood vectors with disjoint supports. Those vectors cannot be proportional, recovering the injectivity requirement for that special case. Noise therefore changes the correct preservation test, not merely the notation used for filtering.

The kernel $\rho$ is conditional reconstruction for the specified coupling $(M, O, Z)$. The existence of some unrelated procedure simulating an experiment from another record is a different assertion. Neither assertion shows that the real sensor has the stated likelihood or that a learner inferred it correctly.

### 12.2 An exact measure of the lost update

A coarser report may fail the criterion while still retaining much of the useful evidence. There is a natural average loss, though it should not be mistaken for a worst-case bound on every possible observation. For finite distributions use natural logarithms and the convention $0 \log 0 = 0$.

**Proposition 12.2** (Expected posterior loss is conditional information)**.** *Under* (14)*,*

$$\mathbb{E}_{O,Z}\left[D_{\mathrm{KL}}(p^O \| p^Z)\right] = I(M; O \mid Z) = I(M; O) - I(M; Z). \tag{17}$$

*The common value is nonnegative. It is zero exactly when the two posteriors agree on every positive-probability observation pair for that prior. For a fully supported prior, zero is equivalent to the factorization in Theorem 12.1.*

*Proof.* On the support of the joint law, $p(m \mid o, z) = p^o(m)$. Substituting this into the finite definition of conditional mutual information gives the expected divergence. Expanding its logarithm as $\log(p^o(m)/p(m)) - \log(p^z(m)/p(m))$ and summing gives the difference of mutual informations. Terms with zero prior mass carry no weight. Every positive-weight numerator has a positive denominator, so the expression is finite.

For probability distributions, $D_{\mathrm{KL}}(a\|b) \geq 0$, with equality only at $a = b$. This follows from $\log t \leq t - 1$ on the common support, with missing support only increasing the divergence. A finite weighted sum of nonnegative terms is zero exactly when every positive-weight term is zero. The final assertion follows from Theorem 12.1. □

This is the standard information chain rule in the present correction contract, not a new entropy identity. It makes the object of the lost information explicit: the update over the chosen model family. That family may include irrelevant distinctions or omit the true mechanism. A small or zero value does not independently validate the family.

### 12.3 The noisy instrument and a harmless discarded flag

Return to the instrument immediately after a known $J$ intervention. The earlier $(s, b)$ is known, while $c$ is initially uniform on five values. A read-only sensor reports the shifted value $Y$ = reported display $- s$ modulo five. It returns the correct $c$ with probability $3/4$ and each wrong value with probability $1/16$. The physical state remains fixed during repeated read-only observations.

Suppose the raw report also contains a binary flag with probabilities $1/3$ and $2/3$, independent of $c$ and of $Y$. Removing that flag satisfies (16) exactly. Both recipients retain the same posterior over $c$ and hence the same distribution over the post-$J$ state $(s + c - b, b, c)$. The flag does not carry corrective information for this family.

**Counterexample 12.3** (A compact report can destroy the useful noisy distinction)**.** In the same instrument, replace $Y \in \{0, 1, 2, 3, 4\}$ by its ordinary integer parity, not a field operation. On observing the full symbol $Y = 2$, the posterior puts $3/4$ on $c = 2$ and $1/16$ on each alternative. On observing only even parity, it puts $7/24$ on each even candidate and $1/16$ on each odd candidate. The report is correct as a parity statement but fails to preserve the posterior needed to distinguish the three even values.

For $n$ conditionally independent identical reports $Y = 2$, the posterior of $c = 2$ is

$$\frac{12^n}{12^n + 4}.$$

For $n$ reports retained only as even parity, the posterior of that same candidate is

$$\frac{14^n}{3 \cdot 14^n + 2 \cdot 3^n}.$$

At $n = 0$ both equal the prior $1/5$. The full-report value at $n = 3$ is $432/433$, whereas the parity value cannot exceed $1/3$. Repeatedly receiving the same coarse fact does not restore distinctions the reporting language never records.

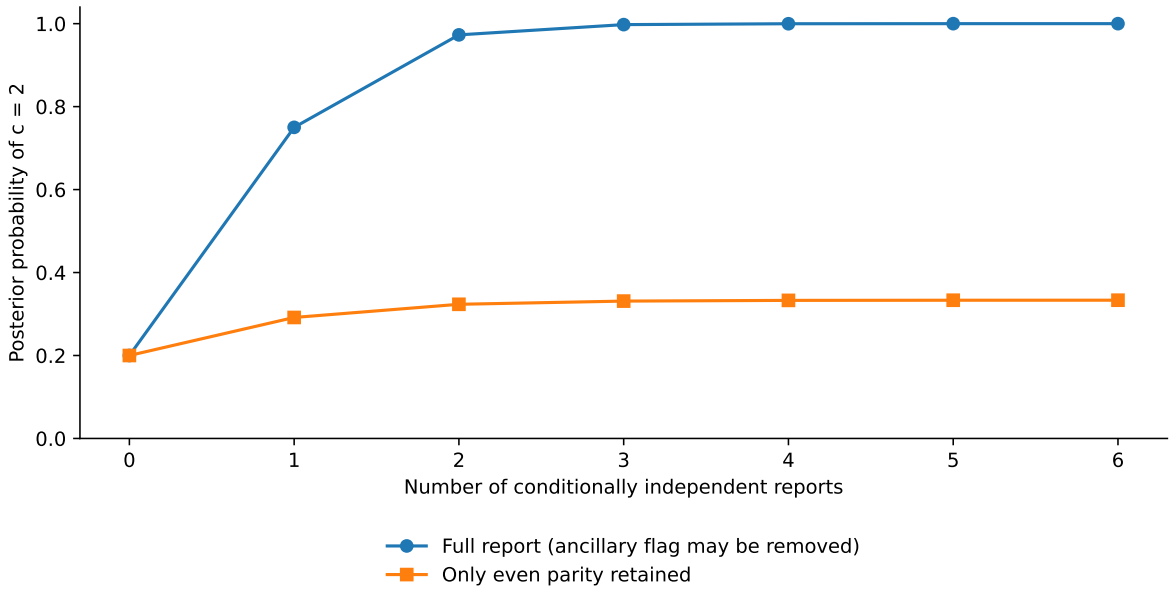



Figure 7: Posterior mass on $c = 2$ along the specified repeated-report histories. Removing the ancillary flag gives exactly the full-report curve. Retaining only parity leaves three even candidates indistinguishable. The curves are analytic conditional calculations, not measured learning trajectories or probabilities of obtaining these histories.

## 13 Approximation, rare evidence, and decision stability

A small error in an unconditional model does not automatically give a small posterior error after an unusual observation. Conditioning divides by the probability of the evidence. A useful correction guarantee must therefore record both an approximation error and the evidence probability at which that guarantee is used.

**Proposition 13.1** (A conditional stability bound with an evidence denominator)**.** *Let $P$ and $Q$ be finite joint laws of a model and an observation, with $d_{\mathrm{TV}}(P, Q) \le \delta$. Let $E$ be an observation event with $a = P(E) > 0$ and $b = Q(E) > 0$. Then*

$$d_{\mathrm{TV}}\big(P(M \mid E), Q(M \mid E)\big) \le \min\left\{1, \frac{\delta}{\max(a, b)}\right\}. \tag{18}$$

*In particular, a justified lower bound $P(E) \ge \gamma > 0$ supplies the weaker usable bound* $\min\{1, \delta/\gamma\}$.

*Proof.* Write $u_m = P(M = m, E)$, $v_m = Q(M = m, E)$, and let $p = u/a$, $q = v/b$. Exchanging $P$ and $Q$ if necessary, assume $a \le b$. The variation inside $E$ is at least $\frac{1}{2}\sum_m |u_m - v_m|$. The variation outside $E$ is at least $|a - b|/2$, since its total masses differ by that amount. Hence

$$\delta \ge \tfrac{1}{2}\sum_m |u_m - v_m| + \tfrac{1}{2}(b - a) = b - \sum_m \min(u_m, v_m).$$

Since $a \le b$, $\min(ap_m, bq_m) \le b \min(p_m, q_m)$. Substitution gives $\delta \ge b\, d_{\mathrm{TV}}(p, q)$. Division by $b = \max(a, b)$ and the universal upper bound one prove the claim. □

The bound is sharp for general joint laws: put the same mass $a$ on $E$ in both laws, choose conditional laws at distance $t$, and make their complements identical. Then the joint distance is $at$. The result neither estimates $\delta$ from data nor proves that the evidence model is calibrated. Those are separate inputs.

**Counterexample 13.2** (Rare evidence can reverse a posterior despite small joint error)**.** For $0 < \epsilon < 1/2$, take a binary model with equal prior probabilities. Under $P$, let $P(M = 0, E) = \epsilon$ and $P(M = 1, E) = 0$. Under $Q$, interchange these two masses. Choose the complement masses so each model retains prior probability $1/2$. Both laws give $E$ probability $\epsilon$, their joint total variation is $2\epsilon$, and their posteriors on $E$ are opposite point masses. Joint error can tend to zero while conditional error remains one.

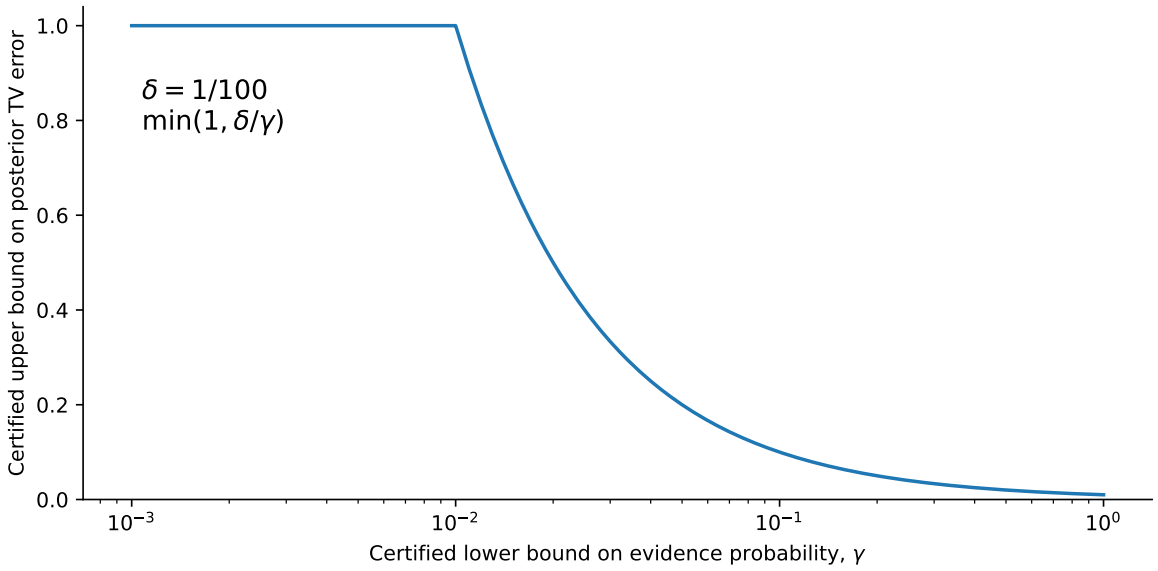



Figure 8: The certified upper bound $\min(1, \delta/\gamma)$ for $\delta = 1/100$. A small joint-law error does not give a useful guarantee when the evidence probability is comparably small. The plot is a bound, not an empirical estimate of posterior error.

### 13.1 From approximate beliefs to a recipient's decisions

Not every posterior distinction matters equally to an action. To link correction with use, fix one finite nonempty action set $A$ and losses $\ell(m, a) \in [0, L_{\max}]$. A receiver evaluates actions from the distribution its actual interpreter reconstructs. The target loss function and available actions must be the same in the comparison.

**Proposition 13.3** (A joint posterior and recipient decision certificate)**.** *Let $p$ be the reference posterior, $\widehat{p}$ an approximate or transmitted posterior, and $r_i$ recipient $i$'s reconstructed posterior. Suppose*

$$d_{\mathrm{TV}}(p, \widehat{p}) \le \epsilon, \qquad d_{\mathrm{TV}}(\widehat{p}, r_i) \le \eta_i.$$

*If $a_i$ minimizes expected loss under $r_i$, then, with $e_i = \min\{1, \epsilon + \eta_i\}$,*

$$\mathbb{E}_p \ell(M, a_i) - \min_{a \in A} \mathbb{E}_p \ell(M, a) \le \min\{L_{\max}, 2L_{\max} e_i\}. \tag{19}$$

*If the expected-loss gap between $a_i$ and every other action under $r_i$ is strictly greater than $2L_{\max} e_i$, then $a_i$ is also the unique optimum under $p$.*

*Proof.* The triangle inequality bounds $d_{\mathrm{TV}}(p, r_i)$ by $e_i$. For a function taking values in $[0, L_{\max}]$, its expectations under these two distributions differ by at most $L_{\max} e_i$. Let $a_*$ minimize expected loss under $p$. Insert the two risks under $r_i$ between the risks of $a_i$ and $a_*$ under $p$. The middle difference is nonpositive by optimality under $r_i$, and the two outside differences are at most $L_{\max} e_i$. Regret is also at most $L_{\max}$. Applying the same two expectation bounds to the difference between $a_i$ and any competitor proves the strict-gap assertion. □

The two errors concern one actual chain of records. One cannot use the best transmission for one recipient and the best posterior approximation for another while reporting a common certificate. When $\widehat{p} = Q(M \mid E)$, Proposition 13.1 can supply $\epsilon$. A separate information-loss estimate or an explicitly verified decoder may instead supply the bound. The result is relative to the stipulated posterior and loss, not a guarantee that a real decision is wise or that a protected real-world outcome can be represented by a bounded scalar loss.

An abstention action can be included with an explicit loss and availability rule. This permits a recipient to decline an unsupported decision within the same model. It does not create a universally correct abstention threshold. Nor does closeness in expected loss establish exact preservation of a logical obligation: an arbitrarily unlikely prohibited outcome can matter under a different contract.

### 13.2 Approximate correction is not arbitrary tolerance

A likelihood that is merely close in average error may have the wrong zeros, changing which observations the model considers possible. Posterior bounds must therefore specify positive evidence, the full observation record, and the domain on which the approximation holds. Confidence produced by normalization is not an independent check of those premises.

The exact factorization criterion in Section 12 is an all-prior statement. The conditional-information identity is an average statement under one prior. The stability bound is an event-specific comparison of two joint laws. The decision certificate concerns one loss family. They are related, but none should be silently substituted for another.

## 14 Sequential inquiry through a retained evidence channel

An explanation intended for further inquiry must support more than one update. The recipient may choose its next experiment after seeing the current report. Exact evidence preservation can survive this adaptation, but only when the experiment-selection and observation assumptions are stated together.

Let the unknown model $M$ remain fixed. At time $t$, a known policy selects experiment $e_t$ using only the prior reported history and recorded experiment names. Randomization is permitted when its conditional law is independent of $M$ given that history. Given $M = m$ and $e_t = e$, a fresh raw observation has law $L_{e,m}$, and its reported form is drawn through $B_e$. Conditional on the model and selected experiment, these draws do not depend on earlier hidden observations. Write $Q_{e,m} = L_{e,m} B_e$.

**Theorem 14.1** (Sequential preservation of noisy correction under adaptive experiments)**.** *Suppose each permitted experiment admits a kernel $\rho_e$ with*

$$L_{e,m}(o) B_e(o, z) = Q_{e,m}(z) \rho_e(o \mid z) \quad \textit{for all } m, o, z.$$

*For every finite positive-probability history under the declared policy, the posterior conditioned on the complete raw and reported record equals the posterior conditioned only on the reported outcomes and experiment names. The claim also holds at a bounded stopping time determined by that reported history.*

*Proof.* For a fixed realized history, its likelihood under model $m$ is the product of the selected-experiment probabilities and the terms $L_{e_t,m}(o_t) B_{e_t}(o_t, z_t)$. The policy factors are independent of $m$ once the reported

history is fixed. Applying the assumed factorization gives a product of the terms $Q_{e_t,m}(z_t)$ multiplied by the model-independent factors $\rho_{e_t}(o_t \mid z_t)$ and the same policy factors. Posterior normalization cancels those factors. The remaining product is exactly the likelihood for the reported history. A bounded stopping decision determined by that history adds no model-dependent factor, so the same calculation applies. □

The theorem gives a reusable correction procedure, not a claim that the selected experiments are optimal. The selection policy can be inefficient while the posterior updates remain exactly preserved. Nor does the theorem cover arbitrary dependent sensor noise. When observation laws depend on hidden physical state or prior observations, that dependence must be represented in the state and likelihood rather than treated as a fresh independent draw.

**Counterexample 14.2** (Copying a record is not repeating an experiment)**.** Let a binary model have equal prior probabilities, and suppose one observed positive report has likelihood $4/5$ under model 1 and $1/5$ under model 0. Its posterior probability for model 1 is $4/5$. Sending two copies of that same report does not add evidence, so the posterior remains $4/5$. Treating the copies as conditionally independent instead gives $16/17$. Both calculations are arithmetically valid for their respective likelihoods, but only the first describes duplication of one record.

A correction record must therefore preserve provenance about which observations are new and which merely repeat an earlier result. A translated explanation that reproduces two identical sentences is not, without that provenance, a reliable record of two independent measurements.

### 14.1 Actions and stopping decisions can also carry information

The theorem records experiment names and restricts selection to information already represented at the recipient. A hidden selector that consults an additional sensor changes the likelihood of which experiment gets selected. Discarding that selection mechanism can bias the update even when the reported sensor itself is well modeled.

A stopping time chosen from the public reported history is different: the history already contains the information used to stop, so the event of stopping contributes no additional model-dependent likelihood on that fixed history. This exact likelihood argument should not be confused with transferring a deterministic-time approximation to an arbitrary stopping time. Small unconditional error at each fixed time still needs a separate argument to yield a useful conditional error on selected histories.

### 14.2 What a correction-preserving artifact must carry

For a recipient to execute this result, an artifact or its declared background must specify the model family, prior, experiment names, likelihoods, reporting channels, and update rule. It must also identify which independence and selection assumptions are being used. Sending only the current most probable answer does not generally preserve the next update.

The recipient does not need every raw token when the factorization is valid. It needs sufficient evidence and an accurate description of the reporting process. In the instrument, the irrelevant flag can be omitted throughout adaptive choices among experiments with verified factorization. A coarse parity language cannot claim the same posterior contract merely because every parity sentence it transmits is true.

# Part III

# Common witnesses and charged resources

## 15 One witness for prediction, transfer, and revision

The separate results now have to refer to one realized system. A good state representation, an elegant translation, and a useful explanation do not establish a combined capability when they require different hidden information, different decoders, or incompatible resource assumptions.

**Definition 15.1** (Integrated adequacy record)**.** An integrated adequacy record for a declared task family identifies a common computational realization, its information boundary, a reusable representation and its updates, permitted intervention implementations, each claimed transport and recipient, the model and evidence maps used for correction, and the cost and scope of those operations. Every claimed guarantee refers to these same objects or to explicitly related replacements.

This is a specification proposal, not a new primitive of cognition. Its purpose is to make the quantifiers visible. The record cannot use a powerful target-specific decoder to establish prediction, then a different decoder to establish transfer, while reporting the result as one reusable method unless the selection procedure is itself part of the method.

### 15.1 What can be checked in the finite core

For explicitly supplied finite tables, the equations in Definition 3.1 can be checked on every state, operation, and query. Theorem 4.1 then proves the guarantee for arbitrary finite continuations. Exact transfer requires corresponding one-step and decoder checks. Correction transport requires the observation square and its stated reflection condition. The recipient must actually construct and execute the declared procedure within its allowed information boundary.

Some obligations remain different in kind. The finite model can list which intervention is admissible, but relating that declaration to a physical device requires evidence about the device. A table can name a sensor reading without proving calibration. A certificate can bind a program version without establishing the semantic meaning of its inputs. The specification should preserve these distinctions rather than present one successful check as universal certification.

| Claim | Required object | What it does not supply |
|---|---|---|
| Reusable prediction | One retained state, updates, decoders, and a fixed future family. | Access to that state or evidence of its actual use. |
| Mechanistic attribution | A realized mapping and discriminating permitted interventions. | Every untested mechanism or unrestricted causal access. |
| Method transfer | Matched states, operations, queries, and coverage. | Reverse reconstruction or new target operations. |
| Recipient explanation | Fixed background, an artifact, an interpreter, and a budget. | Faithfulness to the source solver's actual computation. |
| Corrective transfer | Candidate and evidence maps that preserve filtering. | Model invention, noise calibration, or truth outside the family. |
| Resource feasibility | One accounting boundary for construction and repeated use. | A universal efficiency advantage. |

### 15.2 The complete instrument witness

For the original contract, the instrument witness contains $r(a, b, c) = (a + b, b)$, updates (2), the first-coordinate decoder, and a statement that $c$ is outside the relevant future distinctions. It also names how an observer acquires the initial record. A recipient with that record and the equations can reproduce all declared predictions.

After $J$ is admitted, the witness changes. The old representation does not predict the new reading. An explicit $J$ observation supplies $c$ and updates the first coordinate. The enlarged state has the exact updates in Section 9. The original $H/W$ explanations remain valid on their unchanged domain. New predictions require the acquired evidence, and any translated explanation must preserve that evidence's discriminating meaning.

These are connected properties of one analytical construction. They do not establish that an unspecified learning algorithm will discover the representation, infer the correct transition law, or choose the diagnostic operation. The point is to make those remaining research obligations identifiable rather than confuse them with the solved representation problem.

### 15.3 A cost account for reusable methods

For a sequential workflow with disjoint accounting, total cost includes acquiring information, discovering the representation, encoding state, performing updates, verifying claimed relationships, translating when needed, and interpreting the result. Parallel execution requires work and elapsed time to be recorded separately. Costs in different units should not be added without a stated conversion.

A simple amortization comparison illustrates the issue. Suppose a representation costs $D$ to construct, each later use costs $c$, and the matched baseline costs $b > c$ per use. The representation has lower total cost after $N$ uses exactly when

$$D + Nc < Nb, \quad \text{or equivalently} \quad N > \frac{D}{b - c}.$$

This accounting does not prove that the representation can be discovered for cost $D$, that the uses are comparable, or that all target instances share it. It locates where such evidence would enter.

Understanding need not always make computation faster. A more costly representation can make a proof inspectable, reveal a missing assumption, or support an intervention unavailable to the cheaper one. The relevant benefit must be stated. The theory should not smuggle efficiency into the meaning of understanding, nor infer efficiency from the existence of a short description.

## 16 Joint acquisition and explanatory communication

The acquisition problem and the recipient problem have so far been solved separately. Their composition is not automatic. An inexpensive observation may leave the sender unable to encode what a recipient needs. Conversely, requiring the sender to reconstruct the entire state can demand more evidence than is needed to help recipients who already possess relevant background.

We now combine the two exact finite models without introducing an omniscient encoder. The resulting test gives one witness for the experiment tree, its information boundary, a common artifact, and each recipient's decoding contract.

### 16.1 The graph belongs to the acquired transcript

Fix a finite set $X_0$ of possible initial states, a finite deterministic experiment catalogue as in Section 10, and a finite horizon $h$. A deterministic experimental policy is an observation-history tree. Its choices cannot depend on the hidden initial state. Only uniformly admissible choices are permitted at each reachable node.

For a policy tree $T$ and initial state $x$, write $s_T(x)$ for its terminal sender transcript and $x_T(x)$ for its final physical state. The tree has finitely many reachable leaves. Specify recipient backgrounds $\beta_i^T(x)$ and demands $d_i^T(x)$ by a declared rule. A demand may concern $x_T(x)$, the original $x$, or a stated pair of past and present values. These are different contracts.

The recipient sees only the common artifact and its specified background. If an experiment name, its timing, or another transcript is visible to the recipient, that information belongs in $\beta_i^T$. It must not be used as an uncharged auxiliary message. Let $C(T)$ be the maximum total experimental cost across initial states. Build $G_T$ on the reachable sender transcripts using the graph construction of Section 8, with $\alpha = s_T$.

**Theorem 16.1** (A finite joint certificate for acquisition and explanation)**.** *For finite $X_0$, a finite experiment catalogue, horizon $h$, cost budget $C$, and artifact alphabet bound $K \geq 1$, there exists a uniformly admissible deterministic acquisition policy and one common artifact satisfying every recipient demand exactly if and only if some attainable policy tree $T$ satisfies*

$$C(T) \leq C, \qquad G_T \text{ has no loop}, \qquad \chi(G_T) \leq K.$$

*This is a finite decision problem. An accepting witness consists of the policy tree, a proper coloring, and the induced fixed decoders. Among feasible trees, a minimum experimental cost is attained. The statement does not assert that the search or the resulting decoders are efficient.*

*Proof.* Any feasible system induces a policy tree of depth at most $h$, because its experiment choices depend only on its available history. The sender's final artifact is a function of its terminal transcript. Applying Theorem 8.1 to those transcripts proves the graph conditions. The worst-case experimental cost is $C(T)$, so it satisfies the budget.

Conversely, choose a tree with the stated properties. Execute its uniformly admissible experiments, encode the terminal transcript using its proper coloring, and use the decoders constructed in Theorem 8.1. They give every required answer on every initial state. The tree's cost and horizon meet the contract. There are finitely many experiment and outcome choices at each node and a finite depth bound, hence finitely many candidate trees. Exhaustive admissibility, cost, and coloring checks decide existence. A nonempty finite set of feasible tree costs has an attained minimum. □

The graph is built across all reachable leaves, not separately at each leaf. Two different transcripts can be indistinguishable to a recipient after encoding even when each sender support is locally well understood. A message assignment chosen independently at each leaf can therefore fail the common-artifact requirement.

For a fixed preexisting decoder family, each colored artifact must additionally be checked against that family. For a finite catalogue of allowable implementations, direct verification of those witnesses remains finite. The coloring theorem alone does not make a convenient decoder available, just as the acquisition theorem does not create a sensor.

**Counterexample 16.2** (A useful explanation need not identify the state at its source)**.** Let the hidden state be two bits $(a, b)$. The sender's only experiment reads $a \oplus b$ without changing the state. Repeating it never identifies the complete pair. One recipient already knows $a$ and needs $b$, while another knows $b$ and needs $a$. The single parity reading and its one-bit broadcast satisfy both demands. Before that reading, the sender has only one transcript and the explanation graph has a loop. After the reading, the graph is one edge and the one-bit message is exact.

This is an integrated information construction, not evidence that the sender possesses the recipients' full understanding. The explanatory success is jointly realized by the acquired parity and the recipients' backgrounds. Attribution should identify which participant carries which distinction.

### 16.2 Why noisy belief and exact state recovery remain different

The joint theorem concerns deterministic experiments and exact demands. The Bayesian sections concern noisy records and distributions over a model family. One cannot replace the target state in the joint theorem with a most probable guess and claim the exact guarantee has been preserved.

In the noisy instrument, every readout has positive likelihood under every candidate value of $c$. With a fully supported prior, any finite sequence of those readouts leaves every candidate with positive posterior probability. Repeated reports can make one value overwhelmingly likely, but they cannot give zero-error identification within a fixed finite number of such observations. A decision-loss tolerance, a posterior threshold, and an exact reconstruction target are different terminal conditions.

The decision certificate in Proposition 13.3 supplies one bridge: a recipient can act under an explicit error budget without pretending its estimate is exact. A more general synthesis optimizing noisy experimentation and multi-recipient communication jointly remains a separate problem. That larger stochastic optimization is outside the finite exact synthesis of this edition.

### 16.3 A complete expanded instrument record

The original instrument already separates current display, reusable state, and newly relevant information. The expanded account adds the mechanisms by which that information is acquired, represented probabilistically, and shared.

First, the reasoner retains the base coordinates and the exact scope of $H/W$ reuse. Admitting $J$ splits old behavioral classes. The exact diagnostic plan uses a real $J$ operation and preserves its effect on the first register. If only the noisy sensor is available, the retained object is a posterior over possible post-$J$ states, not an invented exact value of $c$.

Second, the evidence description identifies which readouts are conditionally independent and which tokens are copied or ancillary. Removing the independent flag is correction preserving. Replacing all five possible readings by parity is not. Three identical full readings give the conditional posterior $432/433$ for the corresponding candidate, but that number remains conditional on the declared likelihood model and prior.

Third, recipients receive the method appropriate to their backgrounds. A recipient with the base coordinates needs the newly acquired distinction or its uncertainty distribution. A recipient without those coordinates needs additional state information. A shared artifact is checked across all of those backgrounds at once. Its source encoder may use only the information actually acquired.

Finally, the record includes the decoder, query loss or exactness criterion, operation catalogue, and full resource account. Discovery of the transition law, calibration of the sensor, and mechanistic attribution to a particular neural system are not supplied by this analytical example. They remain separate scientific claims with separate evidence requirements.

### 16.4 A release boundary rather than an inflated definition

The mathematical contribution is a connected set of exact and conditional preservation results. The finite graph and decision tree supply executable witnesses. The likelihood factorization identifies when discarded records do not change Bayesian correction. The error bound exposes when a rare observation makes an approximate model inadequate. The recipient certificate carries those errors into a defined decision criterion.

Together they make a stronger specification of usable understanding. They do not turn every valid compression into a cognitive achievement or every encoded table into an economical explanation. Claims of economy require an explicit interpreter language, charged background information, and a measured or proved implementation cost. Claims of actual use require a realized causal mechanism. Keeping these obligations separate is part of the theory's content, not a reason to postpone its mathematical statements indefinitely.

## 17 What the resource account must charge

A short explanatory message can conceal an expensive recipient. A compact state can require an elaborate encoder. A fast online answer can depend on a table that was expensive to construct and distribute. None of these possibilities makes the answer wrong. They change what has been shown about the economy of the method.

The first-edition resource contract separates four accounts. The *design account* includes discovering and validating the state model, constructing a decoder, and searching for a policy. The *installed account* includes the fixed interpreter, model tables, recipient background, and compiled controller. The *online account* includes observations, physical interventions, communication, and executed instructions. The *revision account* includes obtaining new evidence and replacing invalid assumptions or interfaces. A quantity cannot be declared absent merely because it was paid in a different account.

These accounts use different units. Description bits, controller records, modular operations, physical interventions, and latency are not interchangeable. A weighted scalar cost is legitimate only after its conversion factors have been specified. The synthesis below retains a vector of online charges and a separate controller storage count. It does not silently convert one recipient's computation into the sender's communication budget.

### 17.1 Task-dependent interpreter changes are information

Let $\Theta$ be a finite nonempty family of tasks and let each task have a required answer profile over a fixed query family. A fixed interpreter $U$ receives background $B_\theta$, an installed program string $p_\theta$, an artifact $e_\theta$, and a query. It must return the required answer. The interpreter is fixed across tasks. The strings may depend on the task, but their lengths are charged. Let $N_B$ be the number of distinct required answer profiles among tasks having the same background value $B$.

**Proposition 17.1** (Charging a task-dependent interpreter)**.** *Suppose $p_\theta$ and $e_\theta$ have fixed lengths $d$ and $b$ bits, respectively, and the deterministic interpreter answers every required query exactly. Then*

$$2^{d+b} \geq \max_B N_B.$$

*In particular, representing every Boolean function on $n$ input bits under one fixed background requires $d + b \geq 2^n$. This counts task-dependent information, not the running time of the interpreter or the information needed for any one specially chosen function.*

*Proof.* Fix a background $B$. If two distinct required answer profiles used the same pair $(p_\theta, e_\theta)$, the fixed interpreter would receive identical inputs for a distinguishing query and would have to give the same answer. Exactness would fail for one of the two tasks. There are at most $2^{d+b}$ pairs, proving the first bound. There are $2^{2^n}$ Boolean functions on $n$ input bits, with distinct complete answer profiles. Taking base-two logarithms proves the second assertion. □

This is the counting argument of Proposition 7.2 applied to the complete task-dependent package. It is not a new general description-complexity theorem. The point is to prevent a moving interpreter from becoming an unreported communication channel. Variable-length descriptions require a declared framing or prefix convention. Merely saying that a message has length at most $b$ is not the same fixed-length model, because the length itself can carry information.

A task-independent interpreter can contain rich mathematical knowledge without encoding which task instance is present. Its construction and availability still belong in the installed and design accounts. Conversely, a separate decoder chosen for every task must be identified and charged even when that decoder

is called a background assumption. Control-task critiques of probing and expressivity critiques of alignment make this boundary especially important [8, 9].

### 17.2 Amortization without disappearing costs

Suppose a method has setup cost $S$ and per-use cost $r$, while a comparator costs $c$ on each use. In the same declared cost units and with no other differences, the method costs less over $N$ uses exactly when $S + Nr < Nc$. When $c > r$, this is $N > S/(c - r)$. At equality the costs tie. If $c \le r$ and $S \ge 0$, setup cannot create a strict cost saving in this model.

The formula does not establish that the method can be discovered, that the repeated problems are interchangeable, or that its validity survives a change of task distribution. Its role is bookkeeping. A useful representation can repay a substantial initial investment. It cannot claim that investment never occurred. Likewise, a precomputed table can be a sensible engineering choice for a bounded domain. The comparison should expose its storage and preparation rather than reject it by terminology.

### 17.3 Cost is an attribute of a specified implementation

An algebraic equation does not have one architecture-independent runtime. Addition modulo five can be a primitive instruction, a small Boolean circuit, or a table lookup. The worked implementation in Section 20 counts modular arithmetic operations and physical calls explicitly. It also records the sizes of its actual code and serialized witness. These byte counts describe those files in one environment, not minimum possible descriptions or a measure of cognitive complexity.

Causal attribution remains separate. A cheap decoder may show that information is readily accessible. It does not show that the source solver actually uses that decoder or the selected representation. Resource accounting strengthens an explanatory claim by identifying where work is done. It does not replace the implementation evidence discussed in Section 5.

## 18 Explanations for fixed recipient procedures

The coloring criterion in Theorem 8.1 permits fixed decoders to be designed for the chosen code. A recipient that already exists may not be able to execute those decoders within its budget. We now freeze the decoding procedures before choosing the artifact.

Let $X$ be a finite nonempty set of cases, $\alpha : X \to S$ the sender's access map, and $\mathcal{A}$ a finite nonempty artifact catalogue. Recipient $i$ has a background map $\beta_i : X \to B_i$, a demanded output $d_i : X \to D_i$, and a fixed procedure $L_i$. Evaluate $L_i(\beta_i(x), a)$ under a declared instruction and memory bound. A timeout, an invalid output, or a rejected artifact counts as failure. Since the inputs and execution bounds are finite, the resulting acceptance table is finite even when $L_i$ is written in a general programming language.

Define

$$\mathcal{A}(x) = \{a \in \mathcal{A} : \text{every recipient returns } d_i(x) \text{ within its declared limits}\},$$

and, for each reachable sender value $s$, define

$$\mathcal{A}_s = \bigcap_{x:\alpha(x)=s} \mathcal{A}(x). \tag{20}$$

An artifact selected from $\mathcal{A}_s$ must work for all cases the sender still considers possible. Choosing a different acceptable artifact after secretly learning which hidden case is true is not allowed.

**Theorem 18.1** (The fixed-recipient covering criterion)**.** *Under this finite contract, an exact encoder $E : S \to \mathcal{A}$ exists on the reachable sender values if and only if every $\mathcal{A}_s$ is nonempty. Its smallest possible image size is*

$$\min\{|C| : C \subseteq \mathcal{A},\ C \cap \mathcal{A}_s \neq \varnothing \text{ for every reachable } s\}, \tag{21}$$

*when the feasible family is nonempty. Thus a codebook of at most K artifacts is possible exactly when a set of at most K artifacts meets all these requirements. Every valid encoder induces a proper coloring of the sender-access graph, so its image size is at least that graph's chromatic number. The converse graph condition need not hold for the fixed procedures.*

*Proof.* A sender seeing $s$ must emit one value $E(s)$, and that value must lie in $\mathcal{A}(x)$ for every $x$ with $\alpha(x) = s$. This is exactly membership in $\mathcal{A}_s$. Hence a valid encoder implies nonempty intersections. Conversely, selecting one value from each nonempty finite intersection constructs a valid encoder.

The image of any such encoder intersects every $\mathcal{A}_s$. If a set $C$ intersects every requirement, selecting $E(s)$ from $C \cap \mathcal{A}_s$ gives an encoder whose image is contained in $C$. This proves the minimum formula by finite minimization. If two adjacent sender values received the same artifact, a recipient with the same background would receive identical decoder inputs while requiring distinct outputs. Its fixed procedure could not satisfy both. This proves the graph lower bound. □

The finite optimization in (21) is a set-cover formulation on the reachable sender values. It counts how many distinct catalogue entries are needed. It does not license replacing their actual encodings by short indices unless the corresponding index decoder is installed and charged. A catalogue entry may be a large program or proof object even when only one such entry is used.

**Counterexample 18.2** (Information suffices but the installed decoder needs a larger codebook)**.** Take three cases $x_0, x_1, x_2$ fully visible to the sender. One recipient already has background $i$ in case $x_i$ and always needs output zero. The sender-access graph has no edges, so a designable decoder could ignore a single common message and return zero. Now fix an installed decoder on artifacts $A, B, C$. It returns zero on $A, B$ at background 0, on $B, C$ at background 1, and on $A, C$ at background 2. It fails on the remaining pairs. Every two of the three acceptable-artifact sets intersect, but their common intersection is empty. Two catalogue entries suffice, while one does not. The graph lower bound is one and the fixed-decoder optimum is two.

This construction is intentionally small. It isolates the difference between information-theoretic possibility and compatibility with an already installed procedure. A more capable replacement decoder repairs the example, but replacing it changes the installed account. A procedure that times out on a required input provides an equally direct reason a graph witness can fail at the resource boundary.

### 18.1 Sender uncertainty is a different obstruction

Nonempty casewise sets $\mathcal{A}(x)$ do not suffice. If the sender cannot distinguish two cases whose acceptable-artifact sets are disjoint, their intersection in (20) is empty. More recipient compute cannot repair a distinction missing from both its background and the transmitted artifact. A new observation may repair the sender's access. An added recipient background may repair the receiving side. Neither can be credited without being part of the actual workflow.

The bounded table concerns exact outputs. For approximate explanation, acceptable outputs can instead be defined by a fixed loss tolerance, but that changes the table and the claimed guarantee. A most probable model is not an exact current state merely because the decoder confidently emits it. The Bayesian and decision results retain their own uncertainty contracts.

## 19 Joint acquisition with a compiled resource budget

The preceding theorem freezes recipient procedures. We now combine it with state-changing acquisition and charge the entire compiled observation tree. This closes a specific first-edition gap: a small artifact, an affordable experiment, and a runnable decoder must belong to one attainable workflow.

### 19.1 The explicit finite implementation model

Use a finite augmented state space $X$. Include every origin variable, history flag, recipient background, and permission needed to make the declared transition and terminal tables valid. Each experiment $e$ has the transition, observation, and admissibility semantics of Section 10, together with a fixed nonnegative integer cost vector $c_e \in \mathbb{N}^r$. Its coordinates have declared units. The charge includes the specified source-side interface work for that experiment. We do not assume that acquiring information leaves the physical state unchanged.

For each state $x$ and artifact $a$ accepted by all fixed recipient procedures, let $\gamma(x, a) \in \mathbb{N}^r$ be the terminal charge. It includes transmission and the declared source and recipient execution work. Invalid or out-of-budget decoder executions have no terminal entry. At support $V$, a stop using $a$ is allowed only if it works for every $x \in V$, with worst-case vector

$$g_a(V) = \max_{x \in V} \gamma(x, a),$$

where maxima and comparisons of vectors are componentwise. Different coordinates may attain their maxima on different cases. A coordinatewise budget nevertheless has to cover every one of them.

A controller is an explicit observation tree of depth at most $h$. Each decision node or stop node occupies one fixed-size record, with fields for the operation or artifact and the bounded number of successor pointers. The count $n(T)$ includes all stored branches, not just those visited in one execution. The interpreter, recipient programs, and state tables are installed separately and remain fixed throughout the optimization. Their storage is added to the recorded controller size. Sharing equal subtrees in a directed acyclic graph is a different storage model.

For a tree $T$, let $v(T)$ be its componentwise worst-case charged online cost. Let $\mathcal{P}_h(V)$ be the attainable pairs $(v(T), n(T))$. The stop pairs are

$$\big(g_a(V), 1\big) \quad \text{for each artifact accepted throughout } V. \tag{22}$$

For $h > 0$, each uniformly admissible experiment gives the additional pairs

$$\left(c_e + \max_o v_o,\ 1 + \sum_o n_o\right), \quad (v_o, n_o) \in \mathcal{P}_{h-1}(V_{e,o}), \tag{23}$$

with one child choice for every nonempty observation branch. An experiment contributes no pair when even one possible branch has no completion. At $h = 0$ only stops are allowed. Sets are finite because the catalogue and horizon are finite.

**Theorem 19.1** (A joint resource certificate for fixed recipients)**.** *Recurrences* (22)–(23) *enumerate exactly the cost profiles of the permitted explicit trees. A workflow within online budget $b \in \mathbb{N}^r$ and controller-record budget $N$ exists if and only if some $(v, n) \in \mathcal{P}_h(V)$ satisfies $v \le b$ and $n \le N$. Every such pair has a finite executable tree witness. Keeping only componentwise nondominated pairs at each recursive step preserves feasibility for every pair of budgets $(b, N)$.*

*Proof.* At depth zero, an admissible workflow must stop with one artifact that works for every possible state. Its worst-case terminal charge and its one stored record give exactly (22). For the inductive step, any nonterminal tree first executes an experiment available throughout $V$. Its possible next supports are exactly $V_{e,o}$. By induction each child tree supplies one attainable pair. Every branch is possible, so the online charge is $c_e$ plus the componentwise maximum of the child charges. The installed explicit tree contains the root and every child record, giving the sum in (23). Conversely, combining an admissible root with the witnessing child trees realizes that pair. This proves exact enumeration and the budget equivalence.

If one child profile is dominated by another attainable profile at the same support and remaining horizon, replacing that child cannot increase a parent's componentwise maximum or its sum of records. Terminal admissibility and future observations are unchanged. Repeating this replacement proves that discarding dominated profiles preserves a budget-feasible witness whenever one exists. Finite induction gives the claim at the root. □

This is finite multiobjective dynamic programming under a fixed implementation contract. Its components have the same state-planning and side-information foundations as the earlier results [17, 15, 16]. The new obligation is their simultaneous realization with frozen recipient procedures, online charges, and explicit controller storage. The theorem does not optimize over arbitrary programs, discover missing sensors, or make policy synthesis computationally cheap.

**Counterexample 19.2** (Separate resource optima do not define a feasible common workflow)**.** Let a hidden origin take four values and remain fixed under every experiment. A direct experiment reveals its value at charge four units. Its explicit controller has one decision node and four stop nodes, giving profile $(4, 5)$. Three other experiments ask whether the origin equals 0, 1, or 2, respectively, each at charge one unit. A decision chain of these equality tests distinguishes the four origins with worst-case charge three and seven stored records, giving profile $(3, 7)$. The fixed recipient decoder reads a two-bit origin artifact at a leaf. Its work and the common message length are held fixed and are not included in the one online coordinate shown here.

Within horizon three the nondominated profiles are $(4, 5)$ and $(3, 7)$. A tree with charge at most three cannot use the direct experiment. A binary equality tree distinguishing all four origins needs at least four leaves and hence at least seven records. Conversely, any tree distinguishing four origins needs four distinct stopping artifacts and at least one decision node, so five is the smallest possible record count. Thus the separate minima are three online units and five records, but the combined budget $(3, 5)$ admits no tree.

The direct observation is compact as an installed tree. The equality chain is cheaper to execute in the worst case. Taking the best coordinate from each certificate invents a workflow that neither realizes. Replacing the explicit tree by a program with shared branches changes the storage model and must be assessed separately.

### 19.2 Which implementation costs remain outside the recurrence

Offline search and proof checking do not disappear because the controller is compiled. Their time, software, and input tables are recorded separately. The online interpreter must be specified, and its dispatch costs can be included in $c_e$ and $\gamma$. A record count becomes a storage guarantee only after the fixed record layout and installed tables are included. The number of source-file bytes is an audit quantity, not a theorem of minimal description length.

The result is exact for declared charges. If those charges are conservative upper bounds for an implementation, the certificate is correspondingly a sufficient implementation budget, not necessarily its tight physical cost. If the actual decoder reads an unlisted file or the sensor consults a hidden variable, the implementation is outside the contract. The right response is to expand and recheck the contract, not to call the extra access free.

A source with several possible initial observations can be handled by running the construction on each initial support and specifying how the installed controller dispatches among them. If that observation is public to the recipients, it belongs in their background. If it is private, the fixed decoders still see only their declared inputs. The integrated instrument below makes both sides explicit.

## 20 A costed first-edition instrument certificate

The common realization can now be described without a designable recipient or an omniscient sender. The instrument is still $(a, b, c) \in \mathbb{F}_5^3$. Before the diagnostic, the sender knows only $(s, b) = (a + b, b)$. Recipient

A has the old register pair $(a, b)$. Recipient B has the old display pair $(s, b)$. Neither recipient has $c$. All three parties know the operation laws and the fixed diagnostic protocol before the particular hidden state is selected.

The sole diagnostic in this version applies $J$ exactly once and returns its exact post-operation display $s_J$. The sender computes $m = s_J - s = c$ and broadcasts that one field element. The report $s_J$ and the sender's private controller address are not transmitted separately. The recipients are told by the installed protocol that their demanded record is the state after this one diagnostic, not an arbitrary later state.

The two frozen decoders are

$$L_A((a, b), m) = (a + m, b, m), \qquad L_B((s, b), m) = (s + m, b, m). \tag{24}$$

Recipient A obtains the post-$J$ register description. Recipient B obtains the post-$J$ display-centered description. Both compute one modular addition. Their installed update procedures then support all $H/W/J$ continuations in their respective coordinates. These procedures are supplied, not learned by the diagnostic.

For a fixed old $(s, b)$, five possible values of $c$ remain. Before an observation, no one message can supply both exact demanded records for every one of those cases. After the diagnostic, the five report values lead to five stop nodes and each selects its one valid message. The explicit tree has six records. A fixed three-bit field encodes the five messages. Two bits do not suffice. Three additional unused binary patterns are rejected, not assigned secret meanings.

### 20.1 The execution ledger

The following charges describe the reference procedure rather than a universal machine model. Assignments and interpreter dispatch are recorded as implementation operations, but are not included in the modular-arithmetic coordinate. Each field operation is one modular addition or subtraction, independent of how it is realized by a processor.

| Resource | Count | Meaning and boundary |
|---|---|---|
| Physical diagnostic | 1 | One $J$ operation with one returned exact display. Sensor construction is supplied. |
| Source modular arithmetic | 1 | Subtract the old display from the new one. Selection uses the compiled protocol. |
| Broadcast payload | 3 bits | One fixed-length field for five possible messages. No branch address or raw report is an additional message. |
| Recipient A modular arithmetic | 1 | Compute $a + m$ in its fixed decoder. |
| Recipient B modular arithmetic | 1 | Compute $s + m$ in its fixed decoder. |
| Explicit controller | 6 records | One diagnostic node and five message leaves for the fixed old record. Two records are visited in a run. |
| Installed method | Recorded separately | Both decoder programs, transition semantics, framing, and prior background. Not charged as newly acquired $c$. |

The code enumerates all 25 old-record fibres and all 125 initial instrument states, compiles a resource witness, and verifies its leaves by executing the fixed decoders. It also tests incorrect messages, missing branches, reduced horizons, and insufficient recipient instruction limits. These are exact finite checks of the supplied model. They do not establish that an external reasoner discovered the representation or that a noisy device has this exact sensor.

The six-record count applies to one fixed old record and a literal observation tree. A generic arithmetic controller can share structure across all 25 old records and need not store 25 separate trees. The reference

program implements that arithmetic procedure as well and records its code size. Neither representation is declared the universally smallest controller. The distinction between tree storage and an arithmetic program is retained instead of combining their best costs.

### 20.2 Reuse has an explicit implementation

In B's coordinates $(s, b, c)$, each $H$, $W$, or $J$ update uses one modular addition or subtraction and the query reads the first coordinate. Thus a length-$n$ continuation uses exactly $n$ modular arithmetic operations after initialization, plus its recorded dispatch, assignment, and output operations. In A's coordinates the current display requires the additional addition $a + b$, while swap can be a coordinate permutation. The programs are different but agree on the declared answers through the exact translation theorem.

The supplementary record has five possible values. A separately stored three-bit message appended to a jointly packed five-bit base record uses eight bits, whereas a repacked complete 125-state record requires seven bits. A native implementation storing three individually encoded residues can use nine bits. These are three distinct storage layouts. The delivered source reports which layout is used rather than subtracting rounded bounds from incompatible layouts.

### 20.3 An uncertain reading requires a different artifact

The deterministic certificate does not apply to the noisy sensor merely because the same physical register is present. Under the noisy contract, the sender carries a posterior over $c$, not a certified exact value. A recipient can use a transmitted likelihood history or a specified posterior representation together with the existing correction and decision guarantees. Its message size and arithmetic work must then be charged again. They are not inherited from the three-bit exact-state message.

The ancillary flag may still be omitted when its verified factorization holds. Copies of one reading may not be treated as independent measurements. When a public adaptive selector chooses the next experiment using reported history, Theorem 14.1 applies under its fresh-observation assumptions. A hidden selector, an unknown likelihood, or an observation deemed impossible by every current model requires additional evidence or a revised model.

### 20.4 What this certificate warrants

The first edition supplies exact witnesses for specified representations, experiment trees, reporting channels, recipient procedures, and resource profiles. These witnesses support different statements. The quotient preserves future answers. An admissible diagnostic acquires a missing distinction. A factorization preserves a Bayesian update. A fixed decoder turns an artifact into a usable record. The common resource witness shows that particular components can be executed together.

None of these conclusions identifies the internal mechanism of an unrelated solver. Nor do they establish subjective experience, a universal ranking of understanding, or accuracy of unstated physical premises. A complete adequacy record should preserve those unresolved questions alongside its positive results. That is how a reusable method remains answerable to evidence rather than becoming an impressive answer with an invisible scope.

# Part IV

# Interpretation and scope

## 21 Limits, objections, and the next mathematical questions

The framework is intentionally operational, but that does not make it philosophically complete. A contract specifies selected consequences. It does not establish that those are all the consequences that matter. A formal model of a subject can be wrong, and a model of a reasoner's computation can be incomplete. A successful finite witness does not answer every question about meaning, intentionality, or experience.

### 21.1 Does this define understanding as success?

Not as a single success rate. The framework separates what a state can support, what a solver actually uses, what can be transported, what a recipient can reconstruct, and what survives correction. The counterexamples show that these properties differ. Nevertheless, every operational criterion remains a criterion for a declared set of abilities. This account does not offer an independent essence of understanding beyond all such criteria.

The resulting restraint is useful. A system may possess strong reusable knowledge in one domain and weak correction behavior in another. A receiver may understand an explanation under a supplied notation while failing to translate it. A process may be correctly modeled by an outside analyst without being able to explain itself. These are distinct descriptions rather than contradictory verdicts about an indivisible trait.

### 21.2 Can a table satisfy the theory?

Yes, for some finite contracts. A table can encode a transition system and support exact future queries. It should not be declared incapable by terminology alone. The stronger questions concern its origin, size, use, reuse across a family, and behavior when the contract changes.

The recoding construction in Section 5 makes this unavoidable. To distinguish a compact transferable method from a complete enumeration, one must specify the comparison class and account for the encoder, decoder, and shared background. Restricting those components is part of the research design. An assertion that a representation is meaningful is not a substitute for doing so.

### 21.3 Do valid proofs establish understanding?

A proof establishes the conclusion licensed by its premises and rules. The recipient may still lack a usable model of why those premises were appropriate or how the proof changes under a new problem. Conversely, a reasoner can use a correct model while being unable to produce a formal proof in a particular proof language.

Formal verification is therefore valuable but not interchangeable with the entire profile. The present proofs establish finite and metric statements under their assumptions. They do not certify that a deployed neural model realizes the maps, that a human reader has understood them, or that the task specification captures all relevant external facts.

### 21.4 Finite models and a changing world

The exact quotient assumes a fixed state space and a fixed operation family. The correction theorem assumes specified model families and outcome maps. The instrument's new operation enlarges one of those families in a controlled way. Real inquiry can also introduce new variables, new causal hypotheses, new measurement channels, and new forms of explanation.

There is no guarantee that the minimal state for every future extension is small. Preserving all conceivable future distinctions can collapse the compression objective back toward retaining the entire accessible state, and even that state may omit external information. A useful theory must allow partial coverage and a principled request for new evidence rather than promise unlimited generality.

### 21.5 What is now resolved, and what remains

**Bounded nonvacuity.** The sender-access graph prevents an encoder from consulting missing information, and the joint certificate keeps acquisition and communication in one contract. Sections 17–19 add fixed interpreter procedures, charged task-dependent descriptions, and exact finite resource frontiers. These do not constitute a general complexity theory of concepts. Useful scalable representation languages, economical decoder discovery, and fair cross-architecture comparisons remain research questions. A finite coloring can be optimal in alphabet size and poor as an explanation to an actual person.

**Noisy and approximate correction.** Theorem 12.1 resolves exact posterior preservation for finite likelihoods and common reporting channels. Proposition 13.1 quantifies event-specific sensitivity, and Proposition 13.3 carries approximation through a recipient to a bounded-loss decision. These results assume the likelihood family and its errors are supplied. Robust model sets, dependent noise, misspecification, and empirically justified uncertainty require additional work.

**Acquiring new distinctions.** Theorem 10.1 resolves optimal worst-case acquisition in the finite deterministic, finite-horizon model, including the experiment's effect on current state. Theorem 16.1 adds a common explanatory artifact. Efficient search, noisy joint experimentation and communication, and the invention of missing sensors or hypotheses are not supplied. A statement about feasible observation cannot be promoted to a claim of autonomous scientific discovery.

**Recipients and revision over changing contracts.** The graph criterion resolves exact communication when decoders can be designed. Theorem 18.1 instead resolves finite artifact selection for preexisting bounded procedures. Neither identifies which procedure is realized by an uninstrumented learned reasoner. The sequential correction theorem assumes a fixed model family and matching experiment semantics. A recipient who changes its conceptual language or acquires a new operation may need a new sufficient record, not just the old certificate copied forward.

These remaining questions define the boundary of the first edition. Version 1.0 is a stable theoretical scope with reproducible artifacts. It is not a claim that every question about understanding has been settled. Later empirical findings or mathematical corrections can revise the account without turning an editorial release number into a scientific guarantee.

## 22 Conclusion

A correct answer is an achievement. Reusable computational understanding concerns what a reasoner can carry beyond that answer: a method for new continuations, a representation another reasoner can use, and an explicit route to correction when further evidence changes what is warranted. Those properties belong to a continuing relationship among a subject, a representation, evidence, and a recipient.

The instrument makes the distinctions concrete. One display answers a present question. Two coordinates support every continuation under the original operations. A third register becomes relevant when a new operation is admitted. An exact diagnostic can acquire that distinction, but also changes the instrument. A noisy reading supplies a posterior instead. Reporting channels preserve correction only under the corresponding

evidence conditions. Repeating a copied report is not a new observation, and a small unconditional error can become decisive after rare evidence.

Explanation adds another realization boundary. A common artifact must fit what the sender actually knows and what each recipient already possesses. Even a message with sufficient information can be unusable by the installed decoder. Changing that decoder can help, but its task-dependent information, installation, and execution must be charged. The first-edition covering and resource constructions make those obligations jointly checkable in finite models. A controller small in stored records need not be the cheapest one to execute. The best separate resource bounds need not describe one possible workflow.

The theory does not select a uniquely human notation or exclude a table merely because it is a table. It asks what each implementation retains, how it was acquired, what work it performs, and which losses remain after translation. Predictive adequacy, causal use, economical realization, recipient usefulness, and responsiveness to evidence are related but distinct. A proof about one is not a certificate for all of them.

Operation Komorebi's realization perspective provides the common discipline: specify the operations that a representation preserves and make unresolved distinctions visible. The resulting framework supports exact mathematical claims without inferring consciousness, universal comprehension, or an unobserved mechanism from formal success alone. Its positive proposal is that a useful account of machine understanding should follow a method through reuse, explanation, and correction, keeping the information and work that make those achievements possible inside the account.

# A A complete trace of the instrument

Start from $(a, b, c) = (1, 3, 2)$. The first display is four. The base record is $(s, b) = (4, 3)$. Apply $H$, then $W$, then the newly admitted $J$.

| Event | Complete state | Display | Base record | Status |
|---|---|---|---|---|
| Initial | $(1, 3, 2)$ | 4 | $(4, 3)$ | $H/W$ contract adequate |
| After $H$ | $(4, 3, 2)$ | 2 | $(2, 3)$ | Exact base update |
| After $W$ | $(3, 4, 2)$ | 2 | $(2, 4)$ | Exact base update |
| Before predicting $J$ | $(3, 4, 2)$ | 2 | $(2, 4)$ | Missing $c$ distinction |
| After $J$ | $(0, 4, 2)$ | 4 | $(4, 4)$ | New reading acquired |

The last base record does not by itself retain all newly relevant information. The previous display was two and the new display is four, so $c = 4 - 2 = 2$. The post-operation first register is $4 - 4 = 0$. The enlarged display-centered record is therefore $(s, b, c) = (4, 4, 2)$.

A subsequent $H$ gives $(s, b, c) = (3, 4, 2)$ and display three. A subsequent $W$ leaves the display at three but changes the second coordinate to $3 - 4 = 4$ modulo five. A subsequent $J$ changes the display to zero. Each assertion follows from the enlarged update equations, not from extrapolating a list of displayed numbers.

## A.1 What was actually learned

The original prediction method already encoded the effect of $H$ and $W$. The $J$ observation supplied a new state distinction. It did not establish the $J$ transition law from scratch, because that law was part of the expanded analytical contract. Learning that law would require a candidate family and experiments capable of distinguishing it from alternatives.

This separation is important. Acquiring a latent value under a known rule, inferring an unknown rule, and transferring a known rule into a new notation are different research achievements. A single successful trace should not be counted as all three.

## A.2 A correction record for the trace

The record should retain the old scope, the new operation, the failed sufficiency condition, the acquired observation, and the exact post-intervention state reconstruction. It should also preserve the unchanged base equations. A concise corrected explanation would say that the third coordinate is irrelevant to $H/W$ continuations but relevant to any contract admitting $J$.

Nothing in that explanation establishes a hidden psychological state of an agent. It establishes the content and limits of a reusable mathematical model and gives a recipient an explicit way to use and revise it.

## A.3 The same trace with an uncertain readout

In the preceding trace, the state immediately before $J$ is $(3, 4, 2)$ and its exact display is two. Suppose the reasoner retains only the base record $(2, 4)$ and models $c$ as uniform. After $J$, the physical state is $(0, 4, 2)$. Under the noisy sensor of Section 12, a reported display of four gives the shifted report two. The resulting belief is over states $(c - 2, 4, c)$, not over unchanged pre-intervention states.

| Independent reports retained | One | Two | Three |
|---|---|---|---|
| Full shifted report 2 | 3/4 | 36/37 | 432/433 |
| Only even parity | 7/24 | 98/303 | 1372/4143 |

The entries are posterior mass on $c = 2$ along the named histories. Their probabilities of occurrence are not the entries in the table. Every other candidate remains possible after finitely many noisy reports. Duplicate copies of one reading count as one observation. The executable demonstration reproduces the fractions, checks the harmless ancillary flag, and distinguishes the full-state target from the recipient-relative communication target.

## B Companion verification and evidence boundaries

The companion program `code/finite_checks.py` uses integers and exact rational arithmetic. It checks the register construction, partition refinement against a separately coded pair-search procedure, finite transcript discrimination, a stochastic quotient, transport-error recurrences, explanation capacity, informative model filtering, and correction transport. The program enumerates small domains stated separately for each check group. It does not train a model or test an empirical claim about machine cognition.

The finite quotient checks compare the refinement procedure with direct search for a distinguishing word for every small machine in the stated domain. The register checks compare microscopic updates with independently written reduced coordinate updates. The correction checks enumerate maps, candidate sets, and possible evidence, including evidence inconsistent with the current set. Negative controls deliberately use a wrong swap update, a collapsed evidence channel, and a delayed disagreement beyond a fixed test horizon.

The expanded program `code/expanded_checks.py` adds rational Bayesian updates, likelihood factorization, posterior-mixture identities, rare-evidence bounds, recipient decision margins, adaptive history checks, graph coloring against direct decodability, and acquisition values against full policy enumeration. The conditional-information identity is checked using exact rational coefficients of logarithms of primes rather than floating-point cancellation. Graph and policy enumerations are exhaustive only over the small domains recorded per group.

The reference module `code/understanding.py` exposes the finite constructions with input validation. Its acquisition routine returns a policy witness or an explicit infeasibility result. A timeout or malformed input is not interpreted as mathematical infeasibility. The demonstration `code/demo_expanded.py` reproduces the noisy instrument and the complementary-background example.

The first-edition module `code/resource_contracts.py` implements fixed-decoder covers and nondominated compiled-tree profiles. Its witness checker follows every actual initial state independently of the optimizer. The suite `code/edition_checks.py` compares the optimizer with separately written full observation-history tree enumeration on the declared small domains, tests invalid witnesses, and checks all instrument states with the two frozen recipient programs. The costed demonstration is `code/demo_first_edition.py`. Source-file sizes and hashes are recorded without calling them minimal description lengths.

The script's result is evidence about the supplied code and finite constructions. The prose arguments supply the general conditional results. A clean compilation, an exact finite check, an independent proof review, and a formal-kernel build remain different kinds of evidence. This edition claims the first two, within the recorded review scope, and does not claim the latter two.

The source package records the actual test output, source hashes, build diagnostics, reference-review scope, and known limitations. Diagrams are generated from mathematical constructions. No synthetic image is offered as an empirical observation. The PDF's version number identifies a consolidated first edition, not a completed priority audit or an accepted publication.

# C Claim map

| Statement | Main assumption | Conclusion and boundary |
|---|---|---|
| Theorem 4.1 | Finite deterministic operations and all declared continuations. | Minimal behavioral quotient. Does not minimize acquisition or runtime. |
| Proposition 4.2 | Expanded contract on the same complete state space. | Necessary supplementary capacity. Sharp reuse construction has an explicit refinement condition. |
| Proposition 5.1 | Fixed complete transcript laws and equal priors. | Optimal binary identification bound. New evidence changes the laws. |
| Proposition 5.2 | Equal projected rows on each fibre. | Exact stochastic history preservation for reduced-history policies. |
| Theorem 6.1 | Matched operations and query decoders. | Exact forward transfer. Does not grant reverse recovery or target coverage. |
| Proposition 6.3 | Uniform residuals and Lipschitz target dynamics. | Finite-horizon error budget for matched operations. |
| Proposition 7.2 | One fixed recipient background and fixed-length messages. | Necessary communication capacity, not efficient decoding. |
| Proposition 9.1 | Exact realizable candidate family and separating experiments. | Finite informative elimination and decision-tree lower bound. |
| Theorem 11.1 | A commuting evidence square, including contradictory observations. | Filtering inclusion. Uniform equality exactly requires reflection of attainable outcomes. |

## C.1 Expanded correction and explanation results

| Statement | Main assumption | Conclusion and boundary |
|---|---|---|
| Theorem 8.1 | Finite source access, recipient backgrounds, and designable decoders. | Exact common alphabet via graph coloring. Does not establish decoder efficiency. |
| Theorem 10.1 | Finite deterministic experiments, explicit disturbance and admissibility, bounded horizon. | Minimum worst-case acquisition cost and an attaining policy. |
| Theorem 12.1 | Finite likelihoods and one reporting channel independent of the model. | All-prior posterior preservation exactly when conditional factorization holds. |
| Proposition 12.2 | The same joint experiment and a stated prior. | Average posterior divergence equals conditional information, not a worst-case bound. |
| Proposition 13.1 | Two joint laws and an event positive under both. | Conditional variation bound with an evidence denominator. |
| Proposition 13.3 | One posterior chain, finite actions, and bounded common losses. | Recipient regret and strict action-gap certificate. |
| Theorem 14.1 | Fresh conditionally independent observations and a reported-history selection policy. | Exact posterior preservation along adaptive inquiry and bounded public stopping. |
| Theorem 16.1 | Finite deterministic trees, costs, and receiver contracts. | Exact joint acquisition and common-message feasibility, not efficient general synthesis. |

### C.2 Fixed procedures and first-edition resource results

| Statement | Main assumption | Conclusion and boundary |
|---|---|---|
| Proposition 17.1 | Fixed interpreter with charged task-dependent fixed-length program and artifact. | Combined information lower bound. Not an individual-function runtime lower bound. |
| Theorem 18.1 | Finite source access, artifact catalogue, and fixed bounded recipient procedures. | Exact artifact intersections and minimum covering codebook. Raw artifact size remains charged. |
| Theorem 19.1 | Finite augmented state, bounded experiments, frozen recipients, vector charges, explicit controller trees. | Exact resource frontier and common workflow witnesses. Not optimal unrestricted program synthesis. |

### C.3 How the certificates depend on one another

The quotient and transfer results concern the declared operations. Their conclusions do not specify how an absent register is measured. The acquisition result adds that physical and informational route, and the correction results determine what an observation can eliminate or how it changes a belief. Neither step by itself guarantees that another reasoner can use the result.

The designable-recipient graph in Theorem 8.1 characterizes a communication alphabet after a decoder may be selected. Theorem 18.1 instead holds the recipient's procedure fixed. Its intersections can be empty even when the designable graph has a coloring. This is why a claim about one implemented recipient needs the stronger contract.

The final recurrence incorporates those same fixed terminal procedures into one experiment tree. Its witness must fit both online charges and the storage of every installed branch. A policy attaining one minimum and another policy attaining a second minimum are not interchangeable evidence for one simultaneous budget. The costed instrument supplies a common witness, while the four-origin counterexample supplies an explicit obstruction.

The probabilistic results are not silently inserted into that exact tree model. A noisy report may support a posterior and a bounded decision loss without identifying a unique state. Approximate posterior transfer therefore belongs to its stated likelihood, evidence, and recipient-loss assumptions. A zero-error state certificate requires its own stronger evidence. This separation is part of the first edition's scope, not an omission disguised by a shared diagram.

## References


[1] Komorebi P. *Machine Societies: A Theory of Persistent Plural Computational Organization*. Operation Komorebi, version 1.0.0, 2026. Companion theoretical monograph.

[2] Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? *Proceedings of ACL*, pp. 4198–4205, 2020. https://aclanthology.org/2020.acl-main.386/.

[3] Cosma Rohilla Shalizi and James P. Crutchfield. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *Journal of Statistical Physics*, 104, 2001. https://doi.org/10.1023/A:1010388907793.

[4] Robert Givan, Thomas Dean, and Matthew Greig. Equivalence Notions and Model Minimization in Markov Decision Processes. *Artificial Intelligence*, 147:163–223, 2003. https://engineering.purdue.edu/~givan/papers/mm.pdf.

[5] Naftali Tishby, Fernando C. Pereira, and William Bialek. The Information Bottleneck Method. *Proceedings of the 37th Allerton Conference*, 1999. Author preprint, 2000. https://arxiv.org/abs/physics/0004057.

[6] Sander Beckers and Joseph Y. Halpern. Abstracting Causal Models. *Proceedings of AAAI*, 2019. https://arxiv.org/abs/1812.03789.

[7] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal Abstraction: A Theoretical Foundation for Mechanistic Interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. https://jmlr.org/papers/v26/23-0058.html.

[8] John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. https://aclanthology.org/D19-1275/.

[9] Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability? Preprint, 2025. Version 1 consulted. https://arxiv.org/abs/2507.08802v1.

[10] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38, 2019. Author preprint: https://arxiv.org/abs/1706.07269.

[11] Dedre Gentner. Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7:155–170, 1983. Author-hosted copy: https://groups.psych.northwestern.edu/gentner/papers/Gentner83.2b.pdf.

[12] Dana Angluin. Learning Regular Sets from Queries and Counterexamples. *Information and Computation*, 75:87–106, 1987. https://doi.org/10.1016/0890-5401(87)90052-6.

[13] David Blackwell. Equivalent Comparisons of Experiments. *Annals of Mathematical Statistics*, 24(2):265–272, 1953. https://doi.org/10.1214/aoms/1177729032.

[14] Eigil F. Rischel and Sebastian Weichwald. Compositional Abstraction Error and a Category of Causal Models. *Proceedings of UAI*, PMLR 161:1013–1023, 2021. https://proceedings.mlr.press/v161/rischel21a.html.

[15] Hans S. Witsenhausen. The Zero-Error Side Information Problem and Chromatic Numbers (Corresp.). *IEEE Transactions on Information Theory*, 22(5):592–593, 1976. https://doi.org/10.1109/TIT.1976.1055607.

[16] Noga Alon, Avinatan Hassidim, Eyal Lubetzky, Uri Stav, and Amit Weinstein. Broadcasting with Side Information. *Proceedings of the 49th IEEE Symposium on Foundations of Computer Science*, pp. 823–832, 2008. https://arxiv.org/abs/0806.3246.

[17] Richard D. Smallwood and Edward J. Sondik. The Optimal Control of Partially Observable Markov Processes over a Finite Horizon. *Operations Research*, 21(5):1071–1088, 1973. https://doi.org/10.1287/opre.21.5.1071.

[18] Dennis V. Lindley. On a Measure of the Information Provided by an Experiment. *Annals of Mathematical Statistics*, 27(4):986–1005, 1956. https://doi.org/10.1214/aoms/1177728069.

[19] Paul R. Halmos and L. J. Savage. Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics. *Annals of Mathematical Statistics*, 20(2):225–241, 1949. https://doi.org/10.1214/aoms/1177730032.