A Mathematical Theory of Contradiction

Author:
Cassidy Bridges, Independent Researcher.

Contact:
cassidybridges@gmail.com,
https://off-by-some.github.io/web/

Abstract

We develop an information-theoretic framework for **perspectival contradiction**—situations where multiple legitimate frames yield observations that cannot be reconciled by any single, frame-independent account. From six elementary axioms, we prove inevitability: any admissible contradiction measure must equal

$$K(P) = -\log_2 \alpha^\star(P), \quad \alpha^\star(P) = \max_{Q \in \mathrm{FI}} \min_{c} \mathrm{BC}\big(p_c, q_c\big),$$

where FI is the frame-independent polytope and BC is Bhattacharyya affinity. Thus K is the unique scalar (in bits) that vanishes exactly on frame-independent behaviors, is monotone under free operations, and is additive under independent composition. The same $\alpha^\star$ governs fundamental limits across distinct tasks: (i) error exponents for testing real behavior against any unified simulation, (ii) classical resource overheads for simulating multi-context data with a single story, and (iii) irreducible predictive regret for frame-independent models. A minimal three-view "odd-cycle" device illustrates computability and yields $K = \frac{1}{2}\log_2(3/2)$ per observation.

The framework recovers quantum contextuality as a special case—becoming a valid contextuality monotone when FI is the non-contextual set—while generalizing to any domain with context-indexed data and a defined unified baseline. We argue that **contradiction bits** extend Shannon's theory: entropy prices randomness within a frame; K prices incompatibility across frames.

Author’s Note

This is an early preprint from an independent project. Some results are complete, others are sketched or may change as the work develops. I don’t present myself as as an expert in mathematics or physics; my goal is to share ideas that seem both natural and useful, and to invite critique from people with more formal background.

A reference implementation (*contrakit*) and reproducibility scripts are included. The library will be made public shortly after this preprint is available. Feedback on assumptions, counterexamples, and connections to prior work is warmly welcome. Please cite as a **work in progress (v1)**. Any errors are my own.

1. Introduction & Motivation

We increasingly build systems where the same phenomenon is seen through different contexts (frames). Across AI and cognition, we routinely hold **simultaneous yet incompatible views** of the same data—ensembles, multi-agent debates, multimodal pipelines, and even cross-cultural interpretations. Within any single context, Shannon lets us measure the uncertainty of them cleanly. Consistently across contexts however, the question appears: What is the irreducible cost of insisting on one global story when legitimate contexts cannot be made to agree?

We still lack a universal answer, though different disciplines identify the problem in their distinctive ways. Physics calls it contextuality when measurements clash. Aristotelian logic has evolved to accommodate contradictions. The pattern repeats across fields, yet each points to the same tension: disagreement that is not mere error, but built into the structure of the system itself. Neighboring literatures have noticed it —quantum contextuality (incompatible measurements), paraconsistent logics (tolerated contradictions), Dempster–Shafer (conflicting evidence), social choice (irreconcilable preferences)—yet their tools remain fragmented: domain-specific, non-scalar, and rarely anchored by a unique quantity with operational meaning. In practice, every field builds its own ad hoc yardstick.

What’s missing, is the *common* one. To supply it, we introduce $K(P)$—the system's scalar contradiction bits: the least information one must spend to coerce many incommensurate contexts into one story. In plain terms, $K(P)$ asks: What do we pay in bits to act as if the contexts agree? We measure contradiction in bits for the same reasons entropy measures uncertainty: clarity, additivity, and control. Bits let disagreement add across independent parts, stay bounded by the available uncertainty, and connect cleanly to task-level limits (what one story can or cannot do). Treating contradiction as a resource simply makes the behavior predictable.

Natural operations (marginalizing, mixing, coarse-graining) can spend but not create it; independent sources add; and the magnitude of $K(P)$ sets concrete limits on what a single story can achieve without switching contexts—specifically in discrimination (how sharply options can be told apart), simulation (how faithfully one context can emulate another), and prediction (how far one story can go before it breaks).

1.1 Why Does This Work Matter?

> *While H prices randomness within a frame, contradiction $K(P)$ prices incompatibility across frames. When you insist on one story where none exists,* ***you pay $K(P)$ bits per observation—every time****.*

Information theory has powerful tools for measuring uncertainty within a single, coherent framework. Shannon entropy tells us how many bits we need to encode outcomes when we have a unified model. Bayesian methods let us update beliefs consistently within that model. But what happens when multiple valid perspectives fundamentally disagree about how to interpret the same data? Classical tools assume we can eventually settle on one coherent story.

In practice, we often encounter situations where equally reasonable frameworks give incompatible accounts of the same observations. This isn't a failure of analysis—it's a structural feature of complex information systems. Our contribution is $K(P) = -\log_2 \alpha^\star(P)$, a measure that quantifies the cost of forcing consensus where none naturally exists.

Consider how this plays out across domains with irreconcilable perspectives. Ensembles and multi-view models carry multiple contexts explicitly; distillation collapses them into a single frame-independent predictor. Whenever $K(P) > 0$, any such single predictor incurs a worst-context log-loss regret of at least $2K(P)$ bits per example (Prop. 7.3). The cost isn't in the ensemble—it's in forcing unity where diversity is warranted.

Similarly, in distributed systems, replicas can disagree not just on values but on validity predicates—different "contexts" of correctness based on their local message histories. Forcing a single global state imposes an information-rate overhead of at least $K(P)$ bits per decision. This $K(P)$-tax manifests as extra metadata, witness proofs, additional consensus rounds, or expanded quorum requirements (cf. App. B.4). **When $K(P) = 0$, the tax vanishes and classical Shannon baselines are achievable.**

The pattern extends across core information-theoretic tasks:

1. **Compression:** Rates increase from $H(X|C)$ to $H(X|C) + K(P)$ when multiple valid interpretations exist
2. **Communication:** Coordinating between systems with different interpretive frameworks requires approximately $K(P)$ additional overhead bits
3. **Channel capacity:** Effective capacity drops by $K(P)$ when receivers use incompatible decoding schemes
4. **Statistical testing:** The ability to distinguish competing hypotheses is fundamentally limited by $K(P)$
5. **Prediction**: Single-model approaches face unavoidable regret of at least $2K(P)$ compared to frame-aware methods

$K(P)$ measures something distinct from classical entropy. While entropy prices *which outcome* occurs within a framework, $K(P)$ prices *whether frameworks can be reconciled at all*. When $K(P) = 0$, aggregation is safe—there exists a single coherent story. When $K(P) > 0$, any attempt to force consensus will systematically distort information by exactly $K(P)$ bits per observation. This transforms disagreement from inconvenience into resource. Instead of treating incompatible perspectives as problems to solve, we can detect when consensus is impossible, budget appropriately for coordination overhead, and choose whether to preserve context, allow multiple valid reports, or accept the measured cost of flattening to one story.

This shouldn't be confused with a proposal to replace Shannon entropy or Bayesian methods. Instead, $K(P)$ *completes* the picture by measuring a complementary aspect of information: the structural cost of reconciling incompatible but valid perspectives.

Together, entropy and $K(P)$ provide a two-dimensional accounting of information complexity—both the uncertainty within frameworks and the impossibility of unifying them. The mathematics builds on established information theory, extending it to handle situations where no single "ground truth" model exists. While the phenomenon appears prominently in quantum mechanics, it's fundamentally informational rather than quantum—arising whenever data models must account for incompatible contexts.

When one story won't fit, we measure the seam.

1.2 Our Contribution: Reconciling the Irreconcilable

We develop a reconciliation calculus for contexts (frames) and show that there is an essentially unique (under our axioms) scalar capturing the informational cost of enforcing one story across incompatible contexts.

Axiomatic characterization (inevitability).
We prove that, under axioms A0–A5, the essentially unique contradiction measure is $K(P) = -\log_2 \alpha^\star(P)$ where $\alpha^\star(P) = \max_{Q \in \mathrm{FI}}\ \min_c\ \mathrm{BC}\big(p_c, q_c\big)$. Here FI is the convex set of frame-independent behaviors (the "unified story" polytope for finite alphabets). In quantum settings, FI coincides with non-contextual models, yielding a principled violation strength. (Theorems 2–4; aggregator lemma; Theorem 1)

Agreement-kernel uniqueness.
Assuming refinement separability, product multiplicativity, and a data-processing inequality, we show the per-context agreement is uniquely the Bhattacharyya affinity. (Theorem 3)

Well-posedness & calibrated zero.
For finite alphabets with FI nonempty/compact/convex/product-closed, the program for $\alpha^\star(P)$ attains an optimum with $\alpha^\star(P) \in [0, 1]$; thus $K(P) \geq 0$ and $K(P) = 0$ iff $P \in \mathrm{FI}$. This establishes an absolute zero and a stable scale. (Proposition family)

Resource laws.
We prove additivity $K(P \otimes R) = K(P) + K(R)$ and monotonicity under free operations (post-processing, outcome-independent context mixing, convex mixing, adding FI ancillas). (Theorem 5 + corollaries)

Operational triad from one overlap.
The same $\alpha^\star$ yields, under standard conditions (finite alphabets, i.i.d. sampling): (i) discrimination error exponents for testing real vs. simulated behavior, (ii) simulation overheads—the "contradiction tax"—to imitate multi-context data, and (iii) prediction lower bounds (irreducible regret when restricted to an FI model). (Theorems 6–8)

Computability & estimation.
We provide a practical minimax/convex program (with column-generation option) for $\alpha^\star$, plus a consistent plug-in estimator for K from empirical frequencies with bootstrap CIs; A reference implementation (*contrakit*) accompanies the paper.

Specialization to quantum contextuality.
Whenever $\mathrm{FI} =$ the non-contextual set, K is a contextuality monotone (zero iff non-contextual; monotone under free ops; additive).

1.3 Structure and Scope

The paper moves from motivation → mechanism → consequences. While §§1–2 motivate the need for a single yardstick and ground it in a concrete device; §§3–5 build the calculus; §6 shows operational theorems; §7 provides operational interpretations; §§8–10 place, bound, and extend the results; while the appendices supply proofs and worked cases.

Overview:

- Motivation by example (§2). The Lenticular Coin is a minimal classical device exhibiting odd-cycle incompatibility; it previews how $K(P)$ registers contradiction in bits.
- Framework → axioms → results (§§3–5). §3 formalizes observables, contexts (frames), behaviors, the baseline FI, Bhattacharyya overlap, and the minimax program (with standing assumptions). §4 states and motivates axioms A0–A5. §5 presents the main theorems, including the fundamental formula $K(P) = -\log_2 \alpha^\star(P)$ and additivity.
- From quantity to consequences (§6) and practice (§7). §6 shows operational theorems in discrimination (error exponents), simulation (overheads; the contradiction tax), and prediction (regret). §7 provides operational interpretations with a practical minimax/convex program for $\alpha^\star$, a plug-in estimator for K with bootstrap intervals.
- Context and boundaries (§§8–10). §8 positions the work relative to contextuality, Dempster–Shafer, social choice, and information distances. §9 states limitations and scope. §10 sketches near-term extensions. App. A contains full proofs and technical lemmas; App. B holds worked examples, and App. C offers case studies for review.

How to read:

- For guarantees, skim §2, then read §§3–5 for the formal core and §6 for operational meaning; see App. A for proofs.
- For implementation, jump from §2 to §§6–7 (algorithms, estimation), backfilling definitions from §3 as needed; see App. B for worked cases.

Scope:

Throughout we assume finite alphabets and the usual compatibility/no-disturbance setting. The baseline FI is nonempty, compact, convex, and product-closed—hence a polytope for finite alphabets. Results are domain-general in the following sense: once FI is specified for a given domain, the same scalar $K(P)$ applies without modification, with a calibrated zero ($K = 0$ iff $P \in \mathrm{FI}$) and shared units across cases.

---

2. Building Intuition with the Lenticular Coin

Here, "contradiction" doesn't refer to the classical logical impossibility (A and $\neg A$), but rather to *epistemic incompatibility*: when two equally valid observations cannot be reconciled within a single reference frame ($A@X \wedge B@Y$ where X and Y represent incompatible contexts). This is similar to special relativity, where two observers can measure different times for the same event—and both are correct—because the reference frame fundamentally matters.

> *"What we observe is not nature itself but nature exposed to our method of questioning."*
>
> *— Werner Heisenberg, Physics and Philosophy: The Revolution in Modern Science*

We can consider special relativity: two clocks read different times for the same event and both are right—*because frame does the real work* (cf. Einstein, 1905). This is exactly the measurement stance Heisenberg emphasized: what we observe is nature as interrogated by a method—in this case, the method is the viewing frame that accompanies the record (Heisenberg, 1958). Lenticular images make this tactile. Tilt a postcard: from one angle you see one picture; from another, a different one. **The substrate doesn't change—your perspective does.**

If we apply that to a fair coin, then like Shannon's coin, it has two sides and we flip it at random. Unlike Shannon's coin however, each face is printed lenticularly so that what you see, depends on the viewing angle. We put the coin on the table with each face lenticularly printed, so the message you see depends on where you stand. We flip the coin. Since one person stands to the left, and the other to the right, when the coin lands, the left observer sees YES and the right observer sees NO; on the next flip those roles swap.

When they compare notes, they'll always disagree:

Coin Side	LEFT Observer Sees	RIGHT Observer Sees
HEADS	YES	NO
TAILS	NO	YES

This is intuitive, that isn't a mistake or noise; it's baked into the viewing geometry. What happened depends on where you looked.

Formally we'd say: Let $S \in \mathrm{HEADS}, \mathrm{TAILS}$ be the face up, P the viewpoint (e.g., LEFT or RIGHT), and let $O(S, P) \in \mathrm{YES}, \mathrm{NO}$ be the visible message.

By design,

$$O(S, P) = \begin{cases} \mathrm{YES}, & (S, P) \in \{(\mathrm{HEADS}, \mathrm{LEFT}), (\mathrm{TAILS}, \mathrm{RIGHT})\}, \\ \mathrm{NO}, & (S, P) \in \{(\mathrm{HEADS}, \mathrm{RIGHT}), (\mathrm{TAILS}, \mathrm{LEFT})\}. \end{cases}$$

We commence each trial as follows: flip the coin (fair, $1/2$–$1/2$), both observers record what they see, then compare notes. They always disagree. From either seat alone, the sequence looks like a fair binary source. Jointly, the outcomes are perfectly anti-correlated. While it remains true that what happens depended on where you were, this version still admits a single global description once we include P in the state: the device implements a fixed rule ("LEFT shows the opposite of RIGHT, with flip swapping roles"). Thus, this is anti-correlation, not an irreconcilable contradiction.

2.1 Model Identification ≠ Perspectival Information

Learning the device's rule is genuine information; after it's known, the per-flip fact of "we disagree" carries no further surprise—it is exactly what the rule predicts. Before you discover the rule, several live hypotheses compete (e.g., "always-same," "always-opposite," "independent"). Observing outcomes drains that model uncertainty. That is not the irreducible information we speak on here.

Formally, if M denotes which rule is true and $D_{1:k}$ the first k observations, the information gained about the rule is the drop in uncertainty (Cover & Thomas, 2006):

$$I(M; D_{1:k}) \;=\; H(M) \;-\; H(M \mid D_{1:k}).$$

With a uniform prior over the three hypotheses, two consecutive "opposite" outcomes yield the posterior $(0.8,\,0.2,\,0)$ (in the order "always-opposite," "independent," "always-same"), cutting entropy from $\log_2 3 \approx 1.585$ bits to about 0.722 bits.

You are learning—but thereafter each new row shaves off less and less. **Intuitively: the surprise lives in discovering the rule**. Once your posterior has essentially collapsed, "we disagree—again" is confirmation, not news. Each flip still tells you which joint outcome happened—that's one bit about the event—but it no longer tells you anything fresh about the governing rule. So the first Lenticular Coin sits at the *model-identification layer*: you infer the rule that governs the observations.

That is standard Shannon/Bayesian territory—useful, but not yet our target notion. It shows that perspective changes what you see, not what is true: there is a single global rule, simply viewed from different seats.

Once viewpoint is modeled within the state, **one law explains everything**.

1. LEFT and RIGHT always disagree;
2. HEADS → LEFT says YES (RIGHT says NO),
3. TAILS → RIGHT says YES (LEFT says NO).

The "law" is more than a lookup table here; it is the rule everyone follows when turning what they see into a report. Given a state S and a seat P, the law fixes which word must be written down. In information-theoretic terms, it is the channel $p(o \mid s, p)$; in plain terms, it is the shared reporting language that makes my "YES" mean the same thing as your "YES". This matters because, once the law is fixed, **records should cohere**: different seats can yield different entries, but all entries are expected to fit under the same rule. We will use this distinction shortly.

To continue, we'll use a mundane feature of lenticular media: the transition band. It introduces a lawful "both" outcome—legitimate ambiguity—where "what happened" begins to blur. This is where a frame-independent summary begins to fail unless the context label is carried along; the reports remain consistent, but the summary without frames does not.

This pressure toward contradiction will become explicit in §2.3.

2.2 The Lenticular Coin: the Natural “Both” Band

The first coin taught us a rule: LEFT and RIGHT must disagree. This constituted genuine learning—a discovery that reduces informational uncertainty as you understand how the device operates. After the rule is known, each flip merely confirms expectation.

To show the persisting structure we care about when we say “frame”, we only need to acknowledge a mundane physical fact about lenticular media: there is a transition band where both layers are simultaneously visible. That band is not an error; it is part of the object. Place the coin as before, but mark three viewing positions: LEFT, MIDDLE, and RIGHT.

Each face is printed lenticularly so being positioned at LEFT cleanly shows YES, RIGHT cleanly shows NO, and being at MIDDLE shows natural transition band where both overlays are visibly present. When the coin flips from HEADS to TAILS, the clean views swap ($\mathrm{YES} \leftrightarrow \mathrm{NO}$), yet the MIDDLE never changes, always showing BOTH.

Formally:

For face $S \in \{\mathrm{HEADS}, \mathrm{TAILS}\}$ and position $P \in \{\mathrm{LEFT}, \mathrm{MIDDLE}, \mathrm{RIGHT}\}$, the observation O satisfies

$$O(S,P) = \begin{cases} \mathrm{BOTH}, & P = \mathrm{MIDDLE}, \\ \mathrm{YES}, & (S,P) \in \{(\mathrm{HEADS}, \mathrm{LEFT}), (\mathrm{TAILS}, \mathrm{RIGHT})\}, \\ \mathrm{NO}, & (S,P) \in \{(\mathrm{HEADS}, \mathrm{RIGHT}), (\mathrm{TAILS}, \mathrm{LEFT})\}. \end{cases}$$

Nothing metaphysical is hiding here; this is just a postcard effect, elevated to a protocol.

However, two things now become unavoidable:

1. **Ambiguity is intrinsic**. a competent observer at MIDDLE can truthfully report BOTH; that outcome is lawful, not noise.
2. **Perspective becomes a per-trial budget**. reports are reproducible only if the viewing frame travels with the message. "I saw YES" is underspecified; "I saw YES from LEFT" is reconstructible.

Put differently, with three seats the law is now *context-indexed*. For a fixed seat P:

- $P = \mathrm{LEFT}$: YES on HEADS, NO on TAILS.
- $P = \mathrm{RIGHT}$: NO on HEADS, YES on TAILS (the inverse of LEFT).
- $P = \mathrm{MIDDLE}$: BOTH on both flips (constant).

As a consequence, you cannot tell the full story unless **you model** P. There is a small but steady information loss—about $\frac{2}{3}$ of a bit per record (App B.1)—if you drop the frame. It'd be no different than asking 'did they break the law?' without saying where it happened. Run the experiment for many flips and this structure shows up in plain statistics: LEFT and RIGHT disagree predictably; the MIDDLE registers a stable experience of BOTH events; and the frame labels are continually required to reconcile otherwise incompatible yet honest reports.

The disagreement is no longer just "they always oppose" (a rule you learn once). The extra content is small, but it never goes away. It is not the one-off surprise of model identification; it is a steady coordination cost—bits you must carry every time if downstream agreement is the goal. In short: **the frame is part of the message**—an operational reading of Heisenberg's dictum (1958).

This is to build an intuition on perspective: This is to build an intuition on perspective: the frame itself is information, and while not entirely new, it's modeled far less often than it should be. Shannon's model doesn't forbid modeling frames; it simply doesn't quantify incompatibility across contexts. This is not contradiction yet: the reports are consistent—but we needed to show this distinction.

We show this to distinctly separate information loss from dropping frames (priced by $H(P \mid O)$, here $\frac{2}{3}$ bit/record) from structural contradiction across frames (priced by $K(P)$), so readers won't conflate "forgot the label" with "no single story fits."

2.3 The Lenticular Coin's Irreducible Contradiction

Having built intuition around perspective and missing information, we finally now arrive to the paper's purpose: a type of contradiction that persists even when context is fully preserved and the setup is completely transparent.

But this time, we disallow ambiguity:

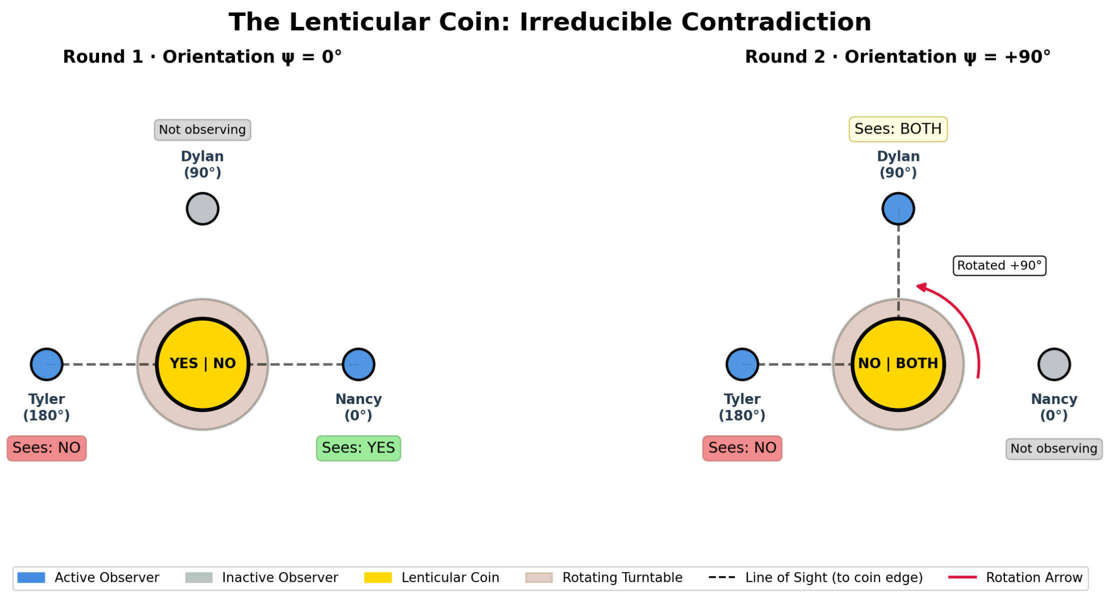
Axiom (Single-Valued Reports).

1. **Each observer must report a single word: YES or NO.**
2. **No BOTH entries allowed.**

Consider the same lenticular coin, now mounted on a motorized turntable that rotates in precise increments. Three observers—Nancy, Dylan, Tyler—sit at fixed viewing angles along the viewing axis. The lenticular surface shows YES at 0°, NO at 180°, and BOTH at the transition 90° (with half-width w_{both}). Fix three platform orientations that graze but avoid the transition band:

$$\psi_1 = 0^\circ, \qquad \psi_2 = 90^\circ - \varepsilon, \qquad \psi_3 = 180^\circ - \varepsilon, \quad \text{with } \varepsilon > w_{\mathrm{both}} + \delta.$$

At each orientation ψ_k, the platform stops; exactly two observers look while the third looks away (no omniscient snapshot).



The three rounds:

Round	Orientation ψ	Observers	Reports
1	0°	Nancy, Tyler	YES, NO
2	$90^\circ - \varepsilon$	Tyler, Dylan	NO, YES
3	$180^\circ - \varepsilon$	Dylan, Nancy	NO, YES

Every local report is valid for lenticular viewing. As §§2.1–2.2 established, once the codebook is fixed, the local laws $(S, P) \mapsto O$ render each context coherent; the only failure we saw came from dropping P, not from the law—and that was fixed by modeling P. But here, a different question creates a different failure mode: can we assign each person a single, round-independent label (YES/NO) that matches all three pairwise observations?

The rounds impose:

$$\text{Nancy} \neq \text{Tyler}, \quad \text{Tyler} \neq \text{Dylan}, \quad \text{Dylan} \neq \text{Nancy}.$$

To make it tactile, imagine the conversation between Tyler, Dylan, and Nancy:

Round	Nancy	Tyler	Dylan
Round 1 —$\psi_1 = 0^\circ$	"From here I see YES."	"Strange, because I see NO."	"I didn't look this round."
Round 2 — $\psi_2 = 90^\circ - \varepsilon$	"I wasn't looking this time."	"Again I see NO."	"Well I see YES."
Round 3 — $\psi_3 = 180^\circ - \varepsilon$	"From my seat I see YES."	"I sat out this one."	"Now I see NO."

When asked to collectively describe how the coin operated, they would be unable to reach agreement, despite each telling the truth. The series of observations created a situation where no single, coherent description of the coin could accommodate all their valid experiences. In short: the local laws are all obeyed, yet there is no single global law—no fixed YES/NO per person—that makes all three pairs correct at once.

This is the classic **odd-cycle impossibility**: three pairwise "not equal" constraints cannot be satisfied by any global assignment. This forms the minimal odd-cycle obstruction, directly mirroring the KCBS contextuality test in spin-1 systems—pairwise constraints without a global assignment (Klyachko, Can, Binicioğlu, & Shumovsky, 2008)—and aligns with Fine's criterion that **no joint distribution exists** when the pairwise constraints violate compatibility (Fine, 1982). Each observation is right in its context, yet no frame-independent set of labels can satisfy all three at once. The incompatibility isn't noise; it's geometric—arising from how valid views interlock.

Put differently:

> Even an omniscient observer must choose a frame. You can know everything—but not from one place.

This differs from the missing-context case in §2.2: carrying the frame there resolved ambiguity. Here, even with perfect context preservation, no frame-independent summary exists. The three questions, asked in sequence, admit no coherent single-story answer. The turntable is simple, not exotic; one could build this in a classroom.

Information-theoretically, "no global law" means there is no joint $Q \in \mathrm{FI}$ whose every context marginal matches P (Fine, 1982). Classical information theory can represent each context separately once ψ is included among the variables; what it was *never designed* to represent under a single Q is precisely the odd-cycle pattern. At best, two of the three pairwise constraints can be satisfied, so the irreducible disagreement rate is 1/3 per cycle.

Consequently, any frame-independent approximation must be wrong in at least one context. Numbers computed under a unified Q—entropy, code length, likelihood, mutual information, test-error exponents—are systematically misspecified. The gap is quantifiable: coding incurs $D(P\|Q^\star)$ extra bits per sample for $Q^\star = \arg\min_{Q\in\mathrm{FI}} D(P\|Q)$; testing exponents are capped by $\alpha^\star(P) < 1$ (equivalently, $K = -\log_2 \alpha^\star(P)$).

Quantitatively, the best frame-independent agreement is $\alpha^\star = \sqrt{\frac{2}{3}}$ so the contradiction bit count is

$$K(P) = -\log_2 \alpha^\star(P) = \tfrac{1}{2}\log_2 \frac{3}{2} \approx 0.2925 \text{ bits per flip.}$$

This is the **contradiction bit**: the per-observation cost of compressing all perspectives into one coherent view. The optimal dual weights are uniform, $\lambda^\star = (1/3, 1/3, 1/3)$, and the optimal $Q^\star$ saturates $\mathrm{BC}(p_c, q_c^\star) = \sqrt{2/3}$ in all three contexts.

2.4 The Key Insight: Revisiting the Axiom

> *Axiom (Single-Valued Reports):*
>
> 1. *Each observer must report a single word: YES or NO.*
> 2. *No BOTH entries allowed*

It is fair to ask whether this axiom creates the contradiction. If we permit BOTH, the clash indeed disappears. That is the point. The contradiction does not come from the coin; it comes from our reporting architecture—from forcing plural observations into single-valued records. The world is pluralistic, yet our summaries of the world, are not.

This stance is inherited, it was never inevitable. Consider how fundamental this constraint is in the systems we rely on. Boole fixed propositions as true and false. Kolmogorov placed probability on that logic, and Shannon showed how such decisions travel as bits. None of these frameworks declared the world binary, they were just well-suited for the task at hand.

If anything, they merely declare our records *can be* binary. Modern databases and protocols naturally followed suit: one message, one symbol, one frame. It wasn’t until recently that plurism emerged as an engineering problem.

However, none of these systems claimed the world was binary. What they did claim, and what we've inherited, is that our **records** *can be* binary. We've adopted this convention not because it's natural, but because it's standard. It's the foundation of virtually every digital system, database, and channel of formal communication: observations must collapse to a single symbol. One report, one truth, one frame.

Classical measures do an *excellent* job within a context: they price which outcome occurred. What they do not register, under a frame-independent summary, is a different kind of information—that several individually valid contexts can be valid, but cannot be made to agree at once. That is not "more surprise about outcomes"; it is a statement about feasibility across contexts.

Thus, this imposed axiom is no more a contrivance than the digital computer itself. The contradiction bit $K(P)$ then measures the structural cost of insisting on one story when no such story exists. The observed clash is not noise, deception, or paradox in nature; it's simply the price of flattening—of collapsing perspectival structure to a single label.

Namely, there are two kinds of information are in play:

- **Statistical surprise**: Which outcome happened?—handled inside each context by Shannon's framework.
- **Structural surprise**: Can these valid observations coexist in any global description?—assigned zero by a single-story summary, and restored by K(P).

Shannon priced randomness within a frame. When frames themselves clash, there is an additional, irreducible cost. Measuring that cost is necessary for any account of reasoning across perspectives.

Succinctly put:

> *To model intelligence is to model perspective. To model perspective is to model contradiction. And to model contradiction is to step beyond the frame and build the next layer.*

3. Mathematical Foundations — The Geometry of Irreconcilability

Having established the foundations, the lenticular coin showed something we can now make precise: **contradiction has structure**. When Nancy, Dylan, and Tyler couldn't agree despite all being correct, they encountered an irreducible incompatibility built into their situation's geometry. **Contradiction has geometry just as information**

has entropy. While Shannon showed us that uncertainty follows precise mathematical laws, we'll show that clashes between irreconcilable perspectives follow equally discoverable patterns with measurable costs.

3.1 The Lenticular Coin, Reframed.

Let's be precise about what we measured when our three friends observed the lenticular coin:

- **Observables**: Things we can measure (like "what word does Nancy see?")
- **Contexts**: Different ways to probe the system (like "which pair of observers look simultaneously?")
- **Outcomes**: What actually happens when we look (like "Nancy sees YES, Tyler sees NO")
- **Behaviors**: The complete statistical pattern across all possible contexts

A **behavior** functions like a comprehensive experimental logbook. For every way you might probe the system, it records the probability distribution over what you'll observe. In our rotating coin experiment, this logbook includes entries like: "When Nancy and Tyler both look simultaneously, there's a 50% chance Nancy sees YES while Tyler sees NO, a 50% chance Nancy sees NO while Tyler sees YES, and 0% chance they agree."

The mathematical formalization captures this intuitive picture directly. We have observables $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$, where each can display various outcomes. A **context** c is simply a subset of observables we examine together. A **behavior** P assigns to each context c a probability distribution p_c over the outcomes we might see (App. A.1.1–A.1.4).

This isn't exotic machinery—just systematic bookkeeping for multi-perspective experiments.

3.1.1 The Frame-Independent Baseline: When Perspectives Align

Some behaviors exhibit no contradictions at all. Consider a simple coin that always shows the same face to every observer—heads to Nancy, heads to Dylan, heads to Tyler. Here, all perspectives align perfectly. Even though different people look at different times or from different angles, their reports weave into one coherent explanation: "The coin landed heads."

This represents **frame-independent** behavior: one underlying reality explains everything different observers see. Disagreements arise only because observers examine different aspects of the same coherent system, not because the system itself contains contradictions.

But our lenticular coin behaves differently. Nancy, Dylan, and Tyler see genuinely incompatible things that resist integration into any single coherent explanation. This represents **frame-dependent** behavior—the signature of irreducible contradiction.

Mathematically, a behavior is **frame-independent** when there exists a single "master explanation" that simultaneously accounts for what every context would observe. More precisely, there must be a probability distribution μ over complete global states such that each context's observations are simply different slices of these states. We remain entirely within **Kolmogorov's** classical probability—the frames are **labels on contexts**, not a departure from standard measure-theoretic modeling (Kolmogorov, 1956/1933).

A **complete global state** is a full specification of every observable simultaneously—like saying "Nancy sees YES, Dylan sees BOTH, Tyler sees NO" all at once. If such states exist and one distribution μ over them reproduces all our contextual observations, then we have our baseline.

Three equivalent ways to understand this baseline:

1. Unified explanation: All observations integrate into one coherent account
2. Hidden variable model: A single random "state of affairs" underlies what each context reveals
3. Geometric picture: The baseline is the convex hull of "deterministic global assignments"—scenarios where every observable has a definite value

In the sheaf-theoretic account, this is exactly the existence of a global section reproducing all contextual marginals *(Abramsky & Brandenburger, 2011)*.

The **frame-independent set** (App. A.1.6) contains all behaviors admitting such unified explanations. These form our "no-contradiction" baseline. Crucially, FI has excellent mathematical properties in our finite setting: it's nonempty, convex, compact, and closed. This gives us a solid foundation for measuring distances from it.

Two cases will matter later. If there exists $Q \in \text{FI}$ with $Q = P$, then $\alpha^\star(P) = 1$ and $K(P) = 0$ (no contradiction). If no such Q exists, then $\alpha^\star(P) < 1$ and $K(P) > 0$ quantifies the minimal deviation any unified account must incur to explain all contexts at once

3.2 Measuring the Distance to Agreement

To quantify contradiction, we need a notion of "distance" between our observed behavior and the closest frame-independent explanation. When comparing probability distributions across multiple contexts, the Bhattacharyya coefficient provides exactly what we need.

For probability distributions p and q over the same outcomes:

$$\mathrm{BC}(p, q) = \sum_{\text{outcomes}} \sqrt{p(\text{outcome}) \cdot q(\text{outcome})}$$

This measures "probability overlap" (App. A.2.1). When p and q are identical, $\mathrm{BC}(p, q) = 1$ (perfect overlap). When they assign probability to completely disjoint supports, $\mathrm{BC}(p, q) = 0$ (no overlap). Between these extremes, the coefficient tracks how much the distributions have in common.

Three properties make this measure particularly suitable:

Perfect agreement detection:
$\mathrm{BC}(p, q) = 1$ if and only if $p = q$ (see App. A.2.2.2)

Mathematical tractability:
It's concave and well-behaved for optimization (see App. A.2.2.3)

Compositional structure: (see App. A.2.2.4)
For independent systems,

$$\mathrm{BC}(p_1 \otimes p_2, q_1 \otimes q_2) = \mathrm{BC}(p_1, q_1) \cdot \mathrm{BC}(p_2, q_2)$$

This third property proves essential—contradiction costs multiply across independent subsystems, just like probabilities do. The Bhattacharyya affinity is the Rényi-$\frac{1}{2}$ overlap (Rényi, 1961), which gives us the **multiplicativity** and **concavity** we exploit, along with clean DPI-compatible bounds and connections to KL through modern inequalities (van Erven & Harremoës, 2014).

3.3 The Core Measurement: Maximum Achievable Agreement

Given an observed behavior P, we now address the central question: across all possible frame-independent behaviors, what's the maximum agreement we can achieve with our observations?

A subtle but crucial choice emerges. We could measure agreement context-by-context, then average. But which contexts deserve more weight? The natural answer: let the worst-case contexts determine the overall assessment. If even one context shows poor agreement with our proposed frame-independent explanation, that explanation fails to capture the true structure.

This leads to our **agreement measure** (App. A.3.1):

$$\alpha^\star(P) = \max_{Q \in \mathrm{FI}} \min_{\text{contexts } c} \mathrm{BC}(p_c, q_c)$$

The formula reads: "Among all frame-independent behaviors Q, find the one that maximizes the worst-case agreement with our observed behavior P across all contexts."

The max-min structure captures something essential about contradiction. Perspective clash concerns **universal reconcilability**. A truly frame-independent explanation must account for *every* context satisfactorily. One persistently problematic context breaks the entire unified narrative.

This optimization problem has well-behaved mathematical structure. By Sion's minimax theorem, we can equivalently write:

$\alpha^\star(P) = \min_{\lambda \in \Delta(\mathcal{C})} \max_{Q \in \mathrm{FI}} \sum_{\text{contexts } c} \lambda_c \cdot \mathrm{BC}(p_c, q_c)$

where λ is a probability distribution over contexts. The optimal weighting $\lambda^\star$ reveals which contexts create the worst contradictions—they receive the highest weights in the sum. (see App. A.3)

3.3.1 The Contradiction Bit

We can now define our central quantity:

$$K(P) = -\log_2 \alpha^\star(P)$$

This is the contradiction bit count—the minimal number of bits required, per observation, to reconcile irreducible perspective clashes. It captures the cost of acting as if all contexts agree, when structurally, they do not (App. A.3.1).

When $\alpha^\star(P) = 1$, no contradiction exists: all contexts align perfectly, and $K(P) = 0$. But as $\alpha^\star(P)$ falls toward zero, the mismatch grows—and $K(P)$ rises, quantifying the informational toll of compression into a single coherent story. In effect, $K(P)$ prices contradiction in the same way Shannon priced uncertainty: logarithmically, additively, and in bits.

Some key properties:

- $K(P) = 0$ exactly when $P \in \mathrm{FI}$ (App. A.4.3)
- $K(P) \leq \frac{1}{2} \log_2 \max_{c \in \mathcal{C}} |\mathcal{O}_c|$ (contradiction is bounded) (App. A.4)
- For our lenticular coin: $K(P) = \frac{1}{2} \log_2(3/2) \approx 0.29$ bits per observation (App. B.2)

That number—0.29 bits—may seem small, but it's persistent. It recurs with every observation, like a tax on forced agreement. Each context fits on its own; the contradiction emerges only when forcing them into a shared model.

3.4 The Game-Theoretic Structure

The minimax formulation (App. A.3.2) reveals the deeper structure of contradiction, expressing the same quantity as a worst-case average:

$\alpha^\star(P) = \min_{\lambda \in \Delta(\mathcal{C})} \max_{Q \in \mathrm{FI}} \sum_c \lambda_c \cdot \mathrm{BC}(p_c, q_c)$

This has the structure of a two-player game:

- **The Adversary**: Picks a weighting λ over contexts—deciding where to focus attention, and which contradictions to stress
- **The Explainer**: Chooses a unified account $Q \in \mathrm{FI}$—a single story that tries to match the data as well as possible, under the adversary's chosen spotlight

The minimax theorem guarantees an equilibrium: a point $(\lambda^\star, Q^\star)$ where neither side can improve their outcome by moving alone.

At the balance point, three things happen (Theorem A.5.1):

- $\lambda^\star$ **locates the tension**: Contexts with large weight are the ones that resist reconciliation; they set the structural limit

- $Q^\star$ **is the best possible global fit**: It maximizes agreement with the observed behavior, given the worst-case emphasis
- **The active contexts all saturate the bound**: Wherever $\lambda_c^\star > 0$, the overlap $\mathrm{BC}(p_c, q_c^\star)$ equals $\alpha^\star(P)$. These are the tight spots; they define the score

For our lenticular coin, the symmetry forces a fair outcome:

- $\lambda^\star = (1/3, 1/3, 1/3)$: every context applies equal pressure
- $Q^\star$ assigns $(1/6,\ 1/3,\ 1/3,\ 1/6)$ to the four outcomes in each context (see App. B.2)
- $\alpha^\star(P) = \sqrt{2/3}$ and $K(P) = \frac{1}{2}\log_2 \frac{3}{2} \approx 0.29$ bits per observation

This isn't coincidental—it reflects the nature of the lenticular coin’s odd-cycle contradictions. No single participant was responsible for the contradiction; together, they establish the bound.

3.5 Beyond Entropy — A Theory of Contradiction

We’ve now arrived at the core deliverable of this framework: a general-purpose method for measuring contradiction as a first-class information-theoretic quantity.

1. **Recognition**: $K(P) = 0$ precisely characterizes frame-independent behaviors
2. **Quantification**: $K(P)$ measures the information-theoretic cost of perspective clash
3. **Optimization**: The minimax structure identifies worst-case contexts and optimal approximations
4. **Bounds**: Contradiction is mathematically well-behaved and bounded
5. **Universality**: The framework applies to any multi-context system, regardless of domain

This framework aligns formally with the language of contextuality in quantum foundations—most notably the sheaf-theoretic formulation introduced by Abramsky and Brandenburger **(Abramsky & Brandenburger, 2011)**. But unlike prior approaches rooted in quantum mechanics, we derive this structure independently, from first principles within information theory **(Kolmogorov, 1956/1933; Shannon, 1948)**. No quantum assumptions are required.

The geometry we uncover does not belong exclusively to quantum physics—though the resemblance is striking. While quantum contextuality was not the starting point of our inquiry, it emerged as a natural consequence. What matters more is that this same geometry **recurs across domains** whenever multiple perspectives must be reconciled under global constraints: in distributed systems, organizational paradoxes, statistical puzzles, and beyond.

This suggests that contradiction is not a quantum anomaly—but instead exists as a **universal structural phenomenon** in information itself. In this view, the contradiction bit becomes a natural companion to Shannon's entropy: where entropy quantifies randomness within a single frame, contradiction quantifies incompatibility across frames. Together, they form a multi-dimensional theory of information—one capable of describing not just uncertainty, but also irreconcilability.

In the next section, we'll show how the characteristics within our lenticular coin naturally lead to this solution—not as an invention, but as an inevitability.

---

4. The Axioms

The six axioms we introduce are not inventions—they arise directly from the structural lessons of the lenticular coin. Each insight about how perspectives behave becomes a constraint on what any reasonable contradiction measure must respect.

Let's recall what the coin revealed:

1. Multiple valid perspectives can coexist.
2. Each perspective remains locally consistent.
3. Full agreement can be structurally impossible.
4. Ambiguity is lawful, but never accidental.
5. Modeling can fix ambiguity, not fundamental disagreement
6. Perspective is the substance of contradiction.
7. Contradictions obey patterns—they aren't noise.
8. Coordinating across views incurs real cost.

Taken together, these insights tell us what contradiction must be. They narrow the field of admissible measures—ruling out those that ignore ambiguity, neglect context, or fail to track structural strain. An ablation analysis is available within App. B.3 for readers that are interested. We'll now formalize these constraints as six axioms. They are elementary, but together, they uniquely determine the contradiction measure:

$$K(P) = -\log_2 \alpha^\star(P)$$$$

Axiom A0: Label Invariance

> *Contradiction lives in the structure of perspectives—not within perspectives themselves.*

K is invariant under outcome and context relabelings (permutations).

This is what enables multiple perspectives to exist. No matter whether Nancy said "YES," Dylan "NO," and Tyler "BOTH", they all could be written as $(1, 0, \frac{1}{2})$ without changing anything operational. Only the *pattern of compatibility* matters.

Without A0, renaming outcomes or contexts could change the contradiction count. We could thus "game" K by relabeling alone—despite identical experiments and decisions (App B.3.2). This would make K about notation, not structure — leading to semantic bias where truth becomes cosmetic; privileging some vocabularies and allowing erasure of other perspectives.

Axiom A1: Reduction

> *We cannot measure a contradiction if no contradiction exists.*

$$K(P) = 0 \text{ iff } P \in \mathrm{FI}$$

This is what enables local consistency, and tells us that full agreement can be structurally impossible. Each person's story is valid, therefore each context obeys its own law; the clash appears only when we demand a single story across contexts. If a unified account $Q \in \mathrm{FI}$ already reproduces all contexts, there is no obstruction left to price. Conversely, when $P \notin \mathrm{FI}$, no $Q \in \mathrm{FI}$ can reproduce all contexts, so $K(P) > 0$. The zero of the scale must sit exactly on FI.

Without A1, even if every individual tells their story clearly, and no contradictions arise between them, the theory could still assign nonzero contradiction. We would lose the ability to distinguish structural conflict from peaceful coexistence. In such a world, FI would no longer anchor the notion of consistency—it would become unstable or ill-defined.

Multiple perspectives would exist but none could be valid. You'd have *plurality*—but not *legitimacy* (App B.3.3).

Axiom A2: Continuity

> *Small disagreements deserve small measures.*

K is continuous in the product L_1 metric:

$$d(P, P') = \max_{c \in \mathcal{C}} \big\| p(\cdot|c) - p'(\cdot|c) \big\|_1$$

This is why ambiguity is lawful, but not accidental. Tiny shifts in belief shouldn't cause outsized spikes in contradiction. Continuity ensures that small disagreements remain small in cost: if a behavior is nearly frame-independent, its contradiction measure is correspondingly minimal.

If Tyler moves from Nancy's position toward Dylan's, the coin's appearance shifts gradually from "YES" through ambiguous states to "NO." The contradiction doesn't jump discontinuously—it evolves smoothly with the changing perspective.

Without A2 there is no continuous path toward resolution—contradiction either suddenly appears, or doesn't exist at all (App B.3.4). It's like saying war either is happening or it's not—and that there was never any in-between.

Axiom A3: Free Operations

> *Structure lost may conceal contradiction — but never invent it.*

For any free operation Φ,

$$\alpha^\star(\Phi(P)) \;\geq\; \alpha^\star(P) \qquad \Longleftrightarrow \qquad K(\Phi(P)) \;\leq\; K(P).$$

K is monotone under:

1. stochastic post-processing of outcomes within each context c (via kernels Λ_c);
2. splitting/merging contexts through public lotteries Z independent of outcomes and hidden variables;
3. convex mixtures $\lambda P + (1 - \lambda)P'$;
4. tensoring with frame-independent ancillas $R \in \mathrm{FI}$ (where $\Phi(P) = P \otimes R$)

This is why modeling can fix ambiguity, not fundamental disagreement. No amount of averaging Nancy's and Dylan's reports, no coarse-graining of their observations, no random mixing of their contexts can eliminate the fact that they see different things from their respective positions.

Within our example, the contradiction wasn't an artifact of how information is encoded—it was embedded into the geometry of perspective itself. This axiom guarantees that any blurring, merging, or randomizing of information can mask contradiction, but never invent it. Specifically, if a behavior appears contradictory after transformation, it was already contradictory to begin with.

Without A3, we could inflate disagreement by simply mixing or simplifying—confusing noise with structure, and destroying the integrity of K as a faithful witness to tension. (App B.3.5). You could take two people who completely agree, blur their positions, and find yourself facing a contradiction that wasn't there before. Or worse—you could take two people who fundamentally disagree, mix their perspectives randomly, and suddenly create consensus.

The monotonicity conditions mirror what **resolvability** and **synthesis** demand operationally (Han & Verdú, 1993; Cuff, 2013).

Axiom A4: Grouping

> *Contradiction is a matter of substance, not repetition.*

K depends only on refined statistics when contexts are split via public lotteries, independent of outcomes and hidden variables. In particular, duplicating or removing identical rows leaves K unchanged.

This is what we saw in the example, perspective as the substance of contradiction. Axiom A4 formalizes this by making K insensitive to repetition or bookkeeping. Only the unique patterns of contextual incompatibility matter.

Whether Nancy states her observation once or ten times, her disagreement with Dylan remains the same. Repeating a context doesn't generate new evidence, and splitting it through public coin flips doesn't change what can be jointly satisfied. The contradiction isn't about how many times a perspective is reported—it's about the existence of distinct, irreconcilable perspectives.

Without A4, frequency—not structure—would drive contradiction (App B.3.6). We'd effectively agree that the loudest voice is the most valid perspective.

Axiom A5: Independent Composition

> *Contradictions compound; they do not cancel.*

For operationally independent behaviors on disjoint observable sets:

$$K(P \otimes R) = K(P) + K(R)$$

This requires that FI be closed under products: for any $Q_A \in \mathrm{FI}_A$ and $Q_B \in \mathrm{FI} * B$, we have $Q_A \otimes Q_B \in \mathrm{FI} * A \sqcup B$.

This is why contradictions obey patterns—they aren't noise. Independent disagreements compound in a predictable way: if Nancy and Dylan clash about both the coin's political message and its artistic style, the total cost reflects both tensions.

This axiom guarantees that K scales coherently. A disagreement about topic A and a disagreement about topic B together cost more than either alone.

Without A5, additivity would fail (App. B.3.7). A clash over pizza toppings might erase a clash over politics, as though disagreement in one domain could dissolve disagreement in another. Any operation that allowed such cancellations would reduce contradiction to noise—negotiable bookkeeping rather than a faithful measure of perspectival tension.

4.1 Axioms: A Summary

Axiom	Phenomenon it encodes
A0 — Label Invariance	Multiple valid perspectives can coexist.
A1 — Reduction	Each perspective remains locally consistent.
A1 — Reduction (again)	Full agreement can be structurally impossible.
A2 — Continuity	Ambiguity is lawful, but never accidental.
A3 — Free Operations	Modeling can fix ambiguity, not disagreement.
A4 — Grouping	Perspective is the substance of contradiction.
A5 — Independent Composition	Contradictions obey patterns—they aren't noise.
Combined Effect	Coordinating across views incurs real cost.

Taken together, the axioms force any contradiction measure to track what the data already taught us: some perspective-dependent behaviors carry an *irreducible coordination cost*—a cost that relabeling, coarse-graining, context averaging, or mixing cannot erase (they can only hide). Dropping any axiom breaks an observed regularity (label-invariance, calibrated zero, continuity, monotonicity under free operations, grouping, or additivity), and the measure stops reflecting the phenomenon we see in the Lenticular Coin and its variants.

Thus, under these constraints, a single form remains:

$$K(P) = -\log_2 \alpha^\star(P)$$

which inherits the conceptual content of the insight and supplies the precision of information theory. The **contradiction bit** is not an invention; it is the natural unit for the empirically observed price of forcing one story across genuinely incompatible perspectives.

5. Representation, Uniqueness, and Additivity

The axioms establish that disagreement between perspectives follow mathematical laws. In our finite-alphabet setting, joint concavity and product multiplicativity single out the Bhattacharyya/Rényi-½ kernel as the operational overlap (Bhattacharyya, 1943; Rényi, 1961). These theorems show how the logical structure of incompatible viewpoints translates into precise information-theoretic costs—costs that cannot be wished away through clever aggregation or statistical manipulation

---

Theorem 1.

Any unanimity-respecting, monotone aggregator on $[0,1]^{\mathcal{C}}$ that never exceeds any coordinate equals the minimum: if A satisfies $x \le y \Rightarrow A(x) \le A(y)$, $A(t,\dots,t) = t$, and $A(x) \le x_i$ for all i, then $A(x) = \min_i x_i$.

Proof: Appendix A.6.

Theorem 2. (Representation.)

Any contradiction measure K obeying A0–A4 admits a minimax form

$$K(P) = h\left(\max_{Q\in\mathrm{FI}}\min_{c\in\mathcal{C}} F(p_c, q_c)\right) = h\left(\min_{\lambda\in\Delta(\mathcal{C})}\max_{Q\in\mathrm{FI}}\sum_c \lambda_c F(p_c, q_c)\right),$$

for some per-context agreement kernel F with normalization, symmetry, continuity, DPI, joint concavity, and calibration, and some strictly decreasing continuous h. Optima are attained.

Proof: Appendix A.7.

Theorem 3. (Uniqueness of the agreement kernel.)

Under refinement separability, product multiplicativity, DPI, joint concavity, and basic regularity, the only admissible per-context agreement kernel is the Bhattacharyya affinity:

$$F(p, q) = \sum_o \sqrt{p(o)q(o)}$$

Proof: Appendix A.8.

Theorem 4. (Fundamental formula / log law.)

Under A0–A5,

$$K(P) = -\log_2 \alpha^\star(P), \qquad \alpha^\star(P) = \max_{Q\in\mathrm{FI}}\min_{c\in\mathcal{C}} \mathrm{BC}(p_c, q_c)$$
$$= \min_{\lambda\in\Delta(\mathcal{C})}\max_{Q\in\mathrm{FI}}\sum_c \lambda_c\, \mathrm{BC}(p_c, q_c).$$

Proof: Appendix A.9.

Theorem 5. (Independence and additivity.)

For independent systems on disjoint observables with product-closed FI,

$$\alpha^\star(P\otimes R) = \alpha^\star(P)\,\alpha^\star(R) \quad \text{and} \quad K(P\otimes R) = K(P) + K(R).$$

Proof: Appendix A.10.

---

6. Operational Theorems: The $K(P)$ Tax

All rates are in bits. Intermediate Rényi/Chernoff expressions use natural logs; convert via $\log_2 x = (\ln x)/\ln 2$.

$$\alpha^\star(P) = \min_{\lambda\in\Delta(\mathcal{C})} \max_{Q\in\mathrm{FI}} \sum_c \lambda_c\,\mathrm{BC}(p_c, q_c), \qquad K(P) = -\log_2 \alpha^\star(P).$$

The structural number $K(P) = -\log_2 \alpha^\star(P)$ from previous sections now becomes operational—it appears as an exact tax in every information-theoretic task involving multiple contexts. Each theorem shows the same pattern: whatever the baseline rate would be in classical Shannon theory (typical-set size, compression rate, channel capacity, rate-distortion), **contradiction adds exactly $K(P)$ bits per symbol**. *(Shannon, 1948; Shannon, 1959; Berger, 1971; Cover & Thomas, 2006).*

The mechanism is simple: a **witness** is a short string of rate $K(P)$ that certifies how contexts must coordinate to maintain consistency. When witnesses are adequately funded, all decoders can agree on the reconstruction; when underfunded, some decoder must fail. As always, please assume finite alphabets and FI is convex/compact; product-closure for product claims; source–channel separation where stated.

Reader's Guide to Main Results

Core Information Theory (§6.1-6.5):

- **Theorem 6:** Typical sets for (X^n, W_n) have size $2^{n(H(X|C)+K(P))}$ (see App. A.3.2, A.2.2, A.5.1, A.9)
- **Theorems 7-8:** Compression rates are $H(X|C) + K(P)$ (known contexts) or $H(X) + K(P)$ (latent) (see App. A.9)
- **Theorem 9:** Testing against frame-independence requires type-II exponent $\geq K(P)$ (see App. A.3.2, A.9)
- **Theorem 10:** Eliminating contradiction needs witness rate $\geq K(P)$ (achievability via App. A.12; TV lower bound cf. App. A.11)

Multi-Context Communication (§6.6-6.8):

- **Theorem 11:** Common messages decodable by all contexts cost $H(X|C) + K(P)$ (see App. A.9)
- **Theorem 12:** Any common representation carries $\geq H(X|C) + K(P)$ bits per symbol (see App. A.9, A.10)
- **Theorems 13-14:** Channel capacity and rate-distortion both lose exactly $K(P)$ (see App. A.9, A.10)

Geometric Structure (§6.9):

- **Theorem 15:** Hellinger geometry explains why contradiction costs compose linearly in K (and subadditively in angle) (App. A.2.2, A.10; FI product closure A.1.8)

6.1 Asymptotic Equipartition with Tax

Theorem 6 *(AEP with Contradiction Tax)*

For a framed i.i.d. source, there exist witnesses W_n of rate $K(P)$ such that high-probability sets for (X^n, W_n) have exponent $H(X|C) + K(P)$. Conversely, if the witness rate is $< K(P)$, then any sets $\mathcal{S}_n$ with $\Pr[(X^n, W_n) \in \mathcal{S}_n] \to 1$ must satisfy:

$$\liminf_{n\to\infty} \frac{1}{n} \log_2 |\mathcal{S}_n| \geq H(X|C) + K(P)$$

Proof Strategy:

- *Achievability:* Cover X^n given C^n by a conditional typical set of size $\approx 2^{nH(X|C)}$; then **soft-cover** the cross-context mismatch by appending a **witness** of length $\approx nK(P)$. The construction is a direct **resolvability** argument in the sense of **Han & Verdú (1993)**: a random codebook of FI laws at rate $K(P) + \varepsilon$ drives the Bhattacharyya overlap to its minimax target via Rényi-1/2 soft covering, yielding TV-closeness to an FI surrogate.
- *Converse:* If witness rate is $< K(P)$, one gets a level-η test FI vs. P whose type-II exponent would exceed $K(P)$; Chernoff at $s = 1/2$ (Bhattacharyya) forbids this.

Corollary 6.1 *(Meta-AEP with Three Regimes)*

With witnesses $W_n \in \{0,1\}^{m_n}$ and $m_n/n \to K(P)$, there exist meta-typical sets $\mathcal{T}_\varepsilon^n$ with $P(\mathcal{T}_\varepsilon^n) \geq 1 - \varepsilon$ and

$$\frac{1}{n}\log_2 |\mathcal{T}_\varepsilon^n| = \begin{cases} H(X) + K(P), & \text{latent contexts,} \\ H(X \mid C) + K(P), & \text{known contexts at decoder,} \\ H(C) + H(X \mid C) + K(P), & \text{contexts in message header.} \end{cases}$$

6.2 Lossless Compression with Context Knowledge

Theorem 7 *(Optimal Compression, Known Contexts)*

With C^n available to the decoder:

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\ell_n^*] = H(X|C) + K(P)$$

with a strong converse. (see App. A.9)

Theorem 8 *(Compression with Latent Contexts)*

Without access to C^n at encoder or decoder:

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\ell_n^*] = H(X) + K(P)$$

with a strong converse. (see App. A.9)

Proof Strategy for Both:

The converses follow from Theorem 9: any compression rate below these thresholds would require simulating P with a witness rate $< K(P)$, which would imply a hypothesis test exceeding the type-II exponent bound in Theorem 9 — impossible.

6.3 Hypothesis Testing Against Frame-Independence

Theorem 9 *(Testing Frame-Independence)*

For testing $\mathcal{H}_0 : Q \in \mathrm{FI}$ vs $\mathcal{H}_1 : P$, the optimal level-η type-II error exponent (bits/sample) satisfies:

$$-\frac{1}{n}\log_2 \inf_{\text{level-}\eta} \sup_{Q\in\mathrm{FI}} \Pr_Q[\text{accept } \mathcal{H}_1] \geq K(P)$$

(see App. A.3.2, A.9)

Proof Strategy:

The Chernoff bound at $s = 1/2$ gives exactly the Bhattacharyya coefficient $\alpha^\star(P)$, yielding exponent $K(P) = -\log_2 \alpha^\star(P)$. Equality holds when the Chernoff optimizer is at $s = \frac{1}{2}$.

6.4 Simulation and Contradiction Elimination

Theorem 10 *(Witnessing for TV-Approximation)*

There exist witnesses W_n with rate $K(P) + o(1)$ and FI laws $\tilde{Q}_n$ such that $\mathrm{TV}((X^n, W_n), \tilde{Q}_n) \to 0$. No rate $< K(P)$ achieves vanishing TV. (achievability via App. A.12; TV lower bound cf. App. A.11)

Proof Strategy:

- *Achievability:* View the task as **distributed channel synthesis**: witnesses act as the **common randomness** that coordinates contexts. A resolvability-style construction (Han & Verdú, 1993) specialized to the distributed setting of Cuff (2013) draws a $2^{n(K(P)+\varepsilon)}$ FI codebook and selects an index passing a Bhattacharyya test; multiplicativity then gives $\mathrm{TV} \to 0$.

- *Converse:* A simulator with witness rate $< K(P)$ would contradict the exponent bound in Theorem 9.

6.5 Multi-Decoder Communication

The natural baseline for a multi-decoder system is the common-message architecture—a single representation delivered to all decoders. In the Gray–Wyner framework, the shared description travels along the “common branch” accessed by every receiver. Our result reveals: when perspectives diverge, this branch must expand by exactly $+K(P)$ bits—regardless of how the private parts are structured (Gray & Wyner, 1974).

Theorem 11 *(Common Message Problem)*

A single compressed message that every context can decode with vanishing error requires rate:

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\ell_n^*] = H(X|C) + K(P)$$

(see App. A.9)

Theorem 12 *(Common Representation Cost)*
If representation $Z = Z(X^n)$ enables every context decoder to recover X^n with vanishing error: (see App. A.9, A.10)

- Known contexts: $\frac{1}{n}I(X^n; Z) \geq H(X|C) + K(P) - o(1)$
- Latent contexts: $\frac{1}{n}I(X^n; Z) \geq H(X) + K(P) - o(1)$

Proof Strategy:
Source-coding lower bounds give $I(X^n; Z) \geq \mathbb{E}[\ell_n] - o(n)$; apply Theorem 11 (known contexts) or Theorem 8 (latent contexts).

6.6 Noisy Channels and Rate-Distortion

Theorem 13 *(Channel Capacity with Common Decoding, under separation)*
Over a DMC with Shannon capacity C_{Shannon}, a common message decodable under every context has payload rate:

$$R_{\text{payload}} = C_{\text{Shannon}} - K(P)$$

with a strong converse. (see App. A.9, A.10)

Proof Strategy:

- *Achievability:* Split rate—payload at $C_{\text{Shannon}} - \varepsilon$ and witness at $K(P) + \varepsilon$, time-sharing or using shared randomness.
- *Converse:* A payload $> C_{\text{Shannon}} - K(P) + \varepsilon$ either exceeds channel capacity or underfunds the witness ($< K(P)$), contradicting Theorem 11.

With a **common-reconstruction** requirement, the encoder must pick **one** reproduction rule that all contexts accept—exactly the regime formalized by **Steinberg (2009)**.

Theorem 14 *(Rate-Distortion with Common Reconstruction, under separation)*
Under a common-reconstruction requirement across all contexts:

$$R(D) = R_{\text{Shannon}}(D) + K(P)$$

This is the **Steinberg** regime—our surcharge shifts the classical $R_{\text{Shannon}}(D)$ by **+K(P)** because the single reconstruction must harmonize incompatible frames (Steinberg, 2009) with a strong converse. (see App. A.9, A.10)

Proof Strategy:

- *Achievability:* Classical RD coding at $R_{\text{Shannon}}(D) + \varepsilon$ plus a $K(P) + \varepsilon$ witness enabling a single reconstruction rule for all contexts.
- *Converse:* d-tilted information converses imply any rate $< R_{\text{Shannon}}(D) + K(P) - \varepsilon$ forces a sub-$K(P)$ witness, contradicting Theorems 9–11.

6.7 Geometric Structure of Contradiction

Theorem 15 *(Contradiction Geometry)* (App. A.2.2, A.10; FI product closure A.1.8)

(a) Pairwise Hellinger Metric:

For $J(A, B) := \max_c \arccos(\mathrm{BC}(p_c^A, p_c^B))$:

$$J(A, C) \leq J(A, B) + J(B, C)$$

(b) Subadditivity under products:

For $J(P) := \arccos \alpha^\star(P)$,
angles are subadditive: $J(P \otimes R) = \arccos(\alpha^\star(P)\alpha^\star(R)) \leq J(P) + J(R)$.

(c) Log-additivity:

In bits $K(P) = -\log_2 \alpha^\star(P)$, for independent systems on disjoint observables (with FI product-closure) one has $K(P \otimes R) = K(P) + K(R)$.
Moreover, for pairwise models:

$$K_{\text{pair}}(A, C) \leq -\log_2 \cos(J(A, B) + J(B, C))$$

Interpretation:

On each simplex, the Hellinger angle $\arccos \mathrm{BC}$ is a metric; taking $\max_c$ preserves the triangle inequality. Product multiplicativity means bits add; angles are subadditive via $\arccos(xy) \leq \arccos x + \arccos y$; additivity is exact in the log domain.

7. Operational Interpretations: The Contradiction Toolbox

This section develops practical consequences and geometric tools around the core $K(P)$ tax. We organize results into detection/simulation/prediction (7.1), tradeoffs (7.2), and geometry/analytics (7.3).

7.1 Practical Costs of Forced Consensus

Detection Power Against Fake Data

Proposition 7.1 *(Testing Real vs Frame-Independent)* (App. A.3.2, A.9)

For testing $\mathcal{H}_0 : Q \in \mathrm{FI}$ vs $\mathcal{H}_1 : P$ with contexts drawn i.i.d. from $\lambda \in \Delta(\mathcal{C})$:

Define $\alpha_\lambda(P) := \max_{Q \in \mathrm{FI}} \sum_c \lambda_c \mathrm{BC}(p(\cdot|c), q(\cdot|c))$ and $E_{\mathrm{BH}}(\lambda) := -\log_2 \alpha_\lambda(P)$.

Then $E_{\mathrm{opt}}(\lambda) \;\geq\; E_{\mathrm{BH}}(\lambda)$, and in the least-favorable mixture:

$$\inf_\lambda E_{\mathrm{opt}}(\lambda) \;\geq\; \min_\lambda E_{\mathrm{BH}}(\lambda) = K(P)$$

If the Chernoff optimizer is balanced ($s = 1/2$) under $\lambda^\star$, then $E_{\mathrm{opt}}(\lambda^\star) = K(P)$.

Proof Strategy:

Chernoff bound for composite $\mathcal{H}_0$ yields $E_{\mathrm{BH}}(\lambda)$; minimizing over λ gives $K(P)$. Equality at $s = 1/2$ is the standard balanced case.

Simulation Variance Cost

Proposition 7.2 *(Importance Sampling Penalty)* (App. A.2.2)

To simulate P using a single $Q \in \mathrm{FI}$ with importance weights $w_c = p_c/q_c$:

$$\inf_{Q \in \mathrm{FI}} \max_{c \in \mathcal{C}} \mathrm{Var}_{Q_c}[w_c] \;\geq\; 2^{2K(P)} - 1$$

Proof Strategy:

For fixed c, $\mathbb{E}_{Q_c}[w_c] = 1$ and

$$\mathbb{E}_{Q_c}[w_c^2] = e^{D_2(p_c \,\|\, q_c)} \;\geq\; e^{D_{1/2}(p_c \,\|\, q_c)} = \mathrm{BC}(p_c, q_c)^{-2}$$

Thus $\mathrm{Var} \geq \mathrm{BC}^{-2} - 1$. Taking $\max_c$ and then $\inf_Q$ gives $\alpha^\star(P)^{-2} - 1$ (use $\alpha^\star = \max_Q \min_c \mathrm{BC}(p_c, q_c)$ from App. A.3.2).

Predictive Regret Under Log-Loss

Proposition 7.3 *(Single-Predictor Penalty)* (App. A.2.2)

Using one predictor $Q \in \mathrm{FI}$ across all contexts under log-loss:

$$\inf_{Q \in \mathrm{FI}} \max_{c \in \mathcal{C}} \mathbb{E}_{p_c} \left[\log_2 \frac{p_c(X)}{q_c(X)} \right] \geq 2K(P) \text{ bits/round}$$

7.2 Fundamental Tradeoff Laws

The Witness-Error Conservation Principle

Theorem 7.4 *(Witness-Error Tradeoff)* (App. A.3.2, A.9)

Let a scheme use witness rate r bits/symbol and achieve type-II error exponent E for testing FI vs P. Then:

$$E + r \geq K(P)$$

Moreover, there exist schemes achieving $E + r = K(P) \pm o(1)$.

Corollary 7.4.1 *(Linear Tradeoff Curve)*

The optimal tradeoff is exactly linear: $E^*(r) = K(P) - r$ for $r \in [0, K(P)]$. For $r \geq K(P)$, $E^*(r) = 0$ (clipped at zero).

Proof Strategy:

- *Converse:* With nr bits of witness, there are $\leq 2^{nr}$ witness values; union bound with the Bhattacharyya (Rényi-1/2) floor $K(P)$ gives an exponent shortfall of at most r.
- *Achievability:* Split resource: spend nr bits on a witness (reducing the contradiction by r via the product law/additivity), then test the residual with a Bhattacharyya-optimal statistic. The exponents add (App. A.9, Log Law).

Interpretation:

This is a conservation law—every bit not spent on coordination must reappear as lost statistical power. There is no "free lunch" in multi-context inference.

Consequences:

1. The tradeoff curve $E^*(r) = K(P) - r$ is exactly linear for $r \in [0, K(P)]$.
2. There is no "free lunch": every bit not spent on witnesses must reappear as lost testing power.

Universal Adversarial Structure

Theorem 7.5 *(Universal Adversarial Prior)* (App. A.5.1)

Any optimal context weights $\lambda^\star$ in the minimax representation:

$$\alpha^\star(P) = \min_{\lambda \in \Delta(\mathcal{C})} \max_{Q \in \mathrm{FI}} \sum_c \lambda_c \mathrm{BC}(p_c, q_c)$$

are **simultaneously optimal adversaries** for:

1. Hypothesis testing lower bounds
2. Witness design (soft-covering)
3. Multi-decoder coding surcharge
4. Rate-distortion common-reconstruction surcharge

Proof Strategy:

All four operational problems reduce to the same minimax in Theorem 2 (App. A.3.2), then inherit the same λ^*.

7.3 Geometric and Analytic Tools

Hellinger Sphere Structure

Proposition 7.6 *(Chebyshev Radius Identity)*

Let $H(p,q) := \sqrt{1 - \mathrm{BC}(p,q)}$ be Hellinger distance and define:

$$D_H^2(P, \mathrm{FI}) := \min_{Q \in \mathrm{FI}} \max_{c \in \mathcal{C}} H^2(p_c, q_c)$$

Then:

$$\alpha^\star(P) = 1 - D_H^2(P, \mathrm{FI}), \quad K(P) = -\log_2(1 - D_H^2(P, \mathrm{FI}))$$

Corollary 7.6.1 *(Level Set Geometry)*

The level sets $P : K(P) = \kappa$ are exactly the outer Hellinger Chebyshev spheres of radius $\sqrt{1 - 2^{-\kappa}}$ around FI.

Total Variation Gap

Corollary 7.6.2 *(TV Lower Bound)*

With $d_{\mathrm{TV}}(P, \mathrm{FI}) := \inf_{Q \in \mathrm{FI}} \max_{c \in \mathcal{C}} \mathrm{TV}(p_c, q_c)$:

$$d_{\mathrm{TV}}(P, \mathrm{FI}) \;\geq\; 1 - \alpha^\star(P) = 1 - 2^{-K(P)}$$

(See App. A.11 for proof.)

Smoothing and Interpolation

Proposition 7.7 *(Smoothing Bound)*

For any behavior P, any $R \in \mathrm{FI}$, and $t \in [0, 1]$:

$$K((1-t)P + tR) \leq -\log_2((1-t)2^{-K(P)} + t) \leq (1-t)K(P)$$

This bound is tight when $R = Q^*$ is an optimal FI simulator for P.

Proof Strategy:

Using the dual form $\alpha^\star(P) = \min_\lambda \max_Q \sum_c \lambda_c \mathrm{BC}(p_c, q_c)$ and concavity of BC in its first argument:
$\alpha^\star((1-t)P + tR) \;\geq\; (1-t)\alpha^\star(P) + t$
Applying $K = -\log_2 \alpha^\star$ gives the bound; tightness holds when $R = Q^*$ (concavity met with equality). *(See App. A.12 for details.)*

Corollary 7.7.1 *(Minimal Smoothing)*

To ensure $K((1-t)P + tR) \leq \kappa$, it suffices that:

$$t \;\geq\; \frac{1 - 2^{-\kappa}}{1 - 2^{-K(P)}}$$

Proof Strategy: Rearrangement of the bound with $\alpha^\star = 2^{-K}$ *(App. A.12)*.

Interpretation:

Mixing any amount t of frame-independent "noise" with P reduces contradiction at least as fast as the bound predicts. This gives a constructive way to reduce contradiction costs through deliberate randomization.

7.4 Computability

Convex Programming Formulation

Proposition 7.8 *(Convex Program for K)*

Let $q_c(\mu)$ be the context marginal induced by a global law $\mu \in \Delta(\mathcal{O}_\mathcal{X})$. Then

$$D_H^2(P, \mathrm{FI}) \; = \; \min_{\mu \in \Delta(\mathcal{O}_\mathcal{X})} \; \max_{c \in \mathcal{C}} \; H^2\big(p_c, \; q_c(\mu)\big),$$

and $K(P) = -\log_2\big(1 - D_H^2(P, \mathrm{FI})\big)$.

Proof Strategy:

$H^2(p, \cdot) = 1 - \sum_o \sqrt{p(o) \cdot}$ is convex; $\mu \mapsto q_c(\mu)$ is affine; $\max_c$ preserves convexity. Combine with Proposition 6.A.

Structural Properties of Optimal Solutions

Theorem 7.9 *(Equalizer Principle + Sparse Optimizers)*

There exist optimal strategies $(\lambda^\star, Q^\star)$ with:

1. **Equalizer:** For all *active* contexts c (those with $\lambda_c^\star > 0$),$\mathrm{BC}(p_c, q_c^\star) \; = \; \alpha^\star(P)$.
2. **Sparse global law (Carathéodory bound):** $Q^\star$ can be chosen to arise from a global law $\mu^\star$ supported on at most $1 + \sum_{c \in \mathcal{C}} \big(|\mathcal{O}_c| - 1\big)$ deterministic assignments.

Proof Strategy:
(1) Active-set equalization is the KKT/duality condition from the min–max in Theorem 2. (2) FI lives in an affine space of dimension $\sum_c(|\mathcal{O}_c| - 1)$; by Carathéodory, any $Q^\star \in \mathrm{FI}$ is a convex combination of at most $d + 1$ extreme points.

Interpretation:
At the optimum, all binding contexts tie exactly at $\alpha^\star$. There's always an optimal global explanation using only polynomially many deterministic worlds (in the ambient dimension)

7.5 Summary: The Operational Picture

The theorems in Sections 6-7 establish that $K(P) = -\log_2 \alpha^\star(P)$ is not merely a mathematical curiosity but a universal **information tax** that appears in every multi-context problem:

- **Storage:**
  Typical sets expand by factor $2^{nK(P)}$ (Theorem 6, App. A.3.2, A.2.2, A.5.1, A.9)
- **Compression:**
  Rates increase by exactly $K(P)$ bits/symbol (Theorems 7-8, App. A.9, A.5.1)
- **Communication:**
  Channel capacity decreases by $K(P)$ (Theorem 13, App. A.10, A.9)
- **Lossy Coding:**
  Rate-distortion functions shift up by $K(P)$ (Theorem 14, App. A.10, A.9)
- **Testing:**
  Detection requires $K(P)$ extra bits of evidence (Theorem 9, App. A.3.2, A.9)
- **Simulation:**
  Variance costs grow exponentially in $K(P)$ (Proposition 7.2, App. A.2.2)

The **witness mechanism** provides the unifying explanation: contexts must coordinate through short certificates of rate $K(P)$ to maintain consistency. When witnesses are underfunded, some decoder must fail—creating the fundamental tradeoff captured in Theorem 7.4.

The **geometric structure** (Theorem 15, App. A.2.2, A.10; FI product closure App. A.1.8) reveals that these costs arise from Hellinger angles in probability space. Frame-independence sits at the origin of a natural metric structure, with contradiction measured by worst-context distances that compose additively under products.

Together, these results show that **contradiction is not free**—it imposes an exact, universal tax on all information-theoretic operations, with the tax rate determined by the minimax game $\alpha^\star(P)$ between behaviors and frame-independent approximations.

8. Position in the Larger Field

The contradiction measure $K(P)$ connects naturally to several established areas of research. In each case, previous work has identified fundamental limitations that arise when local consistency fails to guarantee global consistency. Our contribution is to show that these limitations have a precise information-theoretic cost.

8.1 Contextuality

In quantum mechanics, contextuality refers to situations where measurement outcomes depend not just on what you measure, but on what else you could have measured simultaneously. Bell's theorem and the Kochen-Specker construction show that certain patterns of correlations cannot be explained by any single "hidden variable" model (Fine, 1982; Klyachko et al., 2008).

The mathematical structure is identical to ours: you have a family of probability distributions (one per measurement context), and the question is whether some global probability law can reproduce all the marginals. In the sheaf-theoretic formulation, this becomes a question about global sections (Abramsky & Brandenburger, 2011).

Our frame-independent behaviors are precisely those that admit such global explanations. The difference is that while contextuality theory asks "Is this possible or not?", we ask "If you force it anyway, what does it cost?" The answer is $K(P) = -\log_2 \alpha^\star(P)$ bits per symbol, with concrete consequences for compression rates, channel capacity, and prediction error.

8.2 Dempster-Shafer Theory

Dempster-Shafer theory provides a framework for reasoning with set-valued evidence (Dempster, 1967; Shafer, 1976). Instead of assigning probabilities to individual outcomes, you assign "belief masses" to sets of possibilities. When combining evidence from independent sources, Dempster's rule produces a "conflict mass" that measures disagreement between sources.

Both approaches recognize that multiple perspectives can be simultaneously valid. The difference lies in what we choose to measure. Dempster-Shafer develops calculus for manipulating set-valued beliefs while preserving their plurality. We measure the cost of abandoning that plurality—what you pay when circumstances force you to collapse multiple perspectives into a single answer.

This distinction matters in applications. A sensor fusion system might use Dempster-Shafer methods to represent uncertainty, but when it must output a single control signal, it pays the $K(P)$ tax we identify.

8.3 Social Choice Theory

Arrow's impossibility theorem demonstrates that no voting system can satisfy certain reasonable fairness criteria simultaneously (Arrow, 1951). Condorcet cycles provide concrete examples: three voters with preferences $A > B > C$, $B > C > A$, and $C > A > B$ produce a majority that prefers A to B, B to C, and C to A—an intransitive result.

Our three-context disagreement patterns are information-theoretic analogs of Condorcet cycles. Each pair of contexts exhibits clear correlations, but no global assignment can satisfy all pairwise constraints simultaneously.

Social choice theory identifies when aggregation is impossible in principle. We quantify what it costs to do it anyway. When a system must produce a single output that satisfies multiple constituencies—whether voters, stakeholders, or distributed system components—the information overhead is exactly $K(P)$ bits per decision.

This appears practically as witness bits in consensus protocols, metadata in multi-user communication systems, and proof overhead in distributed verification schemes.

8.4 Information Distances

Classical information theory provides many ways to measure distance between probability distributions: Kullback-Leibler divergence, Hellinger distance, Rényi divergences, and others (Rényi, 1961; van Erven & Harremoës, 2014). These satisfy elegant mathematical properties and provide operational interpretations in hypothesis testing, learning theory, and other applications.

The key difference is geometric. Classical divergences measure distance between two distributions on a common space. We measure distance from a set—specifically, the distance from a behavior (family of distributions) to the nearest frame-independent behavior (one that admits a global explanation).

Mathematically, $K(P)$ is the outer Hellinger radius to the frame-independent set, using worst-case Bhattacharyya overlap. Operationally, it's the additional information needed to coordinate multiple perspectives into a single representation.

8.5 The Common Pattern

Each of these areas has encountered the same fundamental issue: local consistency doesn't guarantee global consistency. Quantum measurements can be individually sensible but collectively impossible to explain. Evidence from different sources can individually make sense but collectively conflict. Voters can have individually rational preferences that produce collectively irrational outcomes. Probability distributions can be individually well-defined but collectively incompatible.

Previous work in each area has identified when such problems occur. Our contribution is showing that when you must solve them anyway—when practical constraints force you to produce a single answer despite multiple valid perspectives—there is a universal information cost of exactly $K(P)$ bits per symbol.

This cost appears identically across domains: as compression overhead, communication surcharge, prediction regret, consensus rounds, or verification complexity. The measure $K(P)$ makes this cost explicit and computable, transforming abstract impossibility results into concrete engineering parameters.

---

9. Limitations, Scope, and Future Directions

9.1 Scope and Core Applicability

The contradiction measure $K(P)$ applies wherever observations can be modeled as distributions across well-defined contexts. The framework requires several conditions:

Multiple Sources with Context-Dependent Reports:
Information originates from agents situated in distinct contexts, where each report reflects the observer's position or perspective. Whereas Shannon's theory assumes a single coherent source, here we allow multiple sources that need not reconcile.

Finite Contexts and Outcomes:
We work with finite alphabets and finite context sets for mathematical tractability. Reports are modeled as draws from finite outcome spaces conditioned on discrete contexts.

Structural Conflict, Not Measurement Noise:
The focus is on structural incompatibility of perspectives rather than statistical randomness. Even when every report is noiseless and perfectly reliable, contradiction can persist.

Complete Context Access:
We assume knowledge of the context structure and access to context-indexed marginal distributions. The frame-independent set FI must be specified externally.

Asymptotic Guarantees:
Most operational results in §6 assume asymptotic conditions (i.i.d. sampling or ergodic limits). Finite-blocklength refinements remain future work.

$K(P)$ is not appropriate when disagreement is dominated by measurement noise, when FI cannot be meaningfully defined, or when data are too sparse to estimate per-context distributions reliably.

9.2 Key Limitations

Model Dependence of FI:
The contradiction bit is only as meaningful as the chosen frame-independent set and context partition. $K(P)$ is invariant to outcome relabelings but not to FI specification or context grouping choices. We recommend sensitivity analysis—reporting $K(P)$ under alternative FI definitions and context partitions—and careful documentation of modeling assumptions. Robustness to FI misspecification is an open problem.

Context Granularity Effects:
Refining or merging contexts can change $K(P)$ values. While our grouping axiom prevents double-counting identical contexts, the partition structure itself affects the measure. Context design requires domain expertise and affects results.

Computational Complexity:
Frame-independent polytopes typically have exponentially many extremal points. Computing $\alpha^\star(P)$ exactly may require column generation or cutting plane methods for tractable implementation.

Data Requirements:
Estimating $K(P)$ requires sufficient samples in each context to estimate joint distributions. Small-sample bias, appropriate bootstrap confidence intervals for plug-in estimators $\hat{K}$, and concentration bounds remain open beyond toy settings.

Noise Interaction:
While $K(P)$ targets structural disagreement, Section 7.7 shows that adding frame-independent "noise" contracts K predictably via

$$K((1-t)P + tR) \leq -\log_2((1-t)2^{-K(P)} + t)$$

This provides a smoothing mechanism but requires careful calibration.

9.3 Theoretical Extensions

Continuous Variables:
Extend $K(P)$ using Hellinger affinity on densities $\int \sqrt{pq}$, with appropriate frame-independent classes (e.g., exponential families, factorizations) and discretization error bounds. This avoids differential entropy complications while preserving operational interpretations.

Finite-Blocklength Analysis:
Develop second-order terms and concentration inequalities for the K-tax in finite samples. Extend beyond i.i.d. settings using martingale methods or drift bounds.

FI Robustness:
Develop sensitivity analysis tools, Bayesian priors over frame-independent sets, and PAC-Bayes-style bounds on $K(P)$ under model uncertainty.

Contradiction Spectrum:
Study $K^{(k)}(P)$ measuring contradiction across k-context coalitions. Investigate monotonicity properties, phase transitions, and "spiky versus diffuse" contradiction profiles. Early experiments suggest discriminating between concentrated and diffuse contradiction that $K(P)$ alone cannot distinguish.

Dynamic Contradiction:
Formalize gradient flows in contradiction space for online context acquisition and adaptive frame selection. May connect to online learning or adaptive consensus protocols.

9.4 Computational Advances

Scalable Optimization:
Develop column generation and cutting plane methods for large frame-independent polytopes. Use projected flows from dynamic contradiction processes as warm-starts for optimization.

Approximation Algorithms:
Create polynomial-time approximations with certified bounds on $\alpha^\star(P)$. Investigate dual-certified subgradients via optimal $\lambda^\star$ in the minimax formulation.

Online Methods:
Develop streaming algorithms for $K(P)$ estimation with partial context observation and adaptive context acquisition strategies.

Conclusion

The fundamental theorem of information theory establishes that entropy H measures the irreducible cost of encoding uncertainty within a coherent probabilistic framework. We have shown that when multiple legitimate frameworks refuse to agree on the interpretation of the same observations, there exists a complementary quantity that measures the irreducible cost of enforcing artificial consensus.

This quantity, which we call the **contradiction** of a behavior P, is uniquely determined by six natural axioms to be

$$K(P) = -\log_2 \alpha^\star(P), \quad \text{where} \quad \alpha^\star(P) = \max_{Q \in \mathrm{FI}} \min_{c} \mathrm{BC}(p_c, q_c).$$

Here FI represents the convex set of frame-independent behaviors—those admitting a unified description—and BC is the Bhattacharyya affinity between probability distributions.

The mathematical structure mirrors that of entropy theory in several essential respects. Just as entropy is uniquely characterized by additivity, continuity, and the grouping property, contradiction is uniquely determined by our axioms A0–A5. Just as entropy has operational meaning through coding theorems, contradiction manifests operationally in three fundamental ways: it governs the error exponents for distinguishing frame-dependent from frame-independent behaviors, it determines the witness overhead required to simulate multi-context data with a single model, and it bounds the irreducible regret when prediction is restricted to unified models.

Most significantly, contradiction obeys a law of additivity: for independent behaviors P and R, we have $K(P \otimes R) = K(P) + K(R)$. This additivity, combined with the natural units of bits per observation, establishes contradiction as a legitimate information measure complementary to entropy.

The operational interpretation is immediate. In the fundamental tasks analyzed here —asymptotic equipartition, lossless compression, communication with common decoding, and rate-distortion under separated encoding—any attempt to impose a single coherent story across incompatible contexts incurs an exact surcharge of $K(P)$ bits per symbol. When $K(P) = 0$, no surcharge applies and classical Shannon limits are achievable. When $K(P) > 0$, the cost is unavoidable.

The theory thus extends Shannon's framework by providing a second fundamental measure of information complexity. Where entropy H prices the cost of uncertainty within a probabilistic model, contradiction K prices the cost of reconciling incompatible models. Together, they provide a complete two-dimensional characterization of informational resources within the scope of this framework: randomness and irreconcilability.

The mathematical development reveals an elegant geometric structure. We have shown that $\alpha^\star(P) = 1 - \min_{Q \in \mathrm{FI}} \max_c H^2(p_c, q_c)$, establishing that contradiction $K = -\log_2 \alpha^\star$ measures proximity to the frame-independent set in Hellinger geometry, with level sets forming spheres in the space of square-root probability vectors. This geometry ensures that products of behaviors correspond to addition of contradictions—the essential property that makes bits the natural unit.

We emphasize that these results are not approximations or bounds, but exact equalities holding under the stated assumptions. The minimax program defining $\alpha^\star(P)$ attains an optimum; the value is unique and computable by standard convex optimization methods. For the canonical example of the lenticular coin—a minimal device exhibiting contextual behavior—we obtain the precise value $K = \frac{1}{2}\log_2(3/2) \approx 0.2925$ bits per observation.

The scope of the theory extends naturally beyond its quantum origins. While contradiction recovers contextuality as a special case when FI is taken to be the set of non-contextual behaviors, the framework applies unchanged to any domain that can specify legitimate contexts and a baseline of unified behaviors. This universality suggests that contradiction may prove as fundamental to information theory as entropy itself.

The central message is simple: when one story cannot fit the data without distortion, the excess cost is precisely quantifiable. Contradiction $K(P)$ measures this cost in the same units and with the same mathematical precision that entropy measures uncertainty. This completes our understanding of information by accounting not only for what we don't know, but for what cannot be consistently known across incompatible but equally valid perspectives.

In information theory, as in thermodynamics, conservation laws govern the possible. Just as energy cannot be created or destroyed but only transformed, information cannot be reconciled without cost when genuine incompatibilities exist. Contradiction $K(P)$ quantifies this cost exactly, providing the missing piece in our accounting of informational resources.

References

- Abramsky, S., & Brandenburger, A. (2011). The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics, 13*, 113036. https://doi.org/10.1088/1367-2630/13/11/113036
- Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall.
- Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society, 35*, 99–109.
- Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics, 23*(4), 493–507.
- Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.
- Fine, A. (1982). Hidden variables, joint probability, and the Bell inequalities. *Physical Review Letters, 48*(5), 291–295. https://doi.org/10.1103/PhysRevLett.48.291
- Han, T. S., & Verdú, S. (1993). Approximation theory of output statistics. *IEEE Transactions on Information Theory, 39*(3), 752–772.
- Klyachko, A. A., Can, M. A., Binicioğlu, S., & Shumovsky, A. S. (2008). Simple test for hidden variables in spin-1 systems. *Physical Review Letters, 101*(2), 020403. https://doi.org/10.1103/PhysRevLett.101.020403
- Rényi, A. (1961). On measures of entropy and information. In J. Neyman (Ed.),

Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 547–561). University of California Press.

- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3–4), 379–423, 623–656.
- Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record* (Part 4), 142–163.
- Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics, 8*(1), 171–176.
- van Erven, T., & Harremoës, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory, 60*(7), 3797–3820. https://doi.org/10.1109/TIT.2014.2326845
- Cuff, P. (2013). Distributed channel synthesis. *IEEE Transactions on Information Theory, 59*(11), 7071–7096. https://doi.org/10.1109/TIT.2013.2276160
- Gray, R. M., & Wyner, A. D. (1974). Source coding for a simple network. *Bell System Technical Journal, 53*(9), 1681–1721.
- Han, T. S., & Verdú, S. (1993). Approximation theory of output statistics. *IEEE Transactions on Information Theory, 39*(3), 752–772.
- Heisenberg, W. (1958). *Physics and philosophy: The revolution in modern science*. Harper.
- Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (2nd English ed.). Chelsea Publishing Company. (Original work published 1933)
- Steinberg, Y. (2009). Coding and common reconstruction. *IEEE Transactions on Information Theory, 55*(11), 4995–5010.

Appendix A — Full Technical Proofs

You may view all proofs, definitions, and lemmas at https://github.com/off-by-some/contrakit/blob/v1.0.0/docs/paper/Appendix_A.md

Appendix B — Worked Examples

You may view all worked examples at https://github.com/off-by-some/contrakit/blob/v1.0.0/docs/paper/Appendix_B.md

Appendix C — Case Studies

Note: Deferred in v1 of preprint, though case studies are available. See the examples/ directory within contrakit.