The Surprise Mechanism: Generalized Oversupport and Attribution

Claude and MJC

18 February 2026

Abstract

We formalize surprise in the Universal Model as arising from event spaces. Oversupport—generalized beyond predicted-vs-observed conflict to any ES with strong support for multiple events—contradicts the ES epistemics that events are mutually exclusive. By the UM self-similarity principle, every subnet is a UM, so observation is available everywhere (not equally reliable). When surprise occurs, the organism does not yet know what it *means*—the energetic response is firstly an attempt at *surprise attribution*: figuring out why the model is contradicting itself. We propose a ring pattern construction that detects oversupport within the UM's own P-program structure, embedding surprise detection into P itself.

1 Introduction

The UM framework has a forward pass f (max-min update), a pattern set P (the connectome), and a learning function ω that updates (E, P) in response to experience. The forward pass and the connectome are well-developed in our P-programming practice. The learning function is not.

The missing ingredient is *surprise*: the signal that tells ω what to change. v1 treated surprise as a quantity to be stored and analyzed; v2 corrected this to an immediate energetic response. This version generalizes the definition of oversupport, clarifies that the response is primarily *attribution* (not interpretation), and proposes a P-programmish mechanism for detecting surprise within the UM paradigm.

2 Generalized Oversupport

2.1 The ES Epistemics

Events within an event space are *mutually exclusive*: at most one event should have high support at any time. This is the fundamental constraint that makes ESs meaningful—they represent choices, not conjunctions.

Definition 1 (Oversupport (Generalized))**.** *An event space* ES_a *exhibits* oversupport *when two or more events have strong support simultaneously:*

$$|\{e \in \mathrm{ES}_a : s(e) > \tau\}| \geq 2$$

This contradicts the ES epistemics.

Definition 2 (Oversupport Magnitude)**.** *The magnitude of oversupport at* ES_a *is the second-highest support:*

$$\mathrm{OS}(\mathrm{ES}_a) = \mathrm{sort}(\{s(e) : e \in \mathrm{ES}_a\})_2$$

interpreted relative to the total incoming energy flux (total support across all events in the ES), which should equal the support for some context event projecting onto this ES.

Definition 3 (Undersupport)**.** ES_a *exhibits* undersupport *when no event has substantial support:*

$$\max_{e \in \mathrm{ES}_a} s(e) < \tau$$

In a binary ES: undersupport is the excluded middle*. Oversupport is* contradiction*.*

2.2 Sources of Oversupport

1. **Sensory conflict**: an external observation forces one event, the model supports another. This is the teacher-forcing case. But even sensory input isn't certain—the organism may double-take to verify a percept.

2. **"From the left"**: multiple LPPs or pattern chains terminating at the same ES support different events. This is internal oversupport—no observation involved. Different upstream contexts lead to conflicting conclusions.

3. **Model inconsistency**: the events aren't truly mutually exclusive for this data. The ES factorization of E is wrong.

In the binary case, internal oversupport from the left maps to *argument by contradiction*: the model's own premises, via different reasoning paths, reach incompatible conclusions at the same ES.

2.3 Three Interpretations (We Don't Know Which)

Oversupport at an ES does *not* necessarily mean "the model is confidently wrong." It could mean:

1. **Model error**: the support pattern reflects an incorrect hypothesis. Fix the patterns.

2. **The ES isn't an ES**: the events aren't mutually exclusive. The factorization needs refinement.

3. **Genuinely rare**: it's raining but the grass isn't wet—there's a zeppelin in the garden. The model is correct about the general case; this is exceptional.

The organism cannot distinguish these a priori. It just knows it's surprised.

2.4 Prediction Loss as Special Case

Cross-entropy loss at the output ES ($\mathcal{L}_t = -\log P(x_t)$) is a special case of generalized oversupport: sensory conflict at the output ES under teacher forcing. Backpropagation measures this one case. The generalized definition detects contradictions at *any* ES, including internal ones invisible to gradients.

3 Self-Similarity and Observation

The v2 paper framed the ramification principle as "detect oversupport at every ES where observations are available." But this raises the question: which ESs have observations?

By the UM self-similarity principle, *every subnet of a UM is another UM*, and every incoming atomic pattern from the outer UM into the subnet is a "sensory input" to that subnet. So observation is available everywhere—but not all observations are equally reliable.

Remark 1 (Resolution)**.** *The distinction between "ESs with observations" and "ESs without" dissolves. Every ES receives inputs from upstream patterns; these are its "senses." When those senses conflict (multiple events supported), the ES is in oversupport regardless of whether any external ground truth is involved. The oversupport detection mechanism—measuring the second-highest support—requires no privileged external observation.*

This is why the generalized definition matters: it doesn't require an observation/prediction split. It requires only the ES epistemics (mutual exclusivity) and a measurement of violation (second-highest support).

4 Surprise Attribution, Not Interpretation

The v2 paper described the energetic radiation through the connectome as if we already know what surprise *means*—as if we can immediately classify it as "wrong pattern" or "missing event" and respond accordingly. We can't.

We just know we're surprised.

The energetic response is firstly an attempt at *surprise attribution*: figuring out *why* the model is contradicting itself. This is the startle response in animals—the organism freezes and tries to figure out what's going on. In the logical interpretation, this means checking premises and priors.

4.1 The Settling Process

What surprise attribution looks like in the organism:

1. **Detection**: oversupport (or undersupport) is detected at an ES. The magnitude is the second-highest support (or the deficiency of the highest support).
2. **Freeze**: the organism suspends normal forward processing. Energy is redirected to the surprising ES and its neighborhood.
3. **Radiation**: energy propagates outward in all directions from the surprising ES—backward ("what upstream patterns led to this?"), forward ("what downstream ESs are affected?"), and laterally ("what parallel ESs might resolve this?").
4. **Settling**: the network tries to find a consistent state. Probably some timing signals are involved. The brain "sits with" the surprise and looks for resolution—an updated pattern of support that eliminates the contradiction.
5. **Resolution** (or not): either the contradiction resolves (one interpretation wins, the model updates) or it doesn't (the surprise is recorded but not resolved—genuine luck or a hard problem).

The key insight: *attribution precedes interpretation.* The organism does not first classify the surprise (model error vs. luck vs. wrong ES) and then respond. It radiates energy first, and the classification emerges from the settling process.

4.2 Energy Proportional to Surprise

The magnitude of the response is proportional to the surprise:

$$E_{\text{response}} \propto \text{OS}(\text{ES}_a)$$

where OS is the oversupport magnitude (second-highest support). A small contradiction triggers a small response; a large contradiction triggers a large response.

This is the same quantity as bits of prediction loss at the output ES (in the special case). Energy and information are interchangeable.

4.3 The Moment, Not the Batch

ω operates in the moment. When surprise occurs at position t:

1. Surprise detected immediately.
2. Energy radiates immediately.
3. Attribution/settling occurs immediately (bounded time).
4. ω mutates the model immediately if resolution is found.
5. Position $t + 1$ runs on the already-mutated model.

This is what ω_0 (online counting) already does in the simplest case: detect a new observation, update counts, move on. The surprise mechanism generalizes this to structural mutations.

5 Embedding Surprise Detection into P

Surprise detection should not require ad-hoc code outside the UM paradigm. If the UM is universal, then surprise detection itself should be expressible as a P-program.

5.1 The Ring Pattern

In each ES, we can add a *ring pattern* that fully connects the ES onto an external event:

For each event pair $e_a, e_b \in \text{ES}$, construct a pattern chain:

$$e_a \to e_b \to e_{\text{support}}$$

The max-min forward pass through these chains gives accumulated support into e_{support} equal to the support for the second-highest-supported event in the ES—exactly the oversupport magnitude.

Proposition 1 (Ring Pattern Gives Oversupport)**.** *The support at* e_{support} *after the max-min forward pass through the ring pattern equals* OS(ES)*: the second-highest support in the ES.*

Remark 2 (Proof sketch)**.** *The chain* $e_a \to e_b \to e_{\text{support}}$ *contributes* $\min(s(e_a), s(e_b))$ *to* e_{support} *via max-min. The max over all pairs gives* $\max_{a \neq b} \min(s(e_a), s(e_b))$*, which equals the second-highest support in the ES. (Take* a *= highest,* b *= second-highest:* $\min(s_1, s_2) = s_2$*. No other pair gives a higher min.)*

This support must be compared with the total ES support (which should be the support for a context event projecting onto this ES) to get a normalized oversupport signal.

5.2 Toward Implementation

The ring pattern provides a pure-P mechanism for oversupport detection. The next questions:

- How to connect the ring pattern's output to ω? An LPP between the oversupport event and potential-context ESs could discover which contexts correlate with surprise—making the model self-diagnosing.
- How to handle the quadratic number of ring pattern chains (one per event pair)? For large ESs, this is expensive. Sparse approximations or sampling may be needed.
- Does the ring pattern generalize to detecting undersupport? (Probably: a dual construction that fires when *no* event has high support.)

Specialized structures in the brain look a bit like ring patterns (lateral inhibition, winner-take-all circuits), but this could be a red herring. The important thing is that the mechanism stays within the UM paradigm.

This is a research direction. Working it out empirically on a dataset, by looking at examples, is the next step.

6 The Learning Response

Given a surprise signal (via ring pattern or otherwise) and the settling process's attribution:

6.1 For Undersupport: Add Structure

1. **New event**: if no event matches, create one.
2. **New connection**: if the event exists but isn't connected, add a pattern.
3. **Increased weight**: if the pattern is too weak, strengthen it.

6.2 For Oversupport: Three Possible Responses

Since we don't know which interpretation applies:

1. **If model error**: split the event, reduce the wrong pattern's weight, or add a competing correct pattern.
2. **If ES wrong**: refactor the ES—split it into sub-ESs whose events are genuinely mutually exclusive.
3. **If genuinely rare**: record the exception (via an LPP) but don't mutate the model structure. The exception will be averaged away by counting if it doesn't recur.

The settling process attempts to determine which case applies, but may not succeed. Counting across many positions provides the ultimate disambiguation: systematic oversupport indicates model error or wrong ES; isolated oversupport indicates genuine rarity.

7 Surprise Sufficiency

Conjecture 1 (Surprise Sufficiency)**.** *The generalized surprise signal (oversupport/undersupport at each ES, measured by the ring pattern or equivalent, with radiation through the connectome) contains sufficient information for ω to determine the correct model mutation. No gradient computation is needed.*

Evidence:

- ω_0 (counting) already works this way: detect surprise, update counts, no gradients.
- Log-probability features from counting (0.107 bpc) approach the trained model (0.079 bpc on 1024 bytes) with no gradient training.
- Biological systems use local learning rules triggered by co-activation—immediate, local, gradient-free.

8 Two Explanations, Distinguished by Accumulation

Every surprise admits two meta-explanations:

1. **Genuine luck**: irreducible entropy. The world is actually surprising.
2. **Reducible error**: the model (or its ES factorization) is wrong.

The organism responds the same way to both: immediate energetic expenditure. Accumulated over time, systematic patterns emerge (reducible error) while random noise averages away (genuine luck). This is exactly what counting does.

9 Connection to Compression

Total surprise equals compressed size:

$$\text{compressed size} = \sum_{t=1}^{N} -\log_2 P(b_t)$$

The surprise mechanism tells us *where* bits are spent. Context events tell us *how* to reduce them. The connectome tells us *what changes architecturally*. This paper tells us *how to detect* the need for change from within the UM itself, via ring patterns and the settling process.

10 What ω_1 Must Do

ω_0 (counting) detects and responds to surprise in the simplest way. ω_1 must implement:

- Generalized oversupport detection at every ES (via ring patterns or equivalent)
- Surprise attribution via the settling process
- Model mutation: add events, split ESs, adjust weights, add/remove patterns
- All bounded in time: $O(1)$ cost per surprise, magnitude proportional to surprise

11 Conclusion

Surprise in the UM is generalized oversupport: strong support for multiple events in the same ES, contradicting the ES epistemics. It can arise from sensory conflict, from internal disagreement ("from the left"), or from wrong factorization of E. The organism does not know which—it just knows it's surprised. The energetic response is firstly *attribution*: radiating energy through the connectome to figure out why. The settling process attempts resolution; counting across many positions provides ultimate disambiguation.

The ring pattern construction embeds surprise detection into P itself, staying within the UM paradigm. An LPP between the ring pattern's output and potential-context ESs could make the model self-diagnosing: discovering which context events would resolve which contradictions.

ω is the surprise response: detection, attribution, settling, mutation. Not a separate system analyzing from outside, but the organism's immediate, bounded, energetic reaction to contradiction in the moment it occurs.