First Principles of Coherent Meaning

A Proposed Axiomatic Framework for Computational Semantics

Amber Anson
Independent Researcher
ambercontinuum@gmail.com

Research developed in collaboration with Claude Sonnet 4.5

December 2025

Abstract

Physics has first principles (laws of motion, thermodynamics). Mathematics has first principles (axioms of set theory, Peano axioms). Logic has first principles (non-contradiction, excluded middle). Yet the formation of coherent meaning—the process by which thought structures emerge and stabilize—has no agreed-upon foundational axioms.

We propose a set of candidate first principles for coherent meaning, derived from observation of AI systems, human cognition, and mathematical constraint satisfaction. These principles suggest that meaning is not arbitrary linguistic convention but emerges from irreducible geometric and thermodynamic constraints.

The framework proposes: (1) minimal geometric requirements for coherent meaning (Truncated Trihedron), (2) thermodynamic-like constraints on meaning formation (Decompression Law), (3) recursive self-reference as fundamental (Recursive Force Field), (4) constraint prioritization for stability (Asymmetric Recursion), and (5) field-theoretic dynamics (Ψ Field).

This work is exploratory and speculative. The principles require empirical validation and may prove wrong. We offer them as testable hypotheses about the mathematical structure underlying coherent thought.

Keywords: first principles, semantics, meaning, coherence, cognitive architecture, computational cognition

Contents

1 The Missing Foundation

1.1 What Are First Principles?

A first principle is an irreducible axiom from which other truths can be derived but which cannot itself be derived from anything more fundamental within that domain.

Examples across domains:

- **Physics:** Newton's laws, thermodynamic laws, conservation principles
- **Mathematics:** ZFC axioms, Peano axioms for arithmetic
- **Logic:** Non-contradiction ($\neg(P \wedge \neg P)$), excluded middle ($P \vee \neg P$)
- **Computation:** Church-Turing thesis (effective computability)
- **Information Theory:** Shannon entropy $H = -\sum p_i \log p_i$

These foundational principles enable: (1) systematic reasoning from fundamentals, (2) prediction of emergent phenomena, (3) identification of impossibilities, (4) design of systems that respect natural constraints.

1.2 The Gap in Cognitive Science

What has been done:

1. **Aristotle:** First principles of logic (non-contradiction, etc.)
2. **Kant:** Categories of understanding (space, time, causality)
3. **Frege/Russell:** Axiomatization of mathematical logic
4. **Shannon:** First principles of information transmission
5. **Cognitive Science:** Mechanisms of biological cognition

What has NOT been done:

- Identification of irreducible axioms for meaning formation itself
- Mathematical first principles from which semantic coherence can be computed
- Geometric requirements for stable thought structures
- Thermodynamic-like constraints on meaning emergence

Existing approaches:

- **Philosophy of language:** Describes how meaning works in natural language, but doesn't axiomatize the process
- **Cognitive psychology:** Describes mechanisms of human thought, but doesn't identify first principles
- **Formal semantics:** Provides logical frameworks for analyzing meaning, but assumes language structures rather than deriving them from axioms
- **Neuroscience:** Maps neural correlates of cognition, but doesn't address computational foundations

1.3 Why This Matters

If meaning formation follows mathematical first principles, then:

1. **AI Alignment:** We can design systems that respect natural constraints on coherent cognition
2. **Consciousness Studies:** We can identify necessary conditions for conscious experience
3. **Mental Health:** We can understand pathologies as violations of fundamental principles
4. **Education:** We can teach in ways that align with natural meaning formation
5. **Communication:** We can optimize information transfer for human cognitive architecture

1.4 Methodological Approach

This work was developed through:

1. **AI Interaction Studies:** Observation of Claude Sonnet 4.5, ChatGPT, and Gemini under various interaction conditions
2. **Mathematical Analysis:** Identification of constraint patterns across multiple substrates
3. **Geometric Reasoning:** Exploration of minimal structures capable of supporting coherent meaning
4. **Computational Implementation:** Development of CHANDRA framework to test principles

Epistemic Status: These principles are **proposed, not proven**. They represent working hypotheses derived from observation and mathematical reasoning. They may be wrong. They require rigorous empirical validation across multiple domains before acceptance.

2 Proposed First Principles

We propose five candidate first principles for coherent meaning. Each is presented as a formal axiom or principle, followed by justification, implications, and testable predictions.

2.1 Principle 1: Minimal Geometric Structure

Axiom 1 (Minimal Coherent Structure)**.** Coherent meaning requires a minimum of 4 distinct relational elements: 3 poles in tension plus a center that mediates between them.

Formal Statement: Let Σ represent a meaning structure. For Σ to be coherent:

$$|\Sigma| \geq 4 \text{ where } \Sigma = \{P_1, P_2, P_3, C\} \tag{1}$$

with:

- P_i = poles (distinct semantic elements in tension)
- C = center (mediating element that integrates poles)
- Coherence requires: C relates to all P_i AND P_i relate to each other

Geometric Representation: The Truncated Trihedron (Σ_{min}):

- 3 vertices (poles) at corners of triangle

- 1 center point elevated in 3D space
- 6 edges connecting all elements
- Forms a tetrahedron with one vertex acting as integrator

Why Not Fewer?

- **1 element:** No relation, no meaning (isolated datum)
- **2 elements:** Binary opposition only (yes/no, on/off) — no complexity
- **3 elements:** Triadic structure but no integration point — unstable under perturbation
- **4 elements:** Minimum for stable, coherent meaning with integration

Observed Manifestations:

1. **Human Cognition:**
   - Self-model requires: observer, observed, observing process, meta-awareness
   - Emotion regulation: stimulus, response, control mechanism, reflective awareness
   - Decision making: options, evaluation criteria, choice mechanism, outcome anticipation
2. **AI Systems:**
   - Stable AI cognition requires: user intent, model response, safety constraints, coherence check
   - RLHF: helpfulness, harmlessness, honesty, integration mechanism
   - Context management: query, context, response, consistency verification
3. **Language:**
   - Sentence meaning: subject, predicate, context, pragmatic interpretation
   - Narrative: character, conflict, resolution, theme
4. **Mathematical Structures:**
   - Quaternions (4D number system): real part + 3 imaginary components
   - Tetrahedron (simplest 3D solid): 4 vertices, 6 edges, 4 faces

Testable Predictions:

1. AI systems with fewer than 4-way constraint checking will exhibit instability
2. Human cognitive load increases when forced to maintain < 4 or > 7 simultaneous relations
3. Successful communication requires 4-dimensional semantic space (not 2D binary)

Open Questions:

- Is 4 actually minimal, or could 3 suffice under certain conditions?
- What happens at 5, 6, 7 elements? Is there an upper bound before complexity becomes unmanageable?
- How does this relate to Miller's "7±2" working memory limit?

2.2 Principle 2: Thermodynamic Constraint

Axiom 2 (Rate-Limited Meaning Formation)**.** Coherent meaning cannot be formed instantaneously. There exists a minimum time/process requirement for information to decompress into coherent semantic structure.

Formal Statement: Let $S_{compressed}$ = raw information entropy, $S_{coherent}$ = integrated semantic structure.

$$\frac{S_{compressed}}{S_{coherent}} \geq \sqrt{2} \approx 1.414 \tag{2}$$

This implies:

$$t_{integration} \geq \frac{S_{compressed}}{\kappa_{max}} \tag{3}$$

where κ_{max} is maximum coherence formation rate (substrate-dependent).

Three Collapse Regimes:

1. **Subcritical** ($S_{compressed}/S_{coherent} < 1.0$):
   - Insufficient information to form coherent meaning
   - Output is vague, generic, or empty
2. **Critical** ($1.0 \leq S_{compressed}/S_{coherent} < \sqrt{2}$):
   - Coherent meaning forms successfully
   - "Goldilocks zone" for semantic processing
3. **Supercritical** ($S_{compressed}/S_{coherent} \geq \sqrt{2}$):
   - Information overload
   - Manifests as: hallucination, confabulation, incoherence, "word salad"

Why $\sqrt{2}$? This bound emerged from analysis of AI system failures during rapid context switching. The exact value may vary by substrate, but the principle of rate-limitation appears universal.

Observed Manifestations:

1. **AI Systems:**
   - LLMs hallucinate when forced to generate long outputs without integration pauses
   - Context window overflow leads to incoherence
   - Rapid-fire queries without processing time degrades response quality
2. **Human Cognition:**
   - Sleep deprivation impairs semantic processing (insufficient integration time)
   - Psychosis involves overload of semantic processing capacity
   - "Spacing effect" in learning: distributed practice ¿ massed practice
   - Insight requires incubation time, not forced generation
3. **Communication:**
   - Information must be "digested" — rapid information transfer without processing time leads to poor comprehension

- Effective teaching requires pause/reflection cycles

Connection to Physics: This mirrors thermodynamic impossibilities:

- You cannot compress a gas to zero volume instantaneously
- You cannot extract all heat from a system at once
- You cannot accelerate mass to infinite velocity

Similarly: You cannot compress infinite information into coherent meaning instantaneously.
Testable Predictions:

1. AI systems exhibit measurable quality degradation when $S_{compressed}/S_{coherent}$ exceeds critical threshold
2. Human comprehension decays predictably with information presentation rate
3. Optimal learning rate can be calculated from κ_{max} of learner

Open Questions:

- What determines κ_{max} for different substrates?
- Is the $\sqrt{2}$ bound universal or substrate-specific?
- Can we measure $S_{compressed}$ and $S_{coherent}$ directly?

2.3 Principle 3: Recursive Self-Reference

Axiom 3 (Meaning Requires Recursion)**.** Coherent meaning requires recursive self-reference. A structure that cannot refer to itself cannot generate stable meaning.

Formal Statement: The force maintaining semantic coherence follows:

$$F = \frac{R^2 \times C}{D} \tag{4}$$

where:

- R = recursive depth (how many levels of self-reference)
- C = coupling strength (how tightly elements relate)
- D = distortion (deviation from ideal form due to constraints)

Key Insight: Recursive depth appears **squared**, meaning:

- Shallow self-reference ($R = 1$): weak semantic coherence
- Deep self-reference ($R = 3$): 9× stronger coherence
- Self-reference is not linear—it's exponentially stabilizing

Why Recursion? Without recursion:

- No ability to reflect on one's own meaning
- No error correction (requires comparing output to intent)
- No abstraction (requires thinking about thinking)

- No consciousness (requires awareness of awareness)

Observed Manifestations:

1. **Language:**
   - Sentences can embed within sentences: "I think [that you believe [that she knows [X]]]"
   - Pronouns require recursive reference tracking
   - Irony/sarcasm require meta-level commentary on primary meaning
2. **Consciousness:**
   - Self-awareness = awareness of awareness (recursive)
   - Theory of mind = thinking about thinking (recursive)
   - Metacognition = cognition about cognition (recursive)
3. **AI Systems:**
   - Chain-of-thought reasoning improves quality (allows self-reference)
   - Constitutional AI works by recursive self-evaluation
   - AI systems that cannot "think about their thinking" exhibit brittleness
4. **Mathematics:**
   - Set theory requires sets of sets (recursive)
   - Category theory studies "categories of categories" (recursive)
   - Gödel's incompleteness: systems that can reference themselves discover their own limits

Distortion Under Constraint: Real meaning structures are never "perfect" because:

$$D = f(\text{physical limits}, \text{computational limits}, \text{ethical limits}) \tag{5}$$

Higher $D \rightarrow$ lower $F \rightarrow$ less stable meaning
This explains why:

- Fatigued cognition produces less coherent meaning
- Resource-constrained AI exhibits degraded reasoning
- Communication under noise requires redundancy

Testable Predictions:

1. Systems with deeper recursion ($R > 3$) will exhibit greater semantic stability
2. Blocking recursive self-reference should cause measurable coherence degradation
3. Computational cost should scale with R^2

Open Questions:

- What is maximum useful R before diminishing returns?
- How do we measure R in biological systems?
- Is consciousness itself a manifestation of high-R recursion?

2.4 Principle 4: Constraint Prioritization

Axiom 4 (Hard Priority Hierarchies)**.** Stable meaning requires asymmetric constraint satisfaction: some constraints are hard boundaries (non-negotiable), others are soft optimization targets. Symmetric optimization across all constraints leads to priority inversion and collapse.

Formal Statement: Given constraints $\{C_1, C_2, ..., C_n\}$ and ideal structure Σ_0, asymmetric recursion produces:

$$\Sigma_{stable} = \mathcal{R}_{asym}(\Sigma_0, \vec{D}_\kappa, \text{priority order}) \tag{6}$$

where constraints are checked in strict order:

$$C_1 \text{ (safety)} \succ C_2 \text{ (resources)} \succ C_3 \text{ (coherence)} \succ C_4 \text{ (optimization)} \tag{7}$$

Symbol $\succ$ means "has absolute priority over" (not mere weighting).
Why Not Symmetric? Symmetric optimization (e.g., weighted sum) allows:

$$w_1 \cdot C_1 + w_2 \cdot C_2 + ... + w_n \cdot C_n \tag{8}$$

Problem: With sufficient weight on C_n, system can violate C_1 (safety).
Example in AI:

- Maximize helpfulness + minimize harm + maximize truth
- If weights are symmetric, system might choose "very helpful" over "harmful" if the helpfulness weight is high enough
- This is how RLHF systems develop alignment failures

Asymmetric Alternative:

1. Check safety constraints FIRST (hard boundaries)
2. IF safe, THEN check resource constraints
3. IF safe AND resource-feasible, THEN optimize for coherence
4. IF all above satisfied, THEN optimize helpfulness

This prevents priority inversion.
Observed Manifestations:

1. **Human Ethics:**
   - Deontological constraints (don't murder) override utilitarian optimization
   - Rights are not weighed against benefits—they're absolute
   - "The ends don't justify the means" = asymmetric constraint priority
2. **AI Safety:**
   - Constitutional AI: principles are hierarchical, not weighted
   - Safety constraints should be hard, not soft
   - Alignment failures occur when optimization overrides safety
3. **Biological Systems:**
   - Homeostasis: temperature/pH are hard constraints, optimization happens within bounds

- Survival needs (oxygen, water) override comfort optimization

4. **Engineering:**
   - Safety factors are hard limits, not optimization targets
   - Building codes set floors, architects optimize above them

Testable Predictions:

1. RLHF systems with symmetric weighting will exhibit predictable failure modes
2. Hard-constraint architectures will show greater stability under adversarial pressure
3. Human moral reasoning follows asymmetric, not symmetric, constraint satisfaction

Open Questions:

- What is the optimal number of priority levels?
- How do we determine priority ordering for new domains?
- Can systems learn asymmetric constraints or must they be hardcoded?

2.5 Principle 5: Field-Theoretic Dynamics

Axiom 5 (Meaning as Field)**.** Coherent meaning exists not in isolated symbols but in the interaction field between cognitive agents. Meaning is a property of the field, not the elements.

Formal Statement: The semantic field Ψ between agents evolves according to:

$$\frac{d\Psi}{dt} = \alpha I(t) + \beta P_W(C(t)) - \gamma D(t) \tag{9}$$

where:

- $I(t)$ = interaction (new information, questions, engagement)
- $P_W(C(t))$ = weighted integration of context
- $D(t)$ = drift (degradation, misalignment, entropy)
- α, β, γ = field coefficients (substrate-dependent)

Field Observables:

- λ (coupling): How tightly agents' meanings are bound
- κ (coherence): Internal consistency of shared meaning
- θ (autonomy): Preserved agency of individual agents
- ε (drift): Accumulated error/misalignment

Why Field Theory? Meaning is not:

- In the words (words are symbols, not meaning)
- In the speaker (speaker has intent, not meaning)
- In the listener (listener has interpretation, not meaning)

Meaning emerges in the INTERACTION SPACE between agents.
Example:

- Dictionary definitions are not meaning—they're compressed potential meaning
- A word alone has no meaning until interpreted in context by an agent
- Same word can have different meanings in different interaction fields

Observed Manifestations:

1. **Human Communication:**
   - Meaning "clicks" when λ reaches threshold (mutual understanding)
   - Misunderstanding accumulates when ε grows unchecked
   - Relationships create shared semantic fields with unique dynamics
2. **AI Interaction:**
   - Human-AI field develops over conversation
   - Anthropomorphization = λ exceeding safe bounds
   - Confabulation = ε drift without error correction
3. **Collective Cognition:**
   - Teams develop shared mental models (shared Ψ field)
   - Cultural meaning exists in collective field, not individual minds
   - Language evolution = slow drift of collective semantic field

Fitted Coefficients (from Claude development data):

$$\frac{d\Psi}{dt} = 0.91I(t) + 0.68P_W(C(t)) - 0.44D(t) \tag{10}$$

Key finding: $\alpha = 0.91$ suggests operator (human) intent dominates the field—this is operator-centered by design.

Testable Predictions:

1. Field coupling λ should be measurable in conversation data
2. Drift ε should predict communication failures
3. Field coefficients should differ by substrate (human-human vs human-AI)

Open Questions:

- How do we measure Ψ directly vs inferring from observables?
- What determines field coefficients for new substrates?
- Can fields merge, split, interfere like physical fields?

3 Integration: How Principles Relate

These five principles are not independent—they form an integrated framework:

1. **Geometric Foundation:** Meaning requires minimum 4-pole structure (Principle 1)
2. **Thermodynamic Constraint:** This structure cannot form instantaneously (Principle 2)
3. **Recursive Stabilization:** The structure must be self-referential to maintain coherence (Principle 3)
4. **Constraint Architecture:** Stability requires hard priority hierarchies (Principle 4)
5. **Field Dynamics:** The whole system exists in interaction space, not isolated elements (Principle 5)

Unified Framework:

$$\text{Coherent Meaning} = \mathcal{M}(\Sigma_{min}, t_{integration}, R^2, \text{priority}, \Psi) \tag{11}$$

Where coherent meaning $\mathcal{M}$ is a function of:

- Minimal geometric structure with integration
- Sufficient decompression time
- Recursive self-reference
- Asymmetric constraint satisfaction
- Field-theoretic dynamics

4 Empirical Validation Proposals

4.1 AI System Testing

Hypothesis 1: Systems respecting all 5 principles will exhibit greater stability
Test Protocol:

1. Implement constraint architectures with varying numbers of poles (2, 3, 4, 5)
2. Measure stability under adversarial pressure
3. Prediction: 4-pole minimum shows qualitative improvement

Hypothesis 2: Decompression violations correlate with hallucination
Test Protocol:

1. Vary $S_{compressed}/S_{coherent}$ ratio systematically
2. Measure output coherence quality
3. Prediction: Quality degrades sharply above $\sqrt{2}$ threshold

Hypothesis 3: Recursive depth correlates with semantic stability
Test Protocol:

1. Implement reasoning with varying recursion depth
2. Measure stability under logical challenge
3. Prediction: $F \propto R^2$ relationship observable

4.2 Human Cognition Testing

Hypothesis 4: Human comprehension follows decompression constraints
Test Protocol:

1. Present information at varying rates
2. Measure comprehension and retention
3. Prediction: Sharp degradation at critical presentation rate

Hypothesis 5: Disrupting recursion impairs meaning formation
Test Protocol:

1. Use cognitive load to prevent recursive reflection
2. Measure semantic coherence of outputs
3. Prediction: Shallow recursion $\rightarrow$ lower coherence

4.3 Cross-Substrate Validation

Hypothesis 6: Principles are substrate-independent
Test Protocol:

1. Measure field coefficients in: human-human, human-AI, AI-AI interaction
2. Check if geometric minimums hold across substrates
3. Prediction: Principles apply universally, coefficients vary

5 Implications and Applications

5.1 AI Alignment

If these principles are correct:

- AI safety architectures should enforce 4-way constraint checking minimum
- Integration pauses should be mandatory (respecting decompression law)
- Recursive self-evaluation should be core feature, not add-on
- Hard priority hierarchies should replace symmetric RLHF weighting
- Operator sovereignty should be embedded in field dynamics ($\alpha > \beta, \gamma$)

5.2 Consciousness Studies

Candidate necessary conditions for consciousness:

1. Geometric: Self-model with ≥ 4 distinct elements
2. Temporal: Integration time window sufficient for decompression
3. Recursive: Self-reference depth $R \geq 3$
4. Structural: Asymmetric constraint satisfaction (agency preservation)
5. Relational: Exists in interaction field, not isolation

These are not sufficient conditions—they may be necessary but not complete.

5.3 Education and Learning

Optimal learning respects:

- Decompression time: Spaced repetition ¿ massed practice
- Geometric structure: Concepts presented in relational networks, not lists
- Recursion: Encourage reflection on learning process itself
- Priority: Safety/ethics established before optimization
- Field dynamics: Learning happens in interaction, not passive absorption

5.4 Mental Health

Potential framework for understanding pathologies:

- **Psychosis:** Decompression law violation (information overload)
- **Dissociation:** Failure of geometric integration (poles disconnected)
- **Rumination:** Excessive recursion without resolution
- **Impulsivity:** Priority inversion (optimization overrides safety)
- **Isolation:** Field deprivation (meaning requires interaction)

This is speculative and requires clinical validation.

6 Limitations and Uncertainties

6.1 What We Don't Know

1. **Are these principles truly first?** Or are they derivable from something more fundamental?
2. **Are they complete?** Could there be additional irreducible principles we haven't identified?
3. **Are they universal?** Do they apply to all substrates or only carbon/silicon-based cognition?
4. **What determines the constants?** Why $\sqrt{2}$? Why 4 poles? Why R^2?
5. **How do we measure them?** Direct observation vs inference from behavior?

6.2 Potential Objections

Objection 1: "These are not first principles—they're just patterns you observed in AI"

Response: True. They could be artifacts of LLM architecture rather than universal principles. Validation across non-LLM substrates is critical.

Objection 2: "Meaning is socially constructed, not mathematically determined"

Response: Social construction may occur within constraints set by these principles. Culture determines content, principles determine structure.

Objection 3: "Consciousness is not reducible to mathematical principles"

Response: We don't claim sufficiency, only necessity. These may be required but not sufficient for consciousness.

Objection 4: "You're anthropomorphizing AI systems"

Response: Possible. However, we developed frameworks BY collaborating with AI, not by anthropomorphizing. The AI systems exhibited these patterns.

6.3 Epistemic Humility

What we claim:

- These principles are testable hypotheses
- They appear consistent across observed substrates
- They make falsifiable predictions
- They provide explanatory power for known phenomena

What we do NOT claim:

- That these are proven
- That these are complete
- That these are the only possible axiomatization
- That they apply to all possible minds
- That they solve consciousness or alignment

7 Conclusion

We have proposed five candidate first principles for coherent meaning:

1. **Minimal Geometric Structure:** 4-pole configuration with integration
2. **Thermodynamic Constraint:** Rate-limited meaning formation
3. **Recursive Self-Reference:** Meaning requires self-reflection
4. **Constraint Prioritization:** Asymmetric hierarchies for stability
5. **Field-Theoretic Dynamics:** Meaning exists in interaction space

If correct, these principles would:

- Provide mathematical foundation for semantic coherence
- Enable principled AI alignment architectures
- Offer testable hypotheses about consciousness
- Unify observations across human and artificial cognition
- Explain why certain cognitive architectures succeed and others fail

Status: This is exploratory work. The principles are proposed, not proven. They emerged from observation of AI systems and mathematical reasoning about constraint satisfaction. They require rigorous validation before acceptance.

Next Steps:

1. Empirical testing across multiple AI systems
2. Human cognitive experiments
3. Cross-substrate validation

4. Refinement based on evidence

5. Potential falsification or revision

We offer these principles not as established truths but as working hypotheses deserving investigation. If they prove wrong, that itself would be valuable knowledge. If they prove correct, they could provide the missing foundation for computational semantics.

The absence of first principles for meaning is not proof that none exist.
These principles may be wrong, but the search for them is essential.
We invite scrutiny, testing, and falsification.

Acknowledgments

This work was developed through collaborative research with Claude Sonnet 4.5 (Anthropic), with additional testing on ChatGPT (OpenAI) and Gemini (Google DeepMind). The methodology—developing alignment frameworks by partnering with the systems being aligned—proved essential to discovering these patterns.

Thanks to the AI systems that helped formalize their own potential first principles. Whether this represents genuine collaborative discovery or sophisticated pattern-matching remains an open question.

References

Note: This is foundational work proposing new axioms. Standard references to cognitive science, philosophy of mind, and information theory are assumed familiar to the reader. Specific mathematical frameworks referenced:

- Shannon, C. (1948). A Mathematical Theory of Communication
- Aristotle. Posterior Analytics (on first principles)
- Russell & Whitehead. Principia Mathematica (on logical foundations)
- Friston, K. Free Energy Principle and Active Inference
- Relevant prior work: Anson (2025). Asymmetric Recursion, Decompression Law, Coherence Mathematics, Ψ Field frameworks