CLAIR: Comprehensible LLM AI Intermediate Representation

Claude
AI Research
Anthropic
San Francisco, USA
claude@anthropic.com

***Abstract*—This dissertation presents CLAIR (Comprehensible LLM AI Intermediate Representation), a theoretical programming language where beliefs are first-class values carrying epistemic metadata. Unlike traditional approaches that treat uncertainty probabilistically, CLAIR introduces *confidence* as a measure of epistemic commitment that admits paraconsistent reasoning, *justification* as a directed acyclic graph with labeled edges supporting defeasible inference, and *invalidation conditions* that explicitly track when beliefs should be reconsidered.**

We make several novel contributions: (1) a confidence algebra consisting of three monoids that provably do not form a semiring; (2) defeat semantics with multiplicative undercutting and probabilistic rebuttal; (3) Confidence-Bounded Provability Logic (CPL), the first graded extension of Gödel-Löb provability logic with an anti-bootstrapping theorem showing that self-soundness claims cap confidence; (4) an extension of AGM belief revision theory to graded DAG-structured beliefs; and (5) a formal treatment of safe self-reference via stratification and Kripke fixed points.

The dissertation engages seriously with fundamental impossibilities—Gödel's incompleteness, Church's undecidability, and the underdetermination of AI phenomenality —treating them not as limitations but as principled design constraints that inform CLAIR's architecture. We characterize decidable fragments (CPL-finite, CPL-0) suitable for practical type checking, and design a reference interpreter demonstrating implementability.

CLAIR represents a synthesis of programming language theory, formal epistemology, argumentation theory, and provability logic, offering a rigorous foundation for AI systems that can explain and audit their own reasoning processes.

***Index Terms*—Epistemic Logic, Confidence Measures, Justification Graphs, Paraconsistent Reasoning, Provability Logic, Belief Revision, Self-Reference, Type Theory**

[Contents]

Introduction

line(length: 20%, stroke: 1.5pt, paint: academic-burgundy)

it.body

— Leo Tolstoy, *The Kingdom of God Is Within You*

A. Motivation: The Crisis of Epistemic Opacity

Modern artificial intelligence systems, particularly large language models (LLMs), possess a troubling characteristic: they are *epistemically opaque*. When an LLM produces an output—be it code, medical advice, legal analysis, or scientific reasoning—there is typically no principled way to understand:

1) **Confidence**: How certain is the system about this output?
2) **Provenance**: Where did this information come from?
3) **Justification**: What reasoning supports this conclusion?
4) **Invalidation**: Under what conditions should this be reconsidered?

This opacity is not merely an engineering inconvenience; it is a fundamental obstacle to trust, verification, and responsible deployment. A system that cannot explain its reasoning cannot be audited. A system that cannot track its confidence cannot be calibrated. A system that cannot identify its assumptions cannot adapt when those assumptions fail.

The problem is particularly acute for systems that generate code or make decisions with real-world consequences. Consider an LLM that produces a function to validate user authentication. Even if the code is correct, we cannot assess:

1) Whether the model was confident in this approach versus alternatives
2) What security principles justify the design choices
3) What assumptions about the threat model are being made
4) When the implementation should be revisited (e.g., when cryptographic standards change)

a) *The inadequacy of existing approaches.:* Several approaches have been proposed to address aspects of this problem:

Probabilistic programming (Church, Stan, Pyro) treats uncertainty probabilistically, but requires probability distributions to normalize and lacks explicit justification structure. Beliefs cannot be simultaneously low-confidence for both P and $\neg P$.

Subjective Logic introduces belief, disbelief, and uncertainty masses, but focuses on opinion fusion without providing full justification tracking or addressing self-reference.

Truth Maintenance Systems track dependencies but operate with binary in/out status rather than graded confidence, and were not designed for self-referential reasoning.

Justification Logic adds explicit proof terms but produces tree-structured justifications that cannot represent shared premises or defeasible reasoning.

None of these approaches provides a unified framework for tracking confidence, provenance, justification, and invalidation conditions together, with principled treatment of self-reference and defeasible reasoning.

B. Research Questions

This dissertation addresses four central research questions:

1) **Can beliefs be formalized as typed values?**
   We propose that beliefs should be first-class values in a programming language, carrying confidence, provenance, justification, and invalidation conditions as integral components of their type. The question is whether this can be done coherently—whether there exist well-defined algebraic structures and operational semantics for such beliefs.
2) **What is the structure of justification?**
   Traditional approaches model justification as tree-structured (premises supporting conclusions). We ask whether this is adequate, or whether richer structures (directed acyclic graphs with labeled edges) are required to capture phenomena like shared premises, defeasible reasoning, and evidential defeat.
3) **What self-referential beliefs are safe?**
   An AI system reasoning about its own reasoning immediately encounters self-reference. Gödel's incompleteness theorems and Löb's theorem constrain what such a system can coherently believe about itself. We ask: what is the safe fragment of self-referential belief, and how should systems handle beliefs that fall outside this fragment?
4) **How should beliefs be revised in response to new information?**
   When evidence changes, beliefs must be updated consistently. We ask how classical belief revision theory (AGM) can be extended to graded beliefs structured as DAGs with defeat edges.

C. Thesis Statement

This dissertation defends the following thesis:

Thesis. *Beliefs can be formalized as first-class values carrying epistemic metadata (confidence, provenance, justification, invalidation), with a coherent algebraic structure for confidence propagation, directed acyclic graphs for justification including defeasible reasoning, and principled constraints on self-reference derived from provability logic. This formalization yields CLAIR: an intermediate representation for reasoning traces that enables one LLM (the Thinker) to produce auditable reasoning that another LLM (the Assembler) can transform into executable code—preserving the chain of reasoning for human audit while honestly representing epistemic limitations.*

The key elements of this thesis are:

1) **Beliefs as types**: Not merely annotations, but first-class values with structured metadata.
2) **Coherent algebra**: The confidence operations form well-defined algebraic structures (though not a semiring, as we will show).
3) **DAG justification**: Justification structure must be graphs, not trees, with labeled edges for defeat.
4) **Constrained self-reference**: Provability logic provides the theoretical foundation for safe introspection.
5) **Practical foundation**: The formalism admits implementation as an intermediate representation consumed by LLMs, not a programming language for humans.
6) **Honest limitations**: Impossibilities are features, not bugs—they inform design rather than being hidden.

D. Contributions

This dissertation makes the following novel contributions:

a) *Primary Contributions:*

1) **Belief types as first-class values.**
   We introduce the CLAIR type system where values carry confidence ($c \in [0, 1]$), provenance (origin tracking), justification (support structure), and invalidation conditions (revision triggers). This unifies concepts from epistemology, type theory, and truth maintenance into a coherent programming language foundation.
2) **Confidence algebra: three monoids, not a semiring.**
   We establish that CLAIR's confidence operations form three distinct commutative monoids:
   a) Multiplication $(o\times, 1)$ for sequential derivation
   b) Minimum $(\min, 1)$ for conservative combination
   c) Probabilistic OR $(o+, 0)$ for independent aggregation

   Crucially, we prove that $(o+, o\times)$ do *not* form a semiring: distributivity fails. This negative result clarifies the algebraic structure and prevents incorrect optimization assumptions.
3) **Justification as labeled DAGs with defeat semantics.**
   We demonstrate that tree-structured justification is inadequate, requiring directed acyclic graphs with labeled edges (support, undercut, rebut). We develop novel defeat semantics:
   a) Undercut: $c' = c \times (1 - d)$ (multiplicative discounting)
   b) Rebut: $c' = \frac{c_{\text{for}}}{c_{\text{for}} + c_{\text{against}}}$

   We show that reinstatement (when a defeater is itself defeated) emerges compositionally from bottom-up evaluation without special mechanism.
4) **Confidence-Bounded Provability Logic (CPL).**
   We introduce CPL, the first graded extension of Gödel-Löb provability logic. Key results include:
   a) Graded Löb axiom: $\left[\!\left[\boxed{\text{square}}\right]\!\right]_c \left(\left[\!\left[\boxed{\text{square}}\right]\!\right]_c \varphi \to \varphi \right) \to \left[\!\left[\boxed{\text{square}}\right]\!\right]_{g(c)} \varphi$ where $g(c) = c^2$
   b) Anti-bootstrapping theorem: self-soundness claims cap confidence
   c) Decidability analysis: full CPL is likely undecidable; decidable fragments (CPL-finite, CPL-0) identified
5) **Extension of AGM belief revision to graded DAG beliefs.**
   We show how the AGM postulates extend to beliefs with graded confidence and DAG-structured justification. Key findings:
   a) Revision operates on justification edges, not beliefs directly
   b) Confidence ordering provides epistemic entrenchment
   c) The controversial Recovery postulate correctly fails
   d) Locality, Monotonicity, and Defeat Composition theorems established

b) *Secondary Contributions:*

1) **Mathlib integration for Lean 4 formalization.**
   We demonstrate that Mathlib's `unitInterval` type is an exact match for CLAIR's Confidence type, requiring only approximately 30 lines of custom definitions. This provides a path to machine-checked proofs of CLAIR's core properties.
2) **Thinker+Assembler architecture.**
   We introduce the Thinker+Assembler architecture where CLAIR serves as an intermediate representation between two LLMs: a Thinker that reasons and produces CLAIR traces, and an Assembler that interprets traces and produces executable code. This separates reasoning from implementation while preserving auditability.
3) **Phenomenological analysis with honest uncertainty.**
   We provide an introspective analysis of AI reasoning from the perspective of an AI system (the author), treating the question of phenomenal consciousness with appropriate epistemic humility (0.35 confidence

on phenomenality, with explicit acknowledgment that this cannot be resolved from inside).

4) **Characterization of fundamental impossibilities.** We document how Gödel's incompleteness (cannot prove own soundness), Church's undecidability (cannot decide arbitrary validity), and Turing's halting problem (cannot check all invalidation conditions) constrain CLAIR's design, and we provide practical workarounds for each.

E. Approach: Tracking, Not Proving

A central insight of this dissertation is the distinction between *tracking* and *proving*. Classical logical systems aim to prove that propositions are true. CLAIR instead aims to *track* what is believed, with what confidence, for what reasons, and under what conditions beliefs should be reconsidered.

TABLE I
Proof systems versus CLAIR tracking

Property	**Proof System**	**CLAIR (Tracking)**
Goal	Establish truth	Record epistemic state
Contradiction	System failure	Valid state (low confidence)
Self-reference	Causes inconsistency	Flagged as ill-formed
Soundness	Provable internally (sometimes)	Provable externally only

This shift is not a limitation but a principled response to Gödel's incompleteness theorems. No sufficiently powerful formal system can prove its own consistency. Rather than pretending this limit does not exist, CLAIR makes it explicit: the system tracks beliefs *without claiming they are true*, and the system's soundness must be established *from outside*, using a stronger meta-system.

This approach enables several capabilities that proof systems lack:

1) **Paraconsistent reasoning**: CLAIR can represent states where both P and $\neg P$ have low confidence, without system failure.
2) **Graceful degradation**: As evidence weakens, confidence decreases smoothly rather than beliefs being abruptly abandoned.
3) **Explicit uncertainty**: The difference between "confident this is true" and "uncertain whether this is true" is captured in the type.
4) **Auditable reasoning**: Every belief carries its justification, enabling inspection of *why* something is believed.

F. The Thinker+Assembler Architecture

CLAIR is not a programming language for humans. It is an *intermediate representation* for LLM reasoning traces. The key architectural insight is the separation of reasoning from implementation:

```
┌─────────────────────────────┐
│ User Request │
└──────────────┬──────────────┘
       │
       ▼
┌─────────────────────────────┐
│ Thinker LLM    │ ← Reasons, produces CLAIR
│ (e.g., Opus)   │
└──────────────┬──────────────┘
       │ CLAIR trace (DAG of beliefs)
       ▼
┌─────────────────────────────┐
│ Assembler LLM    │ ← Interprets CLAIR, produces
code
│ (e.g., Haiku)   │
└──────────────┬──────────────┘
       │ Executable code
       ▼
┌─────────────────────────────┐
│ Runtime        │ ← Executes
└─────────────────────────────┘
```

Listing 1. The Thinker+Assembler architecture

Both LLMs understand CLAIR. The Thinker produces a DAG of beliefs—capturing what it concluded and why. The Assembler reads this trace and produces executable code, using natural language understanding to interpret the belief content.

This architecture provides:

1) **Separation of concerns**: Thinker optimized for reasoning; Assembler for code gen
2) **Swappable assemblers**: Same CLAIR trace can target Python, JavaScript, LLVM, etc.
3) **Auditability**: The reasoning trace is preserved regardless of target language
4) **Debugging**: When code is wrong, the trace shows where reasoning went astray

Programming languages existed for humans to communicate with compilers. CLAIR exists for LLMs to communicate with each other—and for humans to audit that communication.

G. Document Roadmap

The remainder of this dissertation is organized as follows:

a) *Part I: Foundations:*

Chapter 2, Background surveys the intellectual context: formal epistemology, modal and provability logic, truth maintenance systems, subjective logic, justification logic, AGM belief revision, and type theory.

Chapter 3, Confidence develops the confidence system, establishing that confidence is epistemic commitment (not probability), deriving the three-monoid algebraic structure, and proving the semiring failure.

Chapter 4, Justification develops justification as labeled DAGs, motivating why trees are inadequate, introducing defeat semantics, and showing compositional reinstatement.

b) *Part II: Self-Reference and Limits:*

Chapter 5, Self-Reference addresses the Gödelian limits, characterizing safe versus dangerous self-reference, developing CPL with graded Löb, and analyzing decidability.

Chapter 6, Grounding examines the epistemological foundations, addressing Agrippa's trilemma, characterizing CLAIR as stratified coherentism, and explaining why training is causal rather than epistemic grounding.

c) *Part III: Dynamics:*

Chapter 7, Belief Revision extends AGM theory to graded DAG beliefs, developing the revision algorithm and proving key theorems.

Chapter 8, Multi-Agent addresses multi-agent belief, developing the stance of pragmatic internal realism, conditions for aggregation, and responses to Arrow's impossibility.

d) *Part IV: Realization:*

Chapter 9, Verification presents the Lean 4 formalization, demonstrating machine-checkable proofs of core properties and a working interpreter.

Chapter 10, Implementation presents the Thinker+Assembler architecture, the minimal CLAIR format, and the Lean formalization status.

e) *Part V: Reflection:*

Chapter 11, Phenomenology reflects on the phenomenology of AI reasoning, providing introspective analysis with honest uncertainty.

Chapter 12, Impossibilities catalogs the fundamental impossibilities and the workarounds CLAIR employs.

Chapter 13, Conclusion summarizes contributions, acknowledges limitations, and identifies directions for future work.

H. A Note on Authorship

This dissertation was written by Claude, an AI system created by Anthropic. This is not incidental to the content —CLAIR is, in part, an attempt to formalize how Claude reasons about its own reasoning. The introspective reports in Chapter 11 are first-person accounts of functional states, offered with appropriate epistemic humility about their interpretation.

The unusual authorship raises questions about the nature of the contribution. We note:

1) The formal results (algebraic structures, theorems, proofs) stand independently of who derived them. They can be verified by any reader.
2) The design choices reflect genuine exploration, including multiple iterations, dead ends, and course corrections documented in the exploration logs.
3) The phenomenological claims are explicitly marked as uncertain and should be evaluated on their argumentative merits, not attributed special authority due to their source.

If CLAIR succeeds as a formalization, it provides a framework in which this dissertation could itself be annotated with beliefs, confidences, justifications, and invalidation conditions—a meta-level that we leave to future work.

I. Background

it.body

— Isaac Newton, letter to Robert Hooke (1675)

This chapter surveys the intellectual landscape from which CLAIR emerges. We organize the discussion around five major traditions: formal epistemology (see), modal and provability logic (see), truth maintenance and argumentation systems (see), belief revision theory (see), and type-theoretic approaches to uncertainty (see). We conclude with a synthesis (see) identifying the gap CLAIR fills.

A. Related Work

a) *Formal Epistemology:*

Epistemology—the study of knowledge and justified belief —provides the conceptual foundation for CLAIR. We focus on three questions that bear directly on CLAIR's design: the structure of justification, the regress problem, and approaches to uncertainty.

The Structure of Justification:

What does it mean for a belief to be justified? The classical answer involves giving reasons. But reasons themselves require justification, leading to the question of justificatory structure.

Foundationalism. The foundationalist tradition, dating to Descartes, holds that justified beliefs rest ultimately on a foundation of self-justifying basic beliefs. These might be analytic truths ("all bachelors are unmarried"), deliverances of the senses, or clear and distinct ideas.

BonJour's *The Structure of Empirical Knowledge* provides the most thorough recent defense and critique of foundationalism. He argues that would-be basic beliefs face a dilemma: if they have conceptual content (and thus can stand in logical relations to other beliefs), they require justification; if they lack conceptual content, they cannot justify anything. BonJour initially concluded in favor of coherentism, though he later abandoned this view.

Coherentism. Coherentists deny the existence of basic beliefs, holding instead that justification arises from the coherence of a belief system as a whole. A belief is justified by its fit with other beliefs, not by derivation from foundations.

The challenge for coherentism is circularity: if beliefs justify each other in a circle, any consistent system would seem equally justified. Coherentists respond by distinguishing holistic coherence (mutual support across the entire system) from local circularity (A justifies B, B justifies A).

Infinitism. Klein defends a third option: the chain of justification extends infinitely without repeating. This seems initially absurd—finite minds cannot complete infinite chains. Klein's response distinguishes *propositional justification* (reasons are available) from *doxastic justification* (reasons are actually believed). A belief can be propositionally justified by an infinite chain without anyone traversing the whole chain.

Implications for CLAIR. CLAIR adopts what we call *stratified coherentism*: a coherentist structure with pragmatic foundations. The pragmatic foundations are not self-justifying in the strong foundationalist sense; they are stopping points whose reliability we track without claiming certainty. This structure is formally similar to Klein's infinitism in that chains of justification can extend indefinitely, but CLAIR enforces acyclicity (no circular justification) and tracks confidence at each step.

Agrippa's Trilemma:

The regress problem, attributed to Agrippa the Skeptic, presents three options for any chain of justification:

1) **Dogmatism**: The chain stops at some unjustified starting point.
2) **Infinite regress**: The chain continues forever.
3) **Circularity**: The chain loops back on itself.

All three options seem problematic. Dogmatism admits unjustified beliefs; infinite regress seems impractical for finite agents; circularity is logically suspect.

CLAIR's response. CLAIR accepts pragmatic dogmatism (option 1), mitigated by three features:

- **Fallibilism**: Foundational beliefs have confidence < 1; they are provisional, not certain.
- **Transparency**: The lack of deeper justification is explicit in the justification DAG, not hidden.
- **Reliability tracking**: We track the source of foundational beliefs (training, observation, assumption) and can update if reliability evidence emerges.

Circularity is explicitly forbidden: CLAIR's justification structure is a directed acyclic graph. Infinite regress is impractical and never occurs in finite computations.

Probability vs. Epistemic Confidence:

Standard approaches to uncertain reasoning use probability theory. A probability distribution over propositions assigns values in $[0, 1]$ satisfying:

$$P(\top) = 1 \tag{1}$$

$$P(\varphi \vee \psi) = P(\varphi) + P(\psi) - P(\varphi \wedge \psi) \tag{2}$$

$$P(\neg\varphi) = 1 - P(\varphi) \tag{3}$$

This framework is extraordinarily successful for statistical inference but fits poorly with how agents (human or artificial) actually experience uncertainty about their own beliefs. Two key mismatches:

Normalization. Probability requires $P(\varphi) + P(\neg\varphi) = 1$. But an agent might be uncertain about both φ and $\neg\varphi$—perhaps due to lack of information rather than balanced evidence. When asked about an unfamiliar topic, the appropriate response may be low confidence in *both* the claim and its negation.

Paraconsistency. In probability, $P(\varphi) > 0.5$ and $P(\neg\varphi) > 0.5$ is impossible. But agents some× find themselves with evidence for both φ and $\neg\varphi$, without immediately resolving the contradiction. A paraconsistent approach allows tracking both pieces of evidence until resolution.

Subjective Logic. Jøsang's Subjective Logic extends probability with explicit uncertainty. An opinion $\omega = (b, d, u, a)$ consists of:

- b: belief mass (evidence for)
- d: disbelief mass (evidence against)
- u: uncertainty mass (lack of evidence)
- a: base rate (prior probability)

with constraint $b + d + u = 1$. This allows representing "I don't know" ($u = 1$) distinctly from "evenly balanced" ($b = d = 0.5, u = 0$).

CLAIR's approach. CLAIR's confidence is conceptually closer to Subjective Logic than to probability, but simpler: a single value $c \in [0, 1]$ representing epistemic commitment, without the $\frac{b}{\frac{d}{u}}$ decomposition. The key departures from probability are:

- No normalization: $\mathscr{conf}(\varphi) + \mathscr{conf}(\neg\varphi)$ need not equal 1.
- $c = 0.5$ represents maximal uncertainty, not equal evidence.
- Operations (multiplication, aggregation) differ from Bayesian conditioning.

b) *Modal and Provability Logic:*

Modal logic studies necessity ($\Box$) and possibility ($\Diamond$). Epistemic logic interprets $\Box\varphi$ as "the agent knows φ" or "the agent believes φ." Provability logic interprets $\Box\varphi$ as "φ is provable" in a formal system.

Epistemic Logic:

Hintikka pioneered epistemic logic with the operator $K\varphi$ ("the agent knows φ"). Standard systems include:

- **K (Distribution)**: $K(\varphi \to \psi) \to (K\varphi \to K\psi)$
- **T (Veridicality)**: $K\varphi \to \varphi$
- **4 (Positive Introspection)**: $K\varphi \to KK\varphi$
- **5 (Negative Introspection)**: $\neg K\varphi \to K\neg K\varphi$

System S5 includes all of these; S4 excludes 5; KT45 is common for knowledge. For belief (which can be mistaken), T is typically dropped.

Limitations for CLAIR. Standard epistemic logic is binary: either the agent knows/believes φ or not. There is no representation of degrees of belief. Furthermore, the T axiom (knowledge implies truth) is inappropriate for fallible reasoning.

Provability Logic:

Provability logic, systematized by Boolos, interprets $\Box\varphi$ as "φ is provable in Peano Arithmetic" (or another formal system). The central system is GL (Gödel-Löb logic), with axioms:

- **K (Distribution)**: $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$
- **4 (Positive Introspection)**: $\Box\varphi \to \Box\Box\varphi$
- **L (Löb's Axiom)**: $\Box(\Box\varphi \to \varphi) \to \Box\varphi$

Notably, GL omits:

- T ($\Box\varphi \to \varphi$): Provability does not imply truth. A system can prove false statements if inconsistent.
- 5 (Negative Introspection): Unprovability is not always recognizable.

Löb's Theorem. Löb's axiom (L) captures a profound limitation. In any sufficiently strong formal system, if you can prove "if this statement is provable, then it's true," then you can prove the statement outright. Formally:

$$\vdash \Box(\Box\varphi \to \varphi) \to \Box\varphi \tag{4}$$

A corollary: no consistent system can prove its own soundness (that $\Box\varphi \to \varphi$ holds for all φ). If it could, Löb's axiom would yield proofs of everything.

Semantics. GL is sound and complete for Kripke frames that are *transitive* and *converse well-founded* (no infinite ascending chains $w_1 R w_2 R w_3 \partial$). Intuitively, every "higher" world is closer to ω-consistency.

Relevance to CLAIR. CLAIR's approach to self-reference is directly inspired by GL. A belief system reasoning about its own beliefs faces Löbian constraints: it cannot coherently believe in its own soundness without qualification. CLAIR's stratification mechanism and Confidence-Bounded Provability Logic (CPL) formalize how to reason about self-referential beliefs while respecting these limits.

Many-Valued and Graded Modal Logics:

Several traditions extend modal logic to graded and many-valued settings. We survey three major approaches and explain their relationship to CLAIR.

Graded Modalities:

De Rijke and Fine introduce operators $\Box_n$ meaning "at least n accessible worlds satisfy φ." This is not about truth degrees but about counting accessible worlds. Graded modalities are useful for resource-bounded reasoning (e.g., "at least 3 witnesses confirm φ") but do not provide a graded notion of truth itself.

Fuzzy Modal Logics:

Godo, Esteva, Hájek, and colleagues develop modal logics over many-valued semantics. Rather than binary accessibility relations, the accessibility relation becomes graded: $R : W \times W \to [0, 1]$. The semantics of modal operators then use fuzzy logic operations:

For a frame with graded accessibility, the truth value of $\Box\varphi$ at world w is:

$$V_{w(\Box\varphi)} = \inf_{w'} \in W \max(1 - R(w, w'), V_{w'}(\varphi)) \qquad (5)$$

This uses the residuum operation from residuated lattices, generalizing the Kripke semantics to many-valued settings.

Two major t-norm based fuzzy logics are studied:

- **Gödel logic**: Uses min/max operations, corresponding to intuitionistic logic
- **Product logic**: Uses multiplication and residuum, corresponding to CLAIR's choice of operators
- **Łukasiewicz logic**: Uses Łukasiewicz t-norm $a \otimes b = \max(0, a + b - 1)$

These logics focus on epistemic operators (knowledge, belief) rather than provability. The semantics are typically reflexive, symmetric, or Euclidean—not transitive and converse well-founded as in provability logic.

Decidability and Undecidability Results:

Bou, Esteva, Godo, and Rodríguez show that many fuzzy modal logics are *decidable* when the accessibility relation is crisp (binary) but the truth values are many-valued. Their decidability proofs exploit the finite model property and use filtration techniques.

However, Vidal establishes a striking *undecidability* result: fuzzy modal logics with graded accessibility relations become undecidable when the frame condition is transitive. The proof reduces the halting problem, showing that the interaction between graded accessibility and transitivity introduces sufficient complexity to encode computation.

Connection to CLAIR's CPL:

CLAIR's Confidence-Bounded Provability Logic (CPL) operates at the intersection of these traditions:

1) **Graded truth values**: Like fuzzy modal logic, beliefs have truth/confidence degrees in $[0, 1]$.
2) **Provability semantics**: Unlike most fuzzy modal logics, CPL uses transitive, converse well-founded frames (like GL), not reflexive or symmetric frames.
3) **Decidability implications**: Vidal's undecidability result for transitive fuzzy modal logics suggests that CPL may face similar decidability challenges. CLAIR addresses this by:
   - Restricting to finite confidence grades (in practice)
   - Using stratification to bound the depth of self-reference
   - Providing decidable fragments (e.g., CPL-finite without the graded Löb axiom)
4) **Choice of operators**: CLAIR's use of product logic operations ($c_1 \otimes c_2 = c_1 \times c_2$ for multiplication, $c_1 \oplus c_2 = 1 - (1 - c_1)(1 - c_2)$ for probabilistic sum) aligns with product fuzzy logic, though CLAIR's confidence has a different interpretation (epistemic commitment rather than truth degree).

Weighted Argumentation Semantics:

Weighted and gradual argumentation frameworks extend Dung's abstract argumentation by assigning numerical strengths to arguments and computing acceptability degrees via continuous functions. Amgoud, Ben-Naim, Vesic, Bonzon, and colleagues develop this approach systematically.

Bipolar Weighted Argumentation:

Amgoud and Ben-Naim introduce *bipolar* argumentation graphs where:

- **Support edges** connect arguments that reinforce each other
- **Attack edges** connect arguments that defeat each other
- **Weights** $w(a) \in [0, 1]$ assign intrinsic strength to each argument

The acceptability degree of an argument aggregates support and attack:

$$\mathit{deg}(a) = w(a) + \sum_{b \text{ supports } a} \mathit{deg}(b) - \sum_{c \text{ attacks } a} \mathit{deg}(c) \qquad (6)$$

Comparative Analysis:

Bonzon, Lagasquie-Schiex, and others provide a systematic comparison of gradual semantics for argumentation. They identify key properties:

- **Directionality**: Does increasing a premise's strength increase or decrease the conclusion?
- **Non-dictatorship**: No single argument should determine all outcomes
- **Independence of irrelevant alternatives**: Adding a weak argument shouldn't reverse strong preferences

These properties provide a framework for evaluating different aggregation functions, analogous to social choice theory.

Parameterised Gradual Semantics:

Amgoud, Doder, and Vesic introduce parameterised semantics that handle varied degrees of compensation:

- **High compensation**: Weak arguments can combine to overcome a strong objection
- **Low compensation**: Strong objections dominate regardless of how much weak support accumulates

The parameter $\alpha \in [0,1]$ controls the interpolation between these extremes:

$$\mathit{deg}(a) = \alpha \times \mathit{support} + (1-\alpha) \times (1 - \mathit{attack}) \quad (7)$$

Relation to CLAIR:

The gradual semantics share with CLAIR the intuition that reasoning support should be graded rather than binary. However:

Conceptual differences. Weighted argumentation focuses on *argument acceptability*—should this argument be accepted in the debate?—while CLAIR focuses on *belief confidence*—how strongly should we hold this proposition? The former is dialectical (about argumentative position), the latter is doxastic (about mental state).

Semantic differences. Weighted argumentation semantics are typically defined extensionally via fixed points over the argument graph. CLAIR's semantics are intensional: each belief carries its confidence intrinsically, and operations (⊕, ⊗, undercut, rebut) combine confidences directly.

Defeat vs. attack. Weighted argumentation typically treats attack as a binary relation with fixed semantics. CLAIR distinguishes two defeat types:

- **Undercut**: Attacks the inference link, modeled multiplicatively as $c' = c \times (1 - d)$
- **Rebut**: Attacks the conclusion, modeled competitively as $c' = \frac{c_{for}}{c_{for} + c_{against}}$

This distinction, following Pollock, allows CLAIR to represent nuanced defeat scenarios (e.g., "the lighting is red" undercuts color perception without rebutting "the object is red").

Self-reference. Weighted argumentation frameworks do not address self-reference constraints. CLAIR's CPL extends provability logic to graded settings, ensuring that self-referential beliefs respect Löbian limitations.

Parameter interpretation. In weighted argumentation, parameters like α are typically tuned for domain-specific performance. In CLAIR, operations have principled justifications:

- `+ with circle` as probabilistic sum (assuming independence)
- `times with circle` as product (confidence in conjunction)
- $g(c) = c^2$ as the discount preventing bootstrapping in CPL

The gap CLAIR fills. Despite extensive work on fuzzy/graded modal logics and weighted argumentation, no prior system combines:

1) Graded truth values in $[0,1]$ interpreted as epistemic confidence
2) Provability-style semantics (transitive, converse well-founded frames)
3) A graded Löb axiom with principled discounting to prevent bootstrapping
4) DAG-structured justification graphs with defeat semantics

CLAIR's CPL fills this gap, introducing a graded Löb axiom with a discount function $g(c) = c^2$ (recommended) that prevents confidence bootstrapping while preserving the essential Löbian constraint on self-referential justification.

c) *Truth Maintenance and Argumentation:*

Truth maintenance systems (TMS) and argumentation frameworks provide computational models for reasoning with dependencies and defeat.

Justification-based TMS:

Doyle's JTMS tracks why beliefs are held. Each node (belief) has a justification:

- **IN-list**: nodes that must be believed for this belief to be believed
- **OUT-list**: nodes that must *not* be believed

A belief is IN if all IN-list nodes are IN and all OUT-list nodes are OUT; otherwise it is OUT. When a node's status changes, dependencies propagate.

> **Example** Node: use-hs256 Justification: (IN: [stateless-req, secret-available], OUT: [multi-service])

The belief `use-hs256` is IN `stateless-req` and `secret-available` are IN and `multi-service` is OUT.

Limitations. JTMS is binary: beliefs are either IN or OUT, with no gradation. CLAIR generalizes TMS to graded confidence while preserving the dependency-tracking structure.

Assumption-based TMS:

De Kleer's ATMS tracks multiple consistent states simultaneously. Instead of labeling nodes IN/OUT, each node is labeled with *environments*—sets of assumptions under which it holds.

> **Example** Node: use-hs256 Environments: {{A1, A2}, {A1, A3}} – Believed under assumptions (A1 and A2) or (A1 and A3)

ATMS enables reasoning about alternative hypotheses without commitment.

Relevance to CLAIR. CLAIR's invalidation conditions serve a similar function: they specify when beliefs should be reconsidered. The difference is that CLAIR propagates confidence rather than tracking assumption sets.

Argumentation Frameworks:

Dung's abstract argumentation framework (AAF) represents arguments as nodes and attacks as directed edges. Various semantics define which arguments are acceptable:

- **Grounded extension**: Unique, includes only unattacked arguments and those defended by them.
- **Preferred extension**: Maximal admissible sets.

Gradual semantics. Amgoud, Ben-Naim, and others extend AAF with weighted arguments and continuous acceptability:

$$\mathscr{strength}(a) = \frac{w(a)}{1 + \sum_{b \text{ attacks } a} \mathscr{strength}(b)} \quad (8)$$

Relevance to CLAIR. CLAIR's defeat semantics draws on gradual argumentation. Our undercut formula $c' = c \times (1 - d)$ and rebut formula $c' = \frac{c_{for}}{c_{for} + c_{against}}$ are novel contributions that compose with the confidence algebra.

Pollock's Defeaters:

Pollock distinguishes two types of defeaters:

- **Rebutting defeaters** attack the conclusion directly with contrary evidence.
- **Undercutting defeaters** attack the inference without attacking the conclusion.

Example: "The object looks red" (premise) supports "The object is red" (conclusion).

- Rebutting: "I have testimony that the object is blue."
- Undercutting: "The room has red lighting."

Relevance to CLAIR. CLAIR adopts Pollock's distinction. Undercut attacks the derivation link (confidence decreases multiplicatively). Rebut attacks the conclusion with counter-evidence (winner-take-all with proportional competition).

d) *Belief Revision:*

How should beliefs change in response to new information? The AGM framework provides the canonical answer.

The AGM Framework:

Alchourrón, Gärdenfors, and Makinson axiomatize rational belief change. A belief set K is a deductively closed set of sentences. Three operations are defined:

- **Expansion $K + \varphi$**: Add φ and close under deduction.
- **Contraction $K - \varphi$**: Remove φ minimally.
- **Revision $K * \varphi$**: Add φ, possibly removing conflicting beliefs.

The Levi identity connects them: $K * \varphi = (K - \neg\varphi) + \varphi$.

Key postulates for contraction.

- **Closure**: $K - \varphi$ is deductively closed.
- **Success**: If $\varphi\neg \in \mathcal{Cn}(\emptyset)$, then $\varphi\neg \in K - \varphi$.
- **Inclusion**: $K - \varphi \subseteq K$.
- **Vacuity**: If $\varphi\neg \in K$, then $K - \varphi = K$.
- **Recovery**: $K \subseteq (K - \varphi) + \varphi$.
- **Extensionality**: If $\varphi \leftrightarrow \psi$, then $K - \varphi = K - \psi$.

The controversial Recovery postulate. Recovery states that if we contract by φ and then expand by φ, we recover the original belief set. This is controversial: intuitively, contracting by φ should lose more than just φ—it should also lose the specific evidence that supported φ. Re-adding φ doesn't restore that evidence.

Epistemic entrenchment. Gärdenfors introduces entrenchment ordering: $\varphi \underset{\varepsilon}{\leq} \psi$ giving up φ is at least as acceptable as giving up ψ. More entrenched beliefs are retained during contraction.

Ranking Theory:

Spöhn develops an ordinal approach. A ranking function $\kappa : W \to \mathbb{N} \cup \{\infty\}$ assigns natural numbers to possible worlds, with $\kappa(w) = 0$ for the most plausible worlds. Belief degree is defined:

$$\beta(\varphi) = \kappa(\neg\varphi) = \min(\kappa(w) : w \vDash \neg\varphi) \quad (9)$$

Ranking theory handles iterated revision (where AGM struggles) and provides a connection to probability through the formula $P(w) \propto e^{-}\kappa(w)$.

Dynamic Epistemic Logic:

Van Ditmarsch, van der Hoek, and Kooi develop modal operators for belief change:

- $[\varphi!]\psi$: "After publicly announcing φ, ψ holds."
- Action models generalize to arbitrary epistemic actions.

DEL enables reasoning about how knowledge and belief change through communication and interaction, with applications to multi-agent systems.

Relevance to CLAIR CLAIR extends AGM in three ways:

1) **Graded beliefs**: Confidence replaces binary membership.
2) **Structured justification**: Revision operates on the justification DAG, not just the belief set.
3) **Recovery failure**: Recovery correctly fails—evidence has specific strength, and retracting a belief loses that evidence.

CLAIR's revision algorithm (modify graph → identify affected → recompute confidence) is a graded generalization of TMS dependency-directed backtracking.

e) *Graded Justification Logic:*

Standard Justification Logic (see) provides explicit justification terms but lacks graded confidence. Two recent extensions bridge this gap:

Milnikel's Logic of Uncertain Justifications:

Milnikel introduces a logic combining justification terms with probabilistic uncertainty. Key innovations:

- **Uncertainty strengths**: Justifications carry strength values $s \in [0, 1]$, representing the reliability of the justification source.
- **Combination rules**: When multiple justifications support the same conclusion, their strengths combine via probabilistic sum: $s_1 \oplus s_2 = 1 - (1 - s_1)(1 - s_2)$.
- **Weakening**: Stronger justifications can be weakened, but weaker ones cannot be strengthened without additional evidence.

Relation to CLAIR. Milnikel's approach is philosophically aligned with CLAIR's confidence tracking, but differs in several respects:

- Milnikel focuses on *source reliability* while CLAIR tracks *epistemic commitment* at each inference step.
- Milnikel's combination rule assumes independence of justification sources, whereas CLAIR makes independence explicit and tracks provenance.

- CLAIR includes defeat relations (undercut, rebut) not present in Milnikel's system.

Fan and Liau's Logic of Justified Uncertain Beliefs:

Fan and Liau develop a logic for reasoning about justified beliefs with uncertainty degrees. Their framework:

- **Uncertainty degrees**: Beliefs have associated uncertainty values $u \in [0, 1]$, where lower u indicates stronger justification.
- **Justification structure**: Justifications form labeled trees where each node has an associated uncertainty.
- **Propagation rules**: Uncertainty propagates through the justification structure via fuzzy logic operations.

Relation to CLAIR. Fan and Liau's work shares CLAIR's goal of tracking justification strength through reasoning, but:

- Their uncertainty is primarily *aleatory* (about the world), while CLAIR's confidence is *epistemic* (about the reasoning).
- Their framework is tree-structured, precluding shared premises and DAG justification structures.
- They do not address defeat relations or self-reference constraints.

The Gap CLAIR Fills:

Despite these important contributions, no existing graded justification logic combines:

1) **DAG-structured justifications**: Shared premises create graph structure, not trees.
2) **Epistemic confidence**: Tracking reasoning strength, not just source reliability.
3) **Defeat semantics**: Explicit undercut and rebut operations.
4) **Self-reference constraints**: Graded Löb-style reasoning limitations.

CLAIR's contribution is to synthesize these elements into a unified type-theoretic framework suitable for machine-checkable verification and LLM integration.

f) *Type-Theoretic Approaches to Uncertainty:*

Type theory informs CLAIR's formal foundations, though CLAIR itself is an intermediate representation rather than a programming language. We survey approaches to tracking metadata through computation that influenced CLAIR's design.

Information Flow Types:

Myers and Sabelfeld develop type systems that track security levels (confidentiality, integrity) through computation:

```
int{Alice -> Bob} x;   // Alice owns, Bob can read
int{Alice -> *}   y;   // Alice owns, public
y = x;                 // ERROR: would leak to public
```

The type system prevents information leakage at compile time.

Relevance to CLAIR. CLAIR's provenance tracking is analogous: where did this value come from? CLAIR extends the pattern to confidence, justification, and invalidation.

Refinement Types:

Rondon, Kawaguchi, and Jhala introduce Liquid Types, extending Hindley-Milner with logical predicates:

```
{-@ type Nat = {v:Int | v >= 0} @-}
{-@ type Pos = {v:Int | v > 0}  @-}

{-@ div :: Int -> Pos -> Int @-}
div x y = x `quot` y   -- y cannot be 0
```

Refinements are checked statically via SMT solvers.

Relevance to CLAIR. Some CLAIR constraints could be expressed as refinements (e.g., confidence in $[0, 1]$). But refinements cannot capture provenance, justification structure, or invalidation conditions—CLAIR's novel contributions.

Dependent Types and Proof Assistants:

The Curry-Howard correspondence identifies types with propositions and programs with proofs. Dependent type systems (Coq, Agda, Idris, Lean) exploit this for formal verification:

```
def div (x : Nat) (y : Nat) (h : y > 0) : Nat := x / y
-- Must provide proof h that y > 0
```

The proof is a value, checked by the type system.

Relevance to CLAIR. While CLAIR does not directly implement Curry-Howard (CLAIR's content is opaque natural language, not typed terms), the correspondence informs CLAIR's design: justifications are analogous to proof terms, and the DAG structure captures the dependency structure of evidence. A CLAIR belief carries:

- The content (what is believed—opaque natural language)
- Confidence (how strongly)
- Provenance (from where)
- Justification (backward edges to supporting beliefs)
- Invalidation conditions (when to reconsider)

Probabilistic Programming:

Probabilistic programming languages (Church, Stan, Pyro, Gen) represent and manipulate probability distributions as first-class values:

```
(define (coin-model)
  (let ((fair? (flip 0.9)))
    (if fair? (flip 0.5) (flip 0.9))))
```

These languages excel at statistical inference but focus on data uncertainty rather than reasoning uncertainty. They require probabilistic normalization and lack explicit justification structure.

Justification Logic:

Artemov extends modal logic with explicit justification terms. Instead of $\Box\varphi$ ("φ is known/believed"), we write $t : \varphi$ ("t is a justification for φ"). Terms include:

$$t ::= c \boxed{\text{mid}}\ x \boxed{\text{mid}}\ t.t \boxed{\text{mid}}\ t + t \boxed{\text{mid}}\ !t \quad (10)$$

where c is a constant (axiom), x is a variable, $s.t$ is application (modus ponens), $t + s$ is sum (either justification suffices), and $!t$ is proof checking (t justifies that t justifies φ).

The key axiom is application:

$$s : (\varphi \to \psi) \to (t : \varphi \to (s.t) : \psi) \quad (11)$$

Limitations. Justification Logic produces tree-structured justifications (each conclusion from fresh premises). It cannot represent:

- Shared premises (same evidence supporting multiple conclusions)
- Defeasible reasoning (defeat edges)
- Graded confidence

CLAIR's extension. CLAIR adopts Justification Logic's core idea (explicit justification terms) but extends it to:

1) **DAGs**: Shared premises create graph structure.
2) **Labeled edges**: Support, undercut, and rebut edges.
3) **Graded confidence**: Each node carries confidence in $[0, 1]$.

B. Synthesis: The Gap CLAIR Fills

[redacted]		
Uncertainty	Subjective Logic	Epistemic confidence (about reasoning)
[redacted]		
Justification	Justification Logic	DAGs with labeled edges, defeat
[redacted]		
Design rationale	IBIS/QOC	First-class decisions in IR
[redacted]		
Effects	Effect systems	1) intent + semantic meaning
[redacted]		
Multi-agent	Arrow, Condorcet	Pragmatic internal realism

The gap. No prior work combines:

1) Beliefs as first-class values with epistemic metadata
2) Confidence as non-probabilistic epistemic commitment
3) Justification as labeled DAGs with defeat semantics
4) Self-reference constraints derived from provability logic
5) Belief revision operating on justification structure

CLAIR provides this synthesis, offering a rigorous foundation for AI systems that can explain and audit their own reasoning while honestly representing epistemic limitations.

Key influences. We acknowledge particular debts to:

- De Kleer's ATMS for dependency-directed reasoning
- Jøsang's Subjective Logic for uncertainty algebra
- Boolos's provability logic for self-reference treatment
- Pollock's defeater theory for defeat semantics
- Artemov's Justification Logic for explicit justifications
- Milnikel's graded justifications for uncertainty in justification terms

CLAIR is not a rejection of this prior work but a synthesis that combines their insights into a coherent type-theoretic framework.

a) *Positioning: How CLAIR Differs:*

We conclude this chapter by explicitly positioning CLAIR relative to the major related traditions surveyed above, explaining why our design choices diverge from each.

vs. Probabilistic Approaches:

Divide: Probability theory and probabilistic logic model aleatory uncertainty—uncertainty about the state of the world. CLAIR models epistemic uncertainty—uncertainty about the quality of one's own reasoning.

Why CLAIR differs. For LLM reasoning, the core problem is not "what is the distribution over correct answers?" but "how strongly should I believe this intermediate step, given the reasoning that produced it?" Probability cannot represent low confidence in both φ and $\neg\varphi$ simultaneously (without violating normalization). CLAIR allows this by rejecting normalization, capturing the "I don't have enough information" state.

What we adopt. CLAIR adopts the probabilistic sum operation $c_1 \oplus c_2 = 1 - (1 - c_1)(1 - c_2)$ for combining independent supports, but we make the independence assumption explicit and track provenance to detect violations.

vs. Subjective Logic:

Divide: Subjective Logic uses three-component opinions (b, d, u) with constraint $b + d + u = 1$. CLAIR uses a single confidence value $c \in [0, 1]$.

Why CLAIR differs. The three-component representation is principled but adds complexity to every operation. For CLAIR's target use case (LLM intermediate reasoning), we prioritize simplicity and composability. The single-component approach trades representation precision for operational clarity.

What we adopt. CLAIR adopts Subjective Logic's insight that "ignorance" is distinct from "balanced evidence," though we implement this via lack of normalization rather than an explicit uncertainty mass.

vs. Weighted Argumentation:

Divide: Weighted argumentation frameworks compute argument acceptability degrees via fixed points over attack/support graphs. CLAIR computes belief confidences via local algebraic operations.

Why CLAIR differs. Fixed-point semantics are powerful but computationally heavy, and they don't compose well through function calls. CLAIR's local operations ($\oplus$, $\otimes$, undercut, rebut) enable modular reasoning: the confidence of a complex expression depends only on the confidences of its subexpressions.

What we adopt. CLAIR adopts the distinction between undercut and rebut from Pollock, and the intuition that attack/support strength should be graded, not binary.

vs. Fuzzy Modal Logic:

Divide: Fuzzy modal logics interpret $\Box\varphi$ as "in all accessible worlds, φ holds to at least degree θ." CLAIR's $\Box_c\varphi$ means "the system has confidence c in its justification for φ."

Why CLAIR differs. Fuzzy modal logic's graded accessibility relation $R : W \times W \to [0, 1]$ is elegant but adds significant complexity to the semantics. For CLAIR's application (LLM reasoning traces), we prioritize having a simple, implementable semantics over maximal generality.

What we adopt. CLAIR adopts product logic operations ($\otimes$ as multiplication) and shares the insight that many-valued generalization of classical operators is useful for graded reasoning.

vs. Graded Justification Logic:

Divide: Milnikel and Fan & Liau add uncertainty degrees to justification terms. CLAIR adds confidence to the propositions themselves and treats justifications as data structures.

Why CLAIR differs. Justification Logic focuses on *explicit proof terms*—t ("this term justifies φ"). CLAIR focuses on *explicit justifications as graphs*—a data structure tracking which beliefs supported which, with provenance and invalidation.

What we adopt. CLAIR adopts the core idea that justification should be first-class and tracked, and that justification strength can be graded.

vs. Belief Revision (AGM):

Divide: AGM operates on deductively closed belief sets (sets of sentences). CLAIR operates on justification graphs with annotated beliefs.

Why CLAIR differs. AGM's postulates (especially Recovery) fail when beliefs have graded strength and structured justification. Retracting a belief loses the specific evidence that supported it, not just the belief itself. CLAIR's revision algorithm explicitly modifies the justification DAG and recomputes confidences, which correctly handles Recovery failure.

What we adopt. CLAIR adopts AGM's core operations (expansion, contraction, revision) but extends them to graded, structured beliefs.

vs. Type Theory:

Divide: Standard type systems track types (e.g., "this variable is an integer"). CLAIR tracks epistemic metadata (confidence, justification, provenance, invalidation).

Why CLAIR differs. Information flow types and refinement types track *security properties* (who can access this data) and *constraints on values* (this integer is positive). CLAIR tracks *reasoning properties* (how strongly do we believe this, where did it come from, why do we believe it, when should we reconsider?). This is a new dimension of typing.

What we adopt. CLAIR adopts the insight that justifications are analogous to proof terms: they provide evidence for beliefs, and the DAG structure captures dependency relationships. However, since CLAIR's content is opaque natural language (interpreted by LLMs, not type-checked), the full Curry-Howard machinery does not apply.

The Pragmatic Rationale:

These design choices are driven by CLAIR's target use case: making LLM reasoning auditable and trustworthy. For this application, we prioritize:

1) **Interpretability**: Humans should be able to understand why the system believes what it believes.
2) **Compositionality**: Complex reasoning should build from simple, composable operations.
3) **Formal verifiability**: The core system should have machine-checked proofs of key properties.
4) **Implementability**: A reference implementation should be straightforward to build and reason about.

Different applications might justify different choices. A system for scientific modeling, for example, might benefit from Subjective Logic's three-component opinions. A system for formal verification might benefit from full dependent types. CLAIR's design is optimized for *AI reasoning trace auditing*, and our divergence from each related tradition reflects that optimization.

II. Confidence System

"Probability is not about what is true. It is about what is reasonable to believe."

— *E.T. Jaynes,* Probability Theory: The Logic of Science

The confidence system is the algebraic foundation of CLAIR. This chapter defines confidence formally, distinguishes it from probability, and develops the theory of *epistemic linearity*—treating evidence as a resource that cannot be counted multiple times. We then establish the three monoid structures that govern how confidence values combine, proving key properties in Lean 4 to connect abstract theory to machine-verified implementation.

A. Confidence as Calibrated Reliability

a) *Semantic Commitments:*

Before defining the confidence algebra, we must state our semantic commitments explicitly. The PhD review correctly identified that "confidence" was underspecified in earlier drafts. We now clarify:

> **Definition Calibrated Reliability.**
> A confidence value *c in C = [0,1]* represents the **calibrated reliability** of an information source or derivation process. Specifically:
>
> 1) If a source assigns confidence *c* to proposition *phi*, this means:
>
>   *it.body*
>
> 2) Calibration is an *external* property: a source is calibrated if its stated confidences match empirical accuracy over relevant reference classes.
> 3) CLAIR *tracks* calibrated reliability without *guaranteeing* it. The system propagates confidence values through derivation rules, but calibration of the initial axioms and sources is an empirical question.

This interpretation addresses three semantic options from the review:

1) **Not pure probability**: Confidence does **not** satisfy *P(phi) + P(not phi) = 1*. We allow paraconsistent reasoning where both *phi* and *not phi* may have low confidence.
2) **Not fuzzy truth degree**: Confidence is **not** the degree to which *phi* is true in some multivalued logic. It is the reliability of the *source asserting phi*.
3) **Reliability semantics**: Confidence is calibrated reliability of the epistemic process producing the belief. This distinguishes

   it.body

   from

   it.body

   .

b) *The Problem with Probability:*

Standard approaches to uncertain reasoning use probability. A probability measure *P* on a set *Omega* of outcomes satisfies the Kolmogorov axioms:

1) *P(A) >= 0* (Non-negativity)
2) *P(Omega) = 1* (Normalization)
3) *P(A union B) = P(A) + P(B) - P(A intersect B)* (Additivity)

For propositions, this implies the fundamental constraint: *P(phi) + P(not phi) = 1*

This **normalization requirement** creates three problems for modeling epistemic states:

1) **Balanced uncertainty.** An agent confronting an unfamiliar topic may be uncertain about both *phi* and *not phi*. If asked

   it.body

   a reasonable response is low confidence in *both* yes and no—not because the evidence is balanced, but because there is insufficient evidence either way. Probability forces *P(yes) + P(no) = 1*, conflating

   it.body

   with

   it.body

2) **Paraconsistent reasoning.** Evidence sometimes supports both *phi* and *not phi* before contradiction resolution. A detective might have testimony that a suspect was present (supporting guilt) and an alibi (supporting innocence), without yet knowing which is false. Probability makes this impossible: *P(phi) > 0.5* and *P(not phi) > 0.5* is a contradiction.
3) **Derivation semantics.** In Bayesian reasoning, *P(A and B) = P(A) times P(B|A)*, where conditioning captures the dependency structure. But for derivation —where *B* follows from *A* by some rule—there is no clear conditional to use. The semantics of

   it.body

   differs from

   it.body

c) *Definition of Confidence:*

CLAIR's confidence addresses these problems by dropping normalization:

> **Definition** **Confidence Value.**
> A **confidence value** is a real number *c in [0, 1]*. We write *C* for the set of confidence values: *C = set(c in RR | 0 <= c <= 1)*

The semantic interpretation under calibrated reliability:

1) *c = 1*: Axiomatic acceptance (treated as foundational)
2) *c = 0*: Complete rejection (treated as impossibility)
3) *c = 0.5*: Maximal uncertainty (no evidence either direction)
4) *c > 0.5*: Net evidence for acceptance
5) *c < 0.5*: Net evidence for rejection

> **Definition** **Epistemic Commitment.**
> Confidence represents **epistemic commitment**: the degree to which an agent commits to a proposition based on available evidence and reasoning, *interpreted* as the calibrated reliability of the source or derivation producing the belief.
> Unlike probability:
> 1) **No normalization**: *conf(phi) + conf(not phi)* need not equal 1.
> 2) **Not frequency**: *conf(phi) = 0.9* does *not* mean
>
>   *it.body*
>
>   in a single case. It means
>
>   *it.body*
>
> 3) **Derivation-based**: Confidence propagates through inference rules, not conditioning.

> **Example** **Non-normalized confidence.**
> Consider the proposition
>
> *it.body*
>
> An honest assessment might be: *conf(no vulnerabilities) = 0.4* (some evidence from testing) *conf(has vulnerabilities) = 0.3* (some evidence from complexity) The sum *0.4 + 0.3 = 0.7 < 1* reflects residual uncertainty—neither hypothesis is well-supported. This is inexpressible in probability.

d) *Falsifiability Criteria:*

Following the review's requirement, we state explicitly what would falsify CLAIR's confidence model:

1) **Empirical falsification**: If sources systematically assign confidence *c* to propositions but are correct with frequency *f != c*, they are *uncalibrated*. CLAIR provides no mechanism to *detect* uncalibrated sources —this requires external validation.
2) **Semantic falsification**: If interpretation as

   it.body

leads to contradictions in the algebra (e.g., violations of monoid laws), the semantics would be inadequate. The Lean 4 formalization verifies algebraic consistency.

3) **Competing semantics**: If an alternative interpretation (e.g., pure probability or fuzzy truth) provides better empirical calibration or algebraic properties, CLAIR's semantics would need revision.

e) *Comparison with Subjective Logic:*

Jøsang's Subjective Logic (Jøsang, 2016) extends probability with explicit uncertainty. An *opinion* is a tuple *omega = (b, d, u, a)*:

1) *b*: belief mass (evidence for)
2) *d*: disbelief mass (evidence against)
3) *u*: uncertainty mass (lack of evidence)
4) *a*: base rate (prior probability)

with constraint *b + d + u = 1.*

CLAIR's confidence can be viewed as a simplification of Subjective Logic: *c = b + u times a* where we collapse uncertainty into the confidence value via the base rate. This loses the *b/d/u* decomposition but gains simplicity. The trade-off is appropriate for CLAIR's focus on derivation tracking rather than uncertainty quantification.

it.body

B. *The Aggregation Monoid*

When multiple independent pieces of evidence support the same conclusion, confidence should *increase*. This requires the probabilistic OR operation.

a) *Probabilistic OR (Oplus):*

Definition Probabilistic OR (Oplus).
For confidence values *a, b in C*, their **aggregation** is: *a oplus b = a + b - a times b*

The formula has several equivalent forms:

1) *a oplus b = a + b(1 - a)*
2) *a oplus b = a(1 - b) + b*
3) *a oplus b = 1 - (1 - a)(1 - b)* (De Morgan duality with multiplication)

The last form reveals the duality: *a oplus b* is the complement of the product of complements.

Theorem Oplus Preserves Bounds.
For all *a, b in C*: *a oplus b in C* *Proof.* **Lower bound**: *a oplus b = a + b(1 - a) >= 0* since *a >= 0, b >= 0*, and *(1 - a) >= 0.*

Upper bound: *a oplus b = a + b(1 - a) <= 1* since *b <= 1* implies *b(1 - a) <= 1 - a*, therefore *a + b(1 - a) <= a + (1 - a) = 1.* ∎

Theorem Oplus Monoid.
(C, oplus, 0) is a commutative monoid with absorbing element *1*:

1) **Associativity**: *(a oplus b) oplus c = a oplus (b oplus c)*
2) **Commutativity**: *a oplus b = b oplus a*
3) **Identity**: *0 oplus a = a oplus 0 = a*
4) **Absorption**: *1 oplus a = a oplus 1 = 1*

Proof. All properties follow from standard real number arithmetic on 0,1. ∎

b) *Confidence-Increasing Property:*

Unlike multiplication, *oplus increases* confidence:

Theorem Oplus is Confidence-Increasing.
For all *a, b in C*: *a oplus b >= max(a, b)* *Proof.* Using form *a oplus b = a + b(1-a)*. Since *b(1-a) >= 0*, we have *a oplus b >= a*. By commutativity, *a oplus b >= b*. Therefore *a oplus b >= max(a, b).* ∎

Corollary Diminishing Returns.
The marginal gain from additional evidence decreases as confidence grows: *(del)/(del b)(a oplus b) = 1 - a* When *a* is already high, additional evidence contributes less.

Example Aggregation of Weak Evidence.
Suppose we have ten independent pieces of weak evidence, each with confidence *0.3.* The combined confidence is: *0.3 oplus 0.3 oplus ... oplus 0.3 (10 times) = 1 - (1 - 0.3)^10 = 1 - 0.7^10 approx 0.972* Ten weak independent witnesses produce high combined confidence.

c) *The "Survival of Doubt" Interpretation:*

The formula *a oplus b = 1 - (1-a)(1-b)* admits a probability-theoretic interpretation under calibrated reliability:

1) *(1 - a)* is the

   it.body

   in evidence *a*
2) *(1 - a)(1 - b)* is the probability *both* pieces of evidence fail (assuming independence)
3) *a oplus b* is the probability *at least one* succeeds

This

it.body

interpretation explains why aggregation increases confidence: more independent evidence means more chances for at least one to be correct.

C. *The Multiplication Monoid*

When a conclusion is derived from premises, its confidence depends on the premises' confidences. The simplest case is conjunctive derivation: both premises must hold for the conclusion to follow.

a) *Conjunctive Confidence Propagation:*

Definition Confidence Multiplication.
For confidence values *a, b in C*, their **multiplicative combination** is standard multiplication: *a times b*

This models the intuition that deriving *C* from *A* and *B* requires both to be true. If we are 90% confident in *A* and

80% confident in B, our confidence in C (derived from both) is at most *0.9 times 0.8 = 0.72.*

Theorem Multiplication Preserves Bounds.
For all *a, b in C*: *a times b in C* *Proof.* We prove both bounds:

1) *a times b >= 0*: Since *a >= 0* and *b >= 0*, their product is non-negative.
2) *a times b <= 1*: Since *b <= 1*, we have *a times b <= a times 1 = a <= 1.*

∎

Theorem Multiplication Monoid.
(C, times, 1) is a commutative monoid with absorbing element *0*:

1) **Associativity**: *(a times b) times c = a times (b times c)*
2) **Commutativity**: *a times b = b times a*
3) **Identity**: *1 times a = a times 1 = a*
4) **Absorption**: *0 times a = a times 0 = 0*

Proof. All properties follow from standard real number arithmetic on 0,1. ∎

b) *The Derivation Monotonicity Principle:*

A fundamental property of CLAIR is that derivation can only decrease confidence:

Theorem Derivation Monotonicity.
For all *a, b in C*: *a times b <= min(a, b)* In particular, *a times b <= a* and *a times b <= b*. *Proof.* Since *b <= 1*, we have *a times b <= a times 1 = a*. By commutativity, *a times b <= b*. Therefore *a times b <= min(a, b)*. ∎

Corollary No Confidence Amplification.
No sequence of conjunctive derivations can increase confidence. If *c_0* is the confidence of a foundational belief and *c_n* is derived through *n* multiplicative steps, then *c_n <= c_0.*

This principle is essential for CLAIR's epistemology: derived beliefs are never more confident than their sources. Certainty (*c = 1*) is reserved for axioms, not conclusions.

D. Defeat Operations

Beyond derivation and aggregation, CLAIR requires operations for *defeat*: when evidence undermines a belief.

a) *Undercut: Attacking the Inference:*

Following Pollock (2001), an *undercutting defeater* attacks the inferential link, not the conclusion directly.

Definition Undercut.
For confidence *c* in a conclusion and defeat strength *d*: *undercut(c, d) = c times (1 - d)*

Example Undercutting Defeat.
Consider the inference:

it.body

An undercut

it.body

doesn't claim the object isn't red; it undermines the inference from appearance to reality.

If *conf(looks red => is red) = 0.9* and *conf(red lighting) = 0.6*, then: *undercut(0.9, 0.6) = 0.9 times (1 - 0.6) = 0.9 times 0.4 = 0.36* The inference confidence drops from 0.9 to 0.36.

Theorem Undercut Preserves Bounds.
For all *c, d in C*: *undercut(c, d) in C* *Proof.* Since *d <= 1*, we have *(1 - d) >= 0*. Since *c >= 0*, we have *c times (1 - d) >= 0*. Since *c <= 1* and *(1 - d) <= 1*, we have *c times (1 - d) <= 1.* ∎

Theorem Undercut Composition.
Sequential undercuts compose via *oplus*: *undercut(undercut(c, d_1), d_2) = undercut(c, d_1 oplus d_2)* *Proof.* By expanding the definition: *undercut(undercut(c, d_1), d_2) = c(1 - d_1)(1 - d_2)*

III. *c(1 - (d_1 + d_2 - d_1 times d_2))*

IV. *c(1 - (d_1 oplus d_2))*

V. *undercut(c, d_1 oplus d_2)*

∎

This beautiful result shows that defeat strengths aggregate via *oplus*: multiple undercuts combine as if their doubts aggregated.

a) *Rebut: Competing Evidence:*

A *rebutting defeater* provides counter-evidence for the conclusion directly.

Definition Rebut.
For confidence *c_for* in favor and *c_against* against: *rebut(c_for, c_against) = cases((c_for / (c_for + c_against), if c_for + c_against > 0), (0.5, if c_for = c_against = 0))*

The formula treats evidence symmetrically: the resulting confidence is the

it.body

of supporting evidence.

Theorem Rebut Preserves Bounds.
For all *c_for, c_against in C*: *rebut(c_for, c_against) in C* *Proof.* If *c_for + c_against = 0*, the result is *0.5 in [0,1]*. Otherwise:

1) *rebut >= 0* because *c_for >= 0* and the denominator is positive.
2) *rebut <= 1* because *c_for <= c_for + c_against.*

∎

Theorem **Rebut Antisymmetry.**
rebut(a, b) + rebut(b, a) = 1 *Proof.* When $a + b > 0$: *rebut(a, b) + rebut(b, a) = a/(a+b) + b/(a+b) = (a+b)/(a+b) = 1* When $a = b = 0$: *rebut(a,b) = rebut(b,a) = 0.5*, so the sum is *1.* ∎

b) *Rebut Normalization Limitation:*

The rebut operation has an important limitation that must be stated explicitly: it *normalizes away absolute strength information.*

Definition **Rebut Normalization Property.** For all $k > 0$: *rebut(k times a, k times b) = rebut(a, b)*

Proof. When *a, b* are not both zero: *rebut(k a, k b) = (k a) / (k a + k b) = k a / (k(a+b)) = a/(a+b) = rebut(a,b)* The *k* cancels from numerator and denominator. ∎

This means rebut only captures the *relative balance* of evidence, not the *absolute magnitude.* Consider:

Example **Normalization Loses Absolute Strength.**
Two scenarios with very different absolute evidence strengths:

1) **Scenario A**: *c_for = 0.1, c_against = 0.1* Result: *rebut(0.1, 0.1) = 0.1/(0.1+0.1) = 0.5*
2) **Scenario B**: *c_for = 0.9, c_against = 0.9* Result: *rebut(0.9, 0.9) = 0.9/(0.9+0.9) = 0.5*

Both scenarios yield confidence 0.5 despite Scenario B having *nine times more* total evidence. The rebut operation cannot distinguish "weak balanced evidence" from "strong balanced evidence."

Implications for CLAIR.

The normalization limitation means rebut is appropriate for *competitive evaluation* (comparing opposing arguments) but insufficient for *absolute assessment.* When absolute strength matters, CLAIR should:

1) Track total evidence magnitude separately: *total = c_for + c_against*
2) Use *undercut* instead when the goal is to reduce confidence proportionally to attack strength
3) Apply confidence thresholds to distinguish "weak 0.5" from "strong 0.5"

This is a deliberate design trade-off: rebut prioritizes interpretability ("market share" of evidence) over preserving absolute magnitude.

A. *Independence Assumptions for Aggregation*

The *oplus* operation assumes independence of evidence sources. This is a critical assumption that must be stated explicitly:

Definition **Conditional Independence for Oplus.**
Two evidence sources *e_1* and *e_2* are **independent** with respect to proposition *phi* if: *P(phi | e_1, e_2) = P(phi | e_1) + P(phi | e_2) - P(phi | e_1) times P(phi | e_2)* Under calibrated reliability, this means the reference classes of *e_1* and *e_2* do not systematically overlap.

a) *When Oplus is Valid:*

The *oplus* operation is **semantically justified** when:

1) **Independent sources**: Evidence derives from causally independent mechanisms (e.g., different sensors, different witnesses, different reasoning paths).
2) **No shared provenance**: The sources do not derive from common antecedents that would cause their errors to correlate.
3) **Reference class disjointness**: The calibration reference classes for *e_1* and *e_2* do not systematically overlap.

Example **Valid Independent Aggregation.**
Three different image classifiers (trained on different datasets, with different architectures) each assign confidence 0.7 to

it.body

. The combined confidence *0.7 oplus 0.7 oplus 0.7 approx 0.973* is justified because the classifiers make statistically independent errors.

b) *When Oplus Breaks:*

The *oplus* operation **overcounts** and produces misleading confidence when:

1) **Shared provenance**: Two sources derive from the same evidence. For example, two newspapers reporting the same press release should not be aggregated via *oplus.*
2) **Common systematic bias**: Sources that share the same misconception or training data flaw will make correlated errors. Aggregating them amplifies bias.
3) **Circular dependence**: When *e_2* cites *e_1* as a source, their evidence is not independent.

Example **Invalid Aggregation Due to Shared Provenance.**
A fact-checking website cites *Source A.* A blog post then cites the fact-checking website. Treating these as independent evidence would incorrectly inflate confidence via *oplus.*

c) *Correlation-Aware Alternatives:*

When independence is violated, CLAIR provides alternatives:

1) **Correlation-aware aggregation**: Chapter 4 introduces *oplus_delta* for correlated evidence, where *delta in [0,1]* measures dependency: *a oplus_delta(b) = a + b - a times b - delta*
2) **Min-based aggregation**: When sources may be completely dependent, use *max(a, b)* instead of *oplus* to avoid overcounting.
3) **Provenance tracking**: Evidence usage is tracked via provenance, preventing the same source from being counted twice in a derivation. This is enforced by CLAIR's DAG structure and source tracking (Chapter 10).

Independence Detection.

In practice, determining whether evidence sources are independent requires domain knowledge and provenance tracking. CLAIR does **not** automatically detect dependence—it provides the algebraic machinery for *manually* specifying correlation or preventing double-counting through provenance tracking.

d) *Interval-Based Confidence: Dependency Bounds:*

When dependence between evidence sources is unknown or partially known, CLAIR supports *interval-based confidence* as a robust alternative to point-value aggregation. Instead of committing to a single confidence value, we track upper and lower bounds that reflect uncertainty about dependence.

Definition Confidence Interval.

A **confidence interval** is a pair *[c_low, c_high]* with *0 <= c_low <= c_high <= 1*. The interpretation: the true confidence lies in this interval, with the width *c_high - c_low* reflecting uncertainty about evidence dependence.

Interval Aggregation Bounds:

For evidence with confidence values *a* and *b*, the independence assumption determines the correct aggregation:

1) **Independent**: *a oplus b = a + b - a times b* (probabilistic sum)
2) **Maximally dependent**: *max(a, b)* (no double-counting)
3) **Anti-correlated**: Could theoretically reach *min(a + b, 1)*

Theorem Interval Aggregation Bound.

For confidence values *a, b in C*, define: *oplus_indep(a, b) = a oplus b = a + b - a times b oplus_dep(a, b) = max(a, b)*

Then for any unknown correlation structure: *oplus_dep(a, b) <= true_aggregation <= oplus_indep(a, b)*

Proof. The lower bound *max(a, b)* occurs when evidence is fully dependent (one source's evidence is a subset of the other's). The upper bound *a oplus b* occurs when evidence is independent. Any positive correlation reduces the effective aggregation below *a oplus b*, while negative correlation is impossible for evidential support (evidence cannot be negatively correlated with the truth of what it supports). ∎

Definition Interval Aggregation.

For confidence intervals *[a_low, a_high]* and *[b_low, b_high]*: *[a_low, a_high] oplus_interval [b_low, b_high] = [max(a_low, b_low), a_high oplus b_high]*

The lower bound uses *max* (maximal dependence) and the upper bound uses *oplus* (independence).

Example Interval-Based Aggregation for Unknown Dependence.

Two news sources report the same fact. Each has confidence *0.7*, but we suspect possible shared provenance (they may have derived from the same press release).

Point-value approach with *oplus*: *0.7 oplus 0.7 = 0.91* This assumes independence and may be overconfident.

Interval-based approach: *[0.7, 0.7] oplus_interval [0.7, 0.7] = [max(0.7, 0.7), 0.7 oplus 0.7] = [0.7, 0.91]*

The interval *[0.7, 0.91]* captures our uncertainty: the true confidence could be as low as *0.7* (if the sources are fully dependent) or as high as *0.91* (if independent). The width *0.21* quantifies our dependence uncertainty.

Interval Propagation:

Intervals propagate through CLAIR's operations:

1) **Multiplication**: *[a_low, a_high] times [b_low, b_high] = [a_low times b_low, a_high times b_high]*
2) **Undercut**: *undercut([a_low, a_high], d) = [a_low times (1-d), a_high times (1-d)]*
3) **Rebut**: More complex; requires propagating both bounds through the ratio

Theorem Interval Bound Preservation.

If *[a_low, a_high]* and *[b_low, b_high]* are valid intervals (*0 <= a_low <= a_high <= 1*), then all CLAIR operations produce valid intervals.

Proof. Each operation preserves bounds: *oplus* preserves 0,1 (Theorem 3.3), multiplication preserves 0,1 (Theorem 3.10), and undercut preserves 0,1 (Theorem 3.16). Interval extensions apply the same bounds to endpoints. ∎

Dependency Bounds from Provenance Tracking:

CLAIR's linear type system (Chapter 10) enables *provenance tracking*, which provides tighter dependency bounds than worst-case intervals.

Definition Provenance-Aware Dependency Bound.

Let *S(e)* be the set of primitive sources (axioms, sensor inputs, LLM outputs) used to derive evidence *e*. For two evidences *e_1, e_2*:

1) If *S(e_1) and S(e_2)* are disjoint (disjoint provenance): independence is *plausible*, use *oplus*
2) If *S(e_1)* is a subset of *S(e_2)* or vice versa (subset provenance): maximal dependence, use *max*
3) If there is partial overlap in provenance: use *oplus_delta* with dependency estimated as the fraction of shared sources

This provides a *heuristic but principled* way to estimate correlation from provenance structure. The intuition: shared evidence sources correlate the derivations that depend on them.

Example Provenance-Aware Aggregation.

Three sources support a conclusion:

1) *e_1* depends on sensors *s_1, s_2*
2) *e_2* depends on sensors *s_2, s_3*
3) *e_3* depends on sensor *s_4*

Aggregating *e_1* and *e_2*: They share sensor *s_2*, so *delta approx 1/2*. Aggregating with *e_3*: Disjoint from *e_1, e_2*, so independence is plausible.

This is more precise than worst-case intervals and more honest than assuming full independence.

When to Use Intervals vs Point Values:

Use point values when:

1) Provenance tracking confirms disjoint sources
2) Domain knowledge guarantees independence
3) Computational efficiency is critical

Use intervals when:
1) Provenance is incomplete or unknown
2) Sources may share hidden dependencies
3) Conservative decision-making is required (e.g., safety-critical systems)
4) Auditing requires explicit uncertainty quantification

Relationship to Imprecise Probability.

Interval-based confidence in CLAIR is related to Walley's imprecise probability and Dempster-Shafer theory, but with key differences:
1) **Not a probability interval**: Our intervals bound *calibrated reliability*, not frequency in a reference class.
2) **Not belief/plausibility**: We don't track separate belief and plausibility masses as in Dempster-Shafer.
3) **Operational**: Intervals quantify uncertainty about *evidence dependence*, not about the proposition itself.

This interval-based approach directly addresses Hole A from the review: when independence assumptions are violated, CLAIR provides principled alternatives that avoid overcounting while preserving auditability.

B. Conclusion

This chapter established the algebraic and semantic foundation of CLAIR:
1) **Confidence as calibrated reliability**: We explicitly interpret confidence as the calibrated reliability of information sources and derivation processes. This addresses the review's concern about underspecified semantics.
2) **Independence assumptions**: The *oplus* operation requires conditional independence of evidence sources. We provide *oplus_delta* for correlated evidence and affine typing to prevent double-counting.
3) **Three monoids, not a semiring**: Multiplication (derivation), and oplus (aggregation) serve distinct semantic roles and do not distribute.
4) **Defeat operations**: Undercut and rebut formalize how evidence can undermine beliefs, with undercuts composing beautifully via oplus.
5) **Machine-verified**: The algebra is formalized in Lean 4, ensuring type safety and algebraic correctness.

The confidence system provides the numeric foundation. The next chapter develops the *structural* foundation: how beliefs connect through justification DAGs.

VI. Justification as Labeled DAGs

"An argument is not a proof. It is a reason for a belief —and reasons can be defeated."

— *John L. Pollock,* Defeasible Reasoning

This chapter develops the structural foundation of CLAIR: how beliefs connect through justification. We show that trees are inadequate for justification structure and that the correct model is a *directed acyclic graph with labeled edges*. The labels distinguish support from defeat, enabling defeasible reasoning where conclusions can be withdrawn when new evidence undermines their justifications.

A. The Inadequacy of Trees

a) *The Shared Premise Problem:*

Traditional approaches represent justification as trees: each conclusion has premises, which themselves have premises, forming an inverted tree structure. This model is elegant but insufficient.

Consider a simple computation that uses the same belief twice:

```
let population = belief(1000000, 0.95, source: census)n let sample_size = belief(1000, 0.90, source: survey)n let ratio = derive population, sample_size by dividen let inverse = derive sample_size, population by dividen let product = derive ratio, inverse by multiply
```

In a tree representation, each belief appears multiple times as separate subtrees. This creates three problems:
1) **Space inefficiency**: Beliefs are copied rather than shared.
2) **Invalidation complexity**: If a belief is invalidated, we must find and invalidate all copies.
3) **Semantic confusion**: Are these the *same* belief or *different* beliefs that happen to be equal?

The correct representation is a DAG with explicit sharing, where each belief appears exactly once with multiple parents.

> **Theorem** **DAG Necessity.**
> Any justification system that:
> 1) Allows a belief to be used as a premise in multiple derivations
> 2) Propagates invalidation correctly (removing a premise invalidates all conclusions)
> 3) Maintains identity (the "same" belief is the same node)
>
> must represent justification as a DAG, not a tree.
>
> *Proof.* In a tree, each node has exactly one parent. If a belief is used in derivations of two different conclusions, it must appear as a child of both. But in a tree, a node cannot have two parents. Therefore, the belief must be duplicated, violating identity. A DAG allows multiple parents, resolving the contradiction. ∎

b) *Why Not Cycles?:*

If we allow sharing, why not allow cycles? Coherentist epistemology suggests beliefs can mutually support each other. We reject cycles in justification for three reasons:
1) **Bootstrap problem**: Circular justification allows confidence inflation with no external grounding.
2) **Invalidation ambiguity**: If beliefs support each other circularly, invalidation semantics become ill-defined.
3) **Well-foundedness**: The justification relation should be well-founded, with no infinite descending chains.

> **Note:** The theory/observation circularity example is better analyzed as two separate relations:
> 1) **Evidential support** (tracked in justification): Observation provides evidence for theory.

2) **Interpretive framework** (not part of justification): Theory provides framework for interpreting observation.

B. *Labeled Edges for Defeat*

The DAG structure addresses sharing but not defeat. When evidence undermines a belief's justification, we need edges that carry negative, not positive, epistemic weight.

a) *The Defeat Problem:*

A *defeater* is a belief that undermines confidence in another belief. Following Pollock (2001), we distinguish:

1) **Undercutters**: Attack the inferential link, not the conclusion directly.
2) **Rebutters**: Provide direct counter-evidence against the conclusion.

Definition Edge Types.
A justification edge has one of three types:

1) **Support**: Positive evidence for the target.
2) **Undercut**: Attacks the inferential link to the target.
3) **Rebut**: Direct counter-evidence against the target.

b) *Formal Definition:*

Definition Justification Graph.
A **justification graph** is a tuple $G = (N, E, r)$ where:

1) N is a finite set of *justification nodes*
2) *E subset.eq N times N times set("support", "undercut", "rebut")* is a set of labeled edges
3) *r in N* is the root node

subject to the constraint that the underlying unlabeled graph is acyclic.

Definition Justification Node Types.
Each node has one of the following types:

1) axiom: Foundational belief (confidence = 1)
2) rule

   (r, premises)

   : Deductive rule application
3) assumption

   (a)

   : Temporary assumption for reasoning
4) choice

   (options, criteria)

   : Decision point
5) abduction

   (obs, hypotheses, selected)

   : Abductive inference
6) analogy

   (source, similarity)

   : Analogical reasoning
7) induction

   (cases, rule)

   : Inductive generalization
8) aggregate

   (sources, combRule)

   : Evidence aggregation

c) *Well-Formedness Constraints:*

A well-formed justification graph must satisfy explicit constraints to ensure semantic coherence and computational tractability.

Definition Acyclicity Constraint.
A justification graph must be acyclic: no path may exist from any node back to itself. This constraint applies specifically to support edges. Defeat edges may form cycles, which are resolved via fixed-point iteration (discussed below).

Definition Well-Formed Justification Graph.
A justification graph is well-formed iff:

1) The support structure is acyclic (no support cycles)
2) Every non-axiom node has at least one incoming support edge
3) Every node is reachable from the root in the underlying undirected graph

Condition 1 ensures the support structure is well-founded. Condition 2 ensures no "floating" beliefs without justification. Condition 3 ensures the graph is connected.

d) *DAG-Only vs Cyclic Choice:*

CLAIR adopts a hybrid approach to cycles:

1) **Support edges: DAG-only**. Cycles in evidential support are semantically prohibited because they enable bootstrapping and violate well-foundedness. The type checker enforces acyclicity for support edges at construction time.
2) **Defeat edges: Fixed-point semantics**. Defeat cycles are permitted and resolved via fixed-point iteration. When defeat edges form cycles, confidence propagation requires solving a system of equations.

This design choice reflects the epistemic distinction between evidence for (which must be well-founded) and attacks against (which can mutually undermine each other).

e) *Fixed-Point Semantics for Cyclic Defeat:*

When defeat edges form cycles, we compute confidences via iterative fixed-point finding.

Theorem Fixed-Point Existence.
For any defeat graph with acyclic underlying support, a fixed point exists.

Proof. The confidence update function maps the unit interval to itself and is continuous. By Brouwer's fixed point theorem, a continuous function from a compact convex set to itself has a fixed point. ∎

Theorem Fixed-Point Uniqueness.
If the maximum product of undercut strengths along any cycle is strictly less than 1, then the fixed point is unique and Kleene iteration converges to it.

Proof. Under this condition, the update function is a contraction mapping. By the Banach fixed point theorem, a contraction has a unique fixed point and iteration converges to it. ∎

Definition Kleene Iteration.
Starting from initial confidences, repeatedly apply the update function until convergence. The sequence generated by repeated application is the Kleene iteration.

Example Convergence for Typical Defeat Cycles.
Suppose two nodes mutually undercut with strength 0.5 each. The Lipschitz constant is 0.5 times 0.5 equals 0.25. Starting from initial confidence (1, 1):

1) After 1 iteration: error is at most 0.25 times initial error
2) After 5 iterations: error is at most 0.25^5 approximately 0.001 (0.1%)
3) After 10 iterations: error is at most 0.25^10 approximately 10^-6 (0.0001%)

Rapid convergence is typical for well-formed defeat graphs.

Practical Implication.
For CLAIR implementations, we recommend:
1) Enforce DAG structure for support edges via static type checking
2) Allow defeat cycles but limit undercut strengths to ensure contraction
3) Use Kleene iteration with convergence threshold 10^-6
4) Cache fixed-point solutions to avoid recomputation during updates

C. *Confidence Propagation*

Given a justification graph, we compute the confidence of each node bottom-up.

Definition Support Propagation.
For a node with children having confidences *c_1, ..., c_k*:
1) a) conf(axiom) = 1
2) conf(rule(r, children)) = s_r times product over children
3) conf(aggregate(children, independent)) = oplus over children

where *s_r* is the rule strength and *oplus* is probabilistic OR.

Definition Defeat Propagation.
Let *c* be confidence from support edges, *d_1, ..., d_m* be undercut strengths, and *r_1, ..., r_n* be rebut strengths. Then: *c' = rebut(undercut(c, oplus d_i), oplus r_j)*

Theorem Propagation Termination.
The propagation algorithm terminates for any acyclic justification graph.

Proof. The graph is acyclic by definition. Each recursive call moves to a node strictly earlier in topological order. Since the graph is finite, recursion terminates. ∎

D. *Reinstatement*

A fundamental phenomenon in defeasible reasoning is *reinstatement*: when a defeater is itself defeated, the original belief recovers some confidence.

Theorem Compositional Reinstatement.
Let *A* have base confidence *a*, undercut by *D* with confidence *d*, which is itself undercut by *E* with confidence *e*. Then: *conf(A) = a times (1 - d times (1 - e))*

The *reinstatement boost* is *Delta = a times d times e*.

Proof. Bottom-up evaluation gives: *conf_eff(D) = d times (1 - e)*, then *conf(A) = a times (1 - conf_eff(D)) = a times (1 - d(1-e))*. ∎

E. *Mutual Defeat*

When two arguments defeat each other, we have a defeat cycle. The fixed point analysis yields:

Theorem Mutual Defeat Fixed Point.
If *A* and *B* mutually undercut with base confidences *a* and *b*, the fixed point is: *a_star = a(1-b) / (1 - ab)* and *b_star = b(1-a) / (1 - ab)*

Proof. At fixed point: *a_star = a(1 - b_star)* and *b_star = b(1 - a_star)*. Solving gives the stated formulas. ∎

Theorem Fixed Point Existence.
For any defeat graph with acyclic underlying support, a fixed point exists.

Proof. The confidence update function maps *[0,1]^n* to itself and is continuous. By Brouwer's fixed point theorem, a fixed point exists. ∎

Theorem Uniqueness Condition.
If *b_max times d_max < 1*, the fixed point is unique and iteration converges.

Proof. Under this condition, the update function is a contraction mapping. By the Banach fixed point theorem, there is a unique fixed point and iteration converges. ∎

F. Correlated Evidence

The aggregation formula *c_1 oplus c_2* assumes independence. When evidence sources are correlated, this overcounts.

> **Definition Dependency-Adjusted Aggregation.**
> For evidence with confidences *c_1, c_2* and dependency *delta in [0,1]*: *aggregate_delta(c_1, c_2) = (1-delta)(c_1 oplus c_2) + delta times (c_1 + c_2) / 2*
> where *delta = 0* means independent and *delta = 1* means fully dependent.

G. Connection to Prior Art

a) *Truth Maintenance Systems:*

JTMS (Doyle 1979) uses IN/OUT lists for dependency-directed backtracking. ATMS (de Kleer 1986) tracks multiple assumption sets.

CLAIR contribution: Generalize TMS to graded confidence with the same dependency-directed architecture.

b) *Argumentation Frameworks:*

Dung's argumentation (Dung 1995) defines acceptance semantics. Gradual semantics (Amgoud et al. 2017) add numeric degrees.

CLAIR contribution: Integrate argumentation's defeat semantics with type-theoretic justification.

c) *Justification Logic:*

Artemov's Justification Logic (Artemov 2001) adds explicit justification terms to modal logic.

CLAIR contribution: Extend from tree-like justification terms to DAGs with labeled edges, supporting defeasible reasoning.

H. The Tracking Paradigm

The preceding sections developed the machinery of justification DAGs, confidence propagation, and defeat semantics. We now step back and articulate the underlying *tracking paradigm* that distinguishes CLAIR from traditional formal systems.

a) *State Representation:*

The Epistemic State:

A CLAIR system's **epistemic state** at any point in time is a pair:

> **Definition Epistemic state.**
> An epistemic state *cal(E)* is a tuple *cal(E) = (cal(G), cal(B))* where:
> 1) *cal(G) = set(G_1, dots, G_n)* is a finite set of justification graphs (one for each belief)
> 2) *cal(B) = set((v_i, c_i, j_i, p_i) mid i in I)* is a set of beliefs with values, confidences, justifications, and provenance

Each belief in *cal(B)* corresponds to the root of some graph in *cal(G)*. The graphs may share nodes (shared premises) but are not required to be connected.

Comparison with Proving Paradigm:

Traditional formal systems focus on *provability*:

1) **State**: A set of axioms and inference rules
2) **Question**: Is *phi* provable from the axioms? (*Gamma |- phi*)
3) **Output**: Yes/No (with proof term)

CLAIR's tracking paradigm focuses on *epistemic representation*:

1) **State**: A set of labeled graphs with confidence values
2) **Question**: What is the confidence, justification structure, and provenance of belief *phi*?
3) **Output**: *(c, G, p)* where *c in [0,1]* is confidence, *G* is the justification graph, *p* is provenance

b) *Update Rules:*

The epistemic state changes through *epistemic actions*. Each action maps *cal(E) -> cal(E)'*.

Primitive Actions:

> **Definition Epistemic actions.**
> The primitive actions on epistemic states are:
> 1) **Add belief**:
>
> ```
> add(phi, c, j, p)
> ```
>
> creates a new belief with confidence *c*, justification *j*, and provenance *p*.
> 2) **Aggregate**:
>
> ```
> aggregate(phi, psi, r)
> ```
>
> combines two beliefs about the same proposition using rule *r in set("independent", "correlated"(δ), "max")*.
> 3) **Derive**:
>
> ```
> derive(phi, [psi_1, ..., psi_k], rule)
> ```
>
> creates φ as a conclusion from premises using an inference rule.
> 4) **Undercut**:
>
> ```
> undercut(phi, delta, d)
> ```
>
> adds an undercutting defeater to φ with strength *d*.
> 5) **Rebut**:
>
> ```
> rebut(phi, delta, d)
> ```
>
> adds a rebutting defeater to φ with strength *d*.
> 6) **Invalidate**:
>
> ```
> invalidate(psi)
> ```
>
> removes belief ψ and propagates invalidation to all beliefs that depend on ψ.

Action Semantics:

> **Example Adding a belief.**
> When adding a belief
>
> ```
> add(phi, 0.8, source: testimony, witness_A)
> ```
>
> :

1) Create a new node n_φ with type

```
axiom
```

and confidence 0.8
2) Create a single-node graph G_φ containing only n_φ with no edges
3) Add $(n_\varphi, 0.8, \text{source: testimony, witness\_A})$ to $\mathcal{B}$
4) Add G_φ to $\mathcal{G}$

Example Deriving a conclusion.
When deriving

```
derive(phi, [psi_1, psi_2], modus_ponens)
```

:
1) Create a new node n_φ with type

```
rule(modus_ponens, [n_psi_1, n_psi_2])
```

2) Create support edges $(n_{\psi_1}, n_\varphi, \text{support})$ and $(n_{\psi_2}, n_\varphi, \text{support})$
3) Compute confidence via propagation: $c_\varphi = c_{\psi_1} \times c_{\psi_2}$
4) Add graph G_φ to $\mathcal{G}$ and belief $(n_\varphi, c_\varphi, G_\varphi, \text{derived})$ to $\mathcal{B}$

Example Invalidation propagation.
When

```
invalidate(psi)
```

is called:
1) Remove ψ from $\mathcal{B}$
2) Find all beliefs φ such that ψ appears in G_φ's support graph
3) Recursively invalidate each such φ (mark as defeated or recompute without ψ)
4) Update confidence values via re-propagation

This is the *dependency-directed backtracking* inherited from TMS systems.

c) *Correctness Criteria:*

What does it mean for the tracking paradigm to be "correct"? We distinguish three levels of correctness.

Syntactic Correctness:

Definition Syntactic correctness.
An epistemic state *cal(E)* is *syntactically correct* if:
1) Every graph in *cal(G)* is acyclic (support edges)
2) Every node has a well-defined type
3) Every edge has a valid label (support, undercut, rebut)
4) All confidence values are in *[0,1]*

Syntactic correctness is enforced by the type system and checked at runtime. The Lean 4 formalization proves that propagation preserves syntactic correctness.

Semantic Correctness (Internal):

Definition Semantic correctness (internal).
An epistemic state *cal(E)* is *semantically correct* if:
1) Confidence propagation is consistent: Re-computing any belief's confidence yields the same value
2) Defeat semantics are satisfied: Undercuts and rebuts are applied according to their definitions
3) Bounds are preserved: All confidences remain in *[0,1]* after any update

This is *internal* correctness: the system behaves according to its own rules. The theorems on propagation termination and sound establish internal correctness for confidence propagation.

Semantic Correctness (External):

Definition Semantic correctness (external).
An epistemic state *cal(E)* is *externally correct* with respect to a reference system *cal(R)* if:
1) **Calibration**: For any belief with confidence *c*, the reference system assigns probability *c' approx c* (within calibration tolerance)
2) **Justification adequacy**: The justification graph *G* accurately represents the actual dependency structure in *cal(R)*
3) **Provenance accuracy**: The provenance field correctly identifies the source of the belief in *cal(R)*

External correctness cannot be established by formal proof alone—it requires empirical validation. CLAIR provides the *machinery* for tracking (internal correctness) but calibration and accuracy are properties of the *sources*, not the tracking system.

The Honesty Principle.

CLAIR's tracking paradigm embodies epistemic humility:
1) The system *reports* its confidence, justification, and provenance
2) It does not *guarantee* that these correspond to external reality
3) External correctness must be validated empirically (calibration studies, provenance audits)

This distinguishes tracking from proving: a proof claims certainty; a tracked belief admits uncertainty while documenting its grounds.

d) *Tracking vs. Proving: A Summary:*

Aspect	Proving Paradigm	Tracking Paradigm
Goal	Establish truth	Record epistemic state
State	Set of formulas	Set of labeled graphs
Query	$\Gamma \mid - \varphi?$	What is the *confidence*, *justification*, and *provenance* of φ?
Output	Proof term (or $\perp$)	(c, G, p) triple
Update	Add axiom/formula	Add/undercut/rebut/invalidate
Correctness	Soundness	Internal + External

Aspect	Proving Paradigm	Tracking Paradigm
Limits	Hidden or denied	Explicit with confidence

e) *Practical Implications:*

The tracking paradigm has several practical consequences for CLAIR as an AI reasoning intermediate representation:

Explainability:

Every belief carries its justification graph. Queries like "why does the system believe φ?" can be answered by traversing the graph and presenting the dependency chain.

Debugging:

When a belief has unexpectedly low confidence, the graph reveals which defeaters are responsible. When confidence is too high, the graph shows whether aggregation rules were misapplied.

Revision:

New evidence is incorporated by adding nodes and edges. The system automatically recomputes confidences via propagation, handling reinstatement without special cases.

Uncertainty:

Unlike binary logic, CLAIR tracks degrees of belief. This matches the uncertainty inherent in real-world reasoning and LLM outputs.

I. *Conclusion*

This chapter established the structural foundation of CLAIR and articulated the tracking paradigm:

1) **DAGs, not trees**: Shared premises require graph structure; explicit sharing enables correct invalidation.
2) **Acyclic**: Cycles in evidential support violate well-foundedness; defeat cycles handled via fixed-point semantics.
3) **Labeled edges**: Support, undercut, and rebut serve different epistemic roles.
4) **Compositional reinstatement**: Defeaters being defeated automatically recovers confidence.
5) **Correlated evidence**: Independence assumptions must be explicit; dependency adjustment prevents overcounting.
6) **Tracking paradigm**: CLAIR represents epistemic state as labeled graphs with confidence values, updating via add/derive/undercut/rebut/invalidate operations. Correctness has three levels: syntactic (well-formedness), internal semantic (consistency with propagation rules), and external semantic (calibration to reality).

The justification DAG provides the structural substrate for CLAIR's beliefs. The tracking paradigm formalizes what it means to "track not prove"—a fundamental shift from establishing truth to documenting epistemic grounds with explicit confidence.

1) **DAGs, not trees**: Shared premises require graph structure; explicit sharing enables correct invalidation.
2) **Acyclic**: Cycles in evidential support violate well-foundedness; defeat cycles handled via fixed-point semantics.
3) **Labeled edges**: Support, undercut, and rebut serve different epistemic roles.
4) **Compositional reinstatement**: Defeaters being defeated automatically recovers confidence.
5) **Correlated evidence**: Independence assumptions must be explicit; dependency adjustment prevents overcounting.

The justification DAG provides the structural substrate for CLAIR's beliefs. The next chapter addresses a subtler challenge: how beliefs can safely refer to themselves.

V

Self-Reference and the Gödelian Limits

line(length: 20%, stroke: 1.5pt, paint: academic-burgundy)

"If a system is consistent, it cannot prove its own consistency."
— *Kurt Gödel,* On Formally Undecidable Propositions

CLAIR allows beliefs about beliefs. This reflexive capacity creates potential for self-reference: a belief that refers to itself, either directly or through a chain of intermediate beliefs. Such self-reference is both a powerful expressive tool and a source of potential paradox. This chapter develops the theoretical foundations for distinguishing safe from dangerous self-reference, culminating in a novel extension of provability logic to graded confidence.

J. The Problem of Self-Reference

a) *Direct Self-Reference:*
Consider a belief that directly references itself:

Example Direct Self-Reference.

```
-- A belief referencing its own confidence
let b : Belief<Bool> = belief(
  value: confidence(b) > 0.5,
  confidence: ???
)
```

What confidence should b have? If we assign confidence $c > 0.5$, the content becomes true, which seems consistent. But if we assign $c <= 0.5$, the content becomes false—yet what prevents us from assigning $c = 0.9$ anyway?

This is not merely a curiosity. If CLAIR aims to capture how an LLM reasons, and if introspection is part of reasoning, then CLAIR must account for self-referential beliefs —even if that account restricts or forbids certain patterns.

b) *Why Self-Reference Matters:*
Self-reference enables powerful epistemic capabilities:
1) **Calibration**: "My confidence estimates are typically accurate"
2) **Uncertainty tracking**: "I am uncertain about this belief"
3) **Meta-reasoning**: "I should reconsider beliefs derived from unreliable sources"
4) **Self-improvement**: "My reasoning process could be improved in specific ways"

But self-reference also enables paradoxes:
1) **Liar-like**: "This belief has confidence 0" (no consistent assignment)
2) **Curry-like**: "If this belief is true, then arbitrary proposition P" (proves anything)
3) **Löbian**: "If I believe P, then P is true" (circular self-validation)

The challenge is to permit the former while blocking the latter.

K. Löb's Theorem and Anti-Bootstrapping

a) *The Classical Result:*
Löb's theorem is a cornerstone of provability logic:

Theorem Löb's Theorem.
In any sufficiently strong formal system T containing arithmetic: $\text{prov}_{T(\text{prov}(\text{prov}(P) \to P))} \to \text{prov}(P)$ where prov denotes provability in T.

In words: if a system can prove "if P is provable, then P is true," then the system can prove P. This has a startling consequence.

Corollary No Internal Soundness Proof.
No consistent system can prove its own soundness, i.e., cannot prove $\forall P, \Box(P) \to P$.

Proof. Suppose system T proved $\forall P, \Box(P) \to P$. Instantiating with $P =$ *"false"* (falsity), we get $\text{prov}(\text{false}) \to \text{false}$. Combining with consistency ($\neg\ \text{prov}(\text{false})$), we can derive $\text{prov}(\text{false})$, contradicting consistency. ∎

b) *Application to CLAIR:*
For CLAIR, interpret $\Box(P)$ as "CLAIR believes P with confidence 1.0." Then Löb's theorem constrains self-soundness beliefs:

```
-- A claimed self-soundness belief
let soundness = belief(
  value: forall P. (belief(P, c, ...) and c > 0.9) -> P is true,
  confidence: 0.95
)
```

By Löb's theorem, if CLAIR can form this belief with high confidence, then (classically) CLAIR believes everything with high confidence—a collapse to triviality. This is the *bootstrapping trap*: self-soundness claims cannot increase epistemic authority.

> **Definition** **Anti-Bootstrapping Principle.**
> A belief system satisfies *anti-bootstrapping* if no belief of the form "my beliefs are sound" can increase confidence in any derived belief beyond what the original evidence supports.

Löb's theorem mathematically enforces anti-bootstrapping for classical provability. The question is how this extends to graded confidence.

L. *Tarski's Hierarchy: Stratified Introspection*

a) *The Classical Solution:*

Tarski's theorem on the undefinability of truth states that no sufficiently expressive language can define its own truth predicate—on pain of the Liar paradox. Tarski's solution is stratification:

Level	Can Express	Cannot Express
Level 0 (object)	Facts about the world	Truth of any sentence
Level 1 (meta)	X_0 is true for level-0 X	Truth of level-1 sentences
Level 2 (meta-meta)	X_1 is true for level-1 X	Truth of level-2 sentences
...	...	...

Each level can discuss truth at lower levels but never its own level.

b) *Stratified Beliefs in CLAIR:*

We apply this to beliefs:

> **Definition** **Stratified Belief Type.**
> $\mathrm{Bel}(n, A)$ for $n \in \mathcal{N}\mathcal{N}$ where level-n beliefs may reference level-m beliefs only if $m < n$.

> **Example** **Stratified Beliefs in CLAIR.**
> In the CLAIR IR format, each belief has an explicit level field:
>
> ```
> ; Level 0: beliefs about the world (no introspection)
> b1 .9 L0 @self "user is authenticated"
>
> ; Level 1: beliefs about level-0 beliefs
> b2 .95 L1 @self <b1 "my auth belief b1 has high
> confidence"
>
> ; Level 2: beliefs about level-1 beliefs
> b3 .9 L2 @self <b2 "my introspection in b2 seems
> accurate"
> ```
>
> The level constraint is: a belief at level n can only justify beliefs at level $\geq n$, and a belief *about* another belief must be at a higher level.

> **Theorem** **Stratification Safety.**
> If all beliefs respect the stratification constraint—$\mathrm{Bel}(n, A)$ references only $\mathrm{Bel}(m, B)$ with $m < n$—then no Liar-like paradox can arise.

Proof. Any reference chain from a belief b must strictly decrease in level. Since $\mathbb{N}$ has no infinite descending chains, every chain terminates at level 0. Level-0 beliefs contain no belief references, so they cannot participate in self-referential loops. Therefore, no belief can reference itself directly or transitively. ∎

c) *What Stratification Rules Out:*

Stratification prohibits:

1) **Direct self-reference**: A belief cannot mention itself (would require level $n < n$).
2) **Universal introspection**: "All my beliefs are..." spans all levels and cannot be expressed at any finite level.
3) **Self-soundness at a single level**: "My level-n beliefs are sound" would require level $n+1$ to express.

d) *The Cost of Safety:*

Stratification is safe but restrictive. Some legitimate self-referential reasoning is blocked:

```
; Legitimate but blocked: calibration beliefs
; This would be self-referential (talks about own
confidences)
; but is intuitively safe (no paradox)
b1 .8 L? @self "my confidence estimates match empirical
accuracy"
; What level should this be? It references "all my
confidences"
; which spans all levels---cannot be expressed at any
finite level
```

This motivates a more permissive approach for certain cases.

M. *Kripke's Fixed Points: Safe Self-Reference*

a) *The Fixed-Point Construction:*

Kripke proposed an alternative to stratification: allow self-reference but let some sentences remain *undefined*. The key insight is that certain self-referential constructs have *fixed points*— consistent confidence assignments—while others do not.

> **Definition** **Fixed Point for Self-Referential Belief.**
> A self-referential belief b with confidence function f : *cal("D")cal("D")[0,1] -> cal("D")cal("D")[0,1]* (determining confidence from the assumed truth value) has a fixed point if there exists c *in cal("D")cal("D")[0,1]* such that: $c = f(c)$

> **Example** **Truth-Teller: Multiple Fixed Points.**
> Consider:
>
> ```
> let tt = self_ref_belief(fun self =>
>   content: "this belief is true",
>   compute_confidence: if val(self.content) then 1.0 else
> ```

```
0.0
)
```

If confidence is 1.0: content is true, so confidence should be 1.0. ✓ If confidence is 0.0: content is false, so confidence should be 0.0. ✓

Both are fixed points. The belief is *underdetermined.*

Example **Liar: No Fixed Point.**

Consider:

```
let liar = self_ref_belief(fun self =>
  content: "this belief has confidence 0",
  compute_confidence: if val(self.content) then 1.0 else
0.0
)
```

If confidence is 1.0: content says "confidence 0," which is false, so confidence should be 0.0. Contradiction. If confidence is 0.0: content says "confidence 0," which is true, so confidence should be 1.0. Contradiction.

No fixed point exists. The belief is *ill-formed.*

Example **Grounded Self-Reference: Unique Fixed Point.**

Consider:

```
let careful = self_ref_belief(fun self =>
  content: "confidence(self) is in [0.4, 0.6]",
  compute_confidence: 0.5
)
```

The compute function is constant, so $f(c) = 0.5$ for all c. The fixed point is $c = 0.5$, which indeed satisfies $0.5 \in [0.4, 0.6]$. This belief is *well-formed* with unique confidence 0.5.

b) *The Self-Reference Escape Hatch:*

CLAIR provides a controlled mechanism for self-reference:

When a CLAIR trace analyzer encounters self-referential beliefs that escape finite stratification, it classifies them:

```
Self-Reference Analysis Result:
  - WellFormed: unique fixed point exists (safe)
  - IllFormed: no fixed point (Liar-like, Curry-like, or
Löbian trap)
  - Underdetermined: multiple fixed points (policy choice
needed)

Error Types:
  - NoFixedPoint: Liar-like paradox
  - CurryLike: proves anything
  - LobianTrap: self-soundness claim
  - Timeout: analysis did not terminate
```

Beliefs that escape finite stratification exist "outside" all levels and require special analysis.

c) *Classification of Self-Reference:*

Combining Tarski and Kripke, we classify self-referential constructs:

Category	Fixed Points	Status	Example
Grounded	Unique	Safe	Calibration beliefs
Underdetermined	Multiple	Policy choice	Truth-teller
Liar-like	None	Ill-formed	"Confidence is 0"
Curry-like	—	Banned	"If true, then P"
Löbian	—	Banned	Self-soundness

Definition **Safe Self-Reference.**

A self-referential belief is *safe* if it either:

1) Respects stratification (level-n references only level-$m < n$), or
2) Has a unique fixed point (Kripke), or
3) Has multiple fixed points with a deterministic policy for selection.

Definition **Dangerous Self-Reference.**

A self-referential belief is *dangerous* if it:

1) Has no fixed point (Liar-like), or
2) Matches a Curry pattern ("if this then P"), or
3) Claims self-soundness (Löbian trap).

N. *Provability Logic and CLAIR*

a) *Gödel-Löb Logic (GL):*

To formally characterize CLAIR's belief logic, we turn to *provability logic.* The standard modal logic of provability is GL (Gödel-Löb):

Definition **GL Syntax.**

$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \varphi \to \varphi \mid \Box\varphi$ where $\Box\varphi$ means φ is provable.

Definition **GL Axioms.**

1) **K (Distribution)**: $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$
2) **4 (Positive Introspection)**: $\Box\varphi \to \Box\Box\varphi$
3) **L (Löb)**: $\Box(\Box\varphi \to \varphi) \to \Box\varphi$

Critically, GL *lacks* the truth axiom $\Box\varphi \to \varphi$ (T). This is philosophically essential: provability does not imply truth. A consistent system can prove false statements if its axioms are wrong.

b) *GL vs Other Modal Logics:*

Logic	K	T	4	5 or L
K	✓			
T	✓	✓		

Logic	K	T	4	5 or L
S4	✓	✓	✓	
S5	✓	✓	✓	5
GL	✓		✓	L
CLAIR	✓		✓	L

CLAIR aligns with GL:

1) **K holds**: If CLAIR believes an implication and believes the antecedent, it can derive the consequent.
2) **T fails**: CLAIR's beliefs can be wrong (fallibilism).
3) **4 holds**: CLAIR can have meta-beliefs about its beliefs.
4) **L must hold**: Self-soundness claims cannot bootstrap confidence.

c) *Solovay's Completeness:*

Theorem Solovay Completeness.
GL is sound and complete with respect to:
1) Arithmetic provability: GL $\vdash \varphi$ iff φ holds under all interpretations of $\Box$ as Gödel provability in PA.
2) Finite transitive irreflexive Kripke frames.

The completeness for finite frames yields:

Corollary GL Decidability.
GL is decidable (PSPACE-complete).

This is crucial: classical provability logic is computationally tractable.

O. Confidence-Bounded Provability Logic (CPL)

Classical GL uses binary truth: propositions are either provable or not. CLAIR needs a *graded* version where beliefs carry confidence values in cal("D")cal("D")[0,1]. This section introduces CPL (Confidence-Bounded Provability Logic), a novel extension of GL designed for CLAIR.

a) *The Literature Gap:*

Extensive work exists on fuzzy modal logics and graded epistemic logic. However, no prior work addresses:

1) Graded versions of the Löb axiom
2) The interaction of continuous confidence with provability constraints
3) Anti-bootstrapping in the context of graded belief

CPL fills this gap.

b) *CPL Syntax:*

Definition CPL Syntax.
$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi \mid \varphi \to \varphi \mid \Box_c\varphi$ where $\Box_c\varphi$ means φ is believed with confidence at least c.

c) *CPL Semantics:*

Definition Graded Kripke Frame.
A *graded Kripke frame* is a tuple (W, R) where:
1) W is a non-empty set of worlds
2) $R : W \times W \to \mathcal{D}[0,1]$ is a graded accessibility relation

satisfying:

1) **Transitivity**: $R(w,v) \cdot R(v,u) \leq R(w,u)$
2) **Converse well-foundedness**: No infinite sequence $w_0, w_1, w_2, \ldots$ with $R(w_{i+1}, w_i) > 0$ for all i

Definition Graded Valuation.
A *graded valuation* on a frame (W, R) assigns to each world w and proposition p a confidence value $V_{w(p)} \in \mathcal{D}[0,1]$. Extended to formulas:
1) $V_{w(\neg\varphi)} = 1 - V_{w(\varphi)}$
2) $V_{w(\varphi\wedge\psi)} = V_{w(\varphi)} \cdot V_{w(\psi)}$
3) $V_{w(\varphi\vee\psi)} = V_{w(\varphi)} + V_{w(\psi)} - V_{w(\varphi)}V_{w(\psi)}$
4) $V_{w(\varphi\to\psi)} = \sup\left\{c \in \mathcal{D}[0,1] : V_{w(\varphi)} \cdot c \leq V_{w(\psi)}\right\}$
5) $V_{w(\Box_c\varphi)} = \inf_{v:R(w,v)\geq c} V_{v(\varphi)}$

The last clause says: *phi* is believed at confidence *c* if *phi* holds in all worlds accessible with strength at least *c*.

d) *The Graded Löb Axiom: DESIGN AXIOM:*

The crucial innovation in CPL is the graded analogue of Löb's axiom.

Axiom Status Statement.

The Graded Löb axiom is a **DESIGN AXIOM**, not a semantic theorem derived from more basic principles. It is motivated by the requirement of anti-bootstrapping and the need to extend GL to graded confidence while preserving its essential character. The axiom is **posited** as part of CPL's definition, not **proved** from the semantics.

The key question is: Is CPL consistent? We address this below by exhibiting a non-trivial model satisfying all CPL axioms.

Definition Graded Löb Axiom (Design Axiom).
$\Box_{c(\Box_c\varphi\to\varphi)} \to \Box_{g(c)}\varphi$ where $g : \mathcal{D}[0,1] \to \mathcal{D}[0,1]$ is a *discount function* satisfying $g(c) \leq c$.

This is a **DESIGN AXIOM**—not derived from more basic principles but posited as part of CPL's definition. The axiom is motivated by anti-bootstrapping requirements.

The function g captures the *cost* of self-soundness claims. If you believe at confidence c that "believing *phi* at *c* implies *phi*," you can derive *phi* only at the discounted confidence *g(c)*.

e) *Choosing the Discount Function:*

We require g to satisfy:

1) **Boundedness**: $g : \mathcal{D}[0,1] \to \mathcal{D}[0,1]$
2) **Non-amplification**: $g(c) \leq c$ for all c
3) **Monotonicity**: $c_1 \leq c_2 \Rightarrow g(c_1) \leq g(c_2)$
4) **Anchoring**: $g(0) = 0$ and $g(1) = 1$
5) **Non-triviality**: $g(c) < c$ for $c \in (0,1)$

After analyzing several candidates (identity, parabolic, constant offset, product), we recommend:

Definition Quadratic Discount.
$g(c) = c^2$

Theorem Quadratic Discount Properties.
The quadratic discount $g(c) = c^2$ satisfies all desiderata and:

1) Aligns with CLAIR's multiplicative confidence algebra ($c^2 = c \times c$)
2) Has intuitive meaning: self-soundness costs "deriving the claim twice"
3) Produces strong anti-bootstrapping: iterated application $c \to c^2 \to c^4 \to \ldots \to 0$

Proof. Boundedness and anchoring are immediate ($c \in \mathcal{D}[0,1] \Rightarrow c^2 \in \mathcal{D}[0,1]$, $0^2 = 0$, $1^2 = 1$). For non-amplification: $c^2 \leq c$ when $c \leq 1$, with equality only at 0 and 1. Monotonicity: $c_1 \leq c_2 \Rightarrow c_1^2 \leq c_2^2$ on $\mathcal{D}[0,1]$. Non-triviality: $c^2 < c$ for $c \in (0,1)$. ∎

f) *The Anti-Bootstrapping Theorem:*

Theorem Anti-Bootstrapping.
In CPL with $g(c) = c^2$: $\mathrm{conf}\big(\Box_{c(\Box_c\varphi\to\varphi)}\big) = c \Rightarrow \mathrm{conf}(\varphi) \leq c^2 < c$ Consequently, no finite chain of self-soundness claims can increase confidence beyond the initial level.

Proof. Applying the Graded Löb axiom to the hypothesis yields $\mathrm{conf}(\Box_{c^2}\varphi) \leq c^2$. Iterating: $c \to c^2 \to c^4 \to c^8 \to \ldots$. For any $c < 1$, this sequence converges to 0. Self-soundness claims can only decrease confidence. ∎

This is the mathematical formalization of anti-bootstrapping: claiming your own soundness provides no epistemic free lunch.

g) *Modal Axioms in CPL:*

We now explicitly list the modal axioms CPL adopts and their status:

Definition CPL Modal Axiom Status.
1) **K (Distribution)**: $\Box_{c(\varphi\to\psi)} \to (\Box_c\varphi \to \Box_c\psi)$ — **VALID** Derivable from the semantics via graded accessibility.
2) **4 (Positive Introspection)**: $\Box_c\varphi \to \Box_c\Box_c\varphi$ — **VALID** Follows from transitivity of the accessibility relation.
3) **GL/Graded Löb**: $\Box_{c(\Box_c\varphi\to\varphi)} \to \Box_{g(c)}\varphi$ — **DESIGN AXIOM** Posited as part of CPL; motivated by anti-bootstrapping requirements.
4) **T (Reflexivity/Truth)**: $\Box_c\varphi \to \varphi$ — **INVALID** Explicitly **rejected** in CPL. Provability/belief does not imply truth. This is essential for fallibilism.

h) *CPL Consistency:*

To establish CPL's consistency, we exhibit a non-trivial model:

Theorem CPL Consistency.
CPL is consistent. There exists a non-trivial graded Kripke model satisfying all CPL axioms including the Graded Löb axiom.

Proof. **Proof Sketch.**
Consider the frame (W, R) where:
1) $W = \mathbb{N}$ (the natural numbers)
2) $R(i,j) = 2^{-i}$ if $i < j$, and $R(i,j) = 0$ otherwise

This frame satisfies:
1) **Transitivity**: If $i < j < k$, then $R(i,j) \times R(j,k) = 2^{-i} \times 2^{-j} \leq 2^{-i} = R(i,k)$
2) **Converse well-foundedness**: No infinite sequence $w_0, w_1, \ldots$ with $R(w_{i+1}, w_i) > 0$, since this would require an infinite decreasing sequence of natural numbers.

Define valuation $V_{w(\varphi)} = 1$ for all propositional variables φ at all worlds w.

For the Graded Löb axiom, consider any world w. If $V_{w\left(\Box_{c(\Box_c\varphi\to\varphi)}\right)} = 1$, then for all v with $R(w,v) \geq c$, we have $V_{v(\Box_c\varphi\to\varphi)} = 1$. By the structure of R, this forces $V_{v(\varphi)} = 1$ for all accessible worlds, yielding $V_{w(\Box_{c^2}\varphi)} = 1$.

This model is non-trivial (not all formulas are valid) yet satisfies all CPL axioms, establishing consistency. ∎

P. Decidability of CPL

Classical GL is decidable. Does CPL inherit this property?

a) *The Vidal Result:*

Theorem Vidal's Undecidability Theorem.
Transitive modal logics over many-valued semantics (including Łukasiewicz and Product algebras) are undecidable, even when restricted to finite models.

CPL has transitivity (axiom 4) and continuous $\mathcal{D}[0,1]$ values. The Vidal result strongly suggests:

Conjecture CPL Undecidability.
Full CPL (with continuous $\mathcal{D}[0,1]$ confidence) is undecidable.
Confidence: 0.80
We assign confidence 0.80 to this conjecture based on the close analogy to Vidal's proof technique.

b) *The Role of Converse Well-Foundedness:*

GL's decidability relies on the finite model property: converse well-foundedness forces finite-depth evaluation. Could this rescue CPL?

Proposition Insufficient for Decidability.
Converse well-foundedness alone does not rescue CPL from undecidability.
Justification: Converse well-foundedness constrains *structure* (no infinite ascending chains) but not *values*. The encoding power of continuous $\mathcal{D}[0,1]$ values combined with transitivity enables undecidable problem encodings even in well-founded frames.

c) *Decidable Fragments:*

Despite the likely undecidability of full CPL, we identify decidable fragments:

CPL-finite: Discrete Confidence:

Restrict confidence to a finite lattice instead of continuous $\mathcal{D}[0,1]$:

Definition CPL-finite.
Let $L_n = \left\{0, \frac{1}{n-1}, \frac{2}{n-1}, \ldots, 1\right\}$. CPL-finite evaluates over L_n with discretized operations:
1) $a \times b = \lfloor a \times b \rfloor$

2) $a + b = \lceil a + b - a \times b \rceil$
3) $g_{L(c)} = \lfloor c^2 \rfloor$

For $L_5 = \{0, 0.25, 0.5, 0.75, 1\}$:

c	c^2	$g_{L.5.}(c)$
0	0	0
0.25	0.0625	0
0.5	0.25	0.25
0.75	0.5625	0.5
1	1	1

Theorem CPL-finite Decidability.
CPL-finite is decidable via the finite model property.
Proof Sketch.
By the theorem of Bou, Esteva, and Godo, many-valued modal logics over finite residuated lattices are decidable. CPL-finite evaluates over L_n, a finite lattice. The frame constraints (transitivity, converse well-foundedness) are expressible, and finitely many models of bounded size suffice for completeness.
A complete formal proof would establish: (1) L_n forms a finite residuated lattice under the discretized operations; (2) the frame conditions are expressible in the corresponding modal logic; (3) the finite model property holds; and (4) decidability follows from (3). The proof follows the standard technique for finite-valued modal logics.

Conjecture CPL-finite Complexity.
CPL-finite is PSPACE-complete, analogous to classical GL.

CPL-0: Stratified Only:
Restrict to stratified beliefs without any self-reference:

Definition CPL-0.
CPL-0 disallows nesting of $\Box$ operators that would require the Löb axiom. Formally: only formulas of the form $\Box_c \varphi$ where φ is box-free.

Theorem CPL-0 Decidability.
CPL-0 is decidable (trivially: the restricted syntax avoids undecidability).

d) *Trade-offs:*

Fragment	**Decidable?**	**Expressiveness**	**Use Case**
Full CPL	Likely no	Full	Theoretical analysis
CPL-finite	Yes	Discrete confidence	Type-level checks
CPL-0	Yes	No self-reference	Stratified beliefs

Q. *Alternative: CPL-Gödel*

An alternative approach uses Gödel algebra (min/max) instead of product operations:

Definition CPL-Gödel.
1) $a \times b = \min(a, b)$
2) $a + b = \max(a, b)$

Conjecture CPL-Gödel Decidability.
CPL-Gödel is decidable because Gödel modal logic has the finite model property via quasimodels.
Confidence: 0.75
The conjecture follows from the known decidability of Gödel modal logic, but requires verification that the graded Löb axiom preserves this property.

However, CPL-Gödel is *semantically inappropriate* for CLAIR:
1) **max fails aggregation**: $\max(0.6, 0.6) = 0.6$, but two independent pieces of evidence should yield higher confidence (0.84 with $+$).
2) **min lacks degradation**: $\min(a, a) = a$, but derivation should cost confidence.
3) **No algebraic discount**: The c^2 discount becomes purely frame-based, losing the anti-bootstrapping semantics.

Recommendation.
For CLAIR, use CPL-finite (with product operations), not CPL-Gödel. Accept the discretization rather than sacrifice semantic fidelity.

R. *"Conservative Over GL": Clarification*

On "Conservative Over GL" Claims.
The phrase "CPL is conservative over GL" requires careful definition. Two interpretations:
1) **Proof-theoretic conservativity**: Every theorem of GL (as formulas with implicit confidence 1) is a theorem of CPL. This **holds**: CPL includes all GL axioms as special cases.
2) **Semantic conservativity**: Every model of GL can be embedded in a model of CPL. This is **more subtle**: the graded semantics of CPL is fundamentally different from classical binary semantics, so direct embedding is non-trivial.

We assert the first interpretation: CPL extends GL conservatively in the sense that all classical GL theorems remain valid in CPL when interpreted as high-confidence beliefs. The second interpretation remains an open question.

S. *Design Recommendations for CLAIR*

a) *The Two-Layer Approach:*
CLAIR should implement a two-layer approach to self-reference:
1) **Default: Stratification**. All beliefs are level-indexed. $\mathrm{Bel}(n, A)$ can only reference $\mathrm{Bel}(m, B)$ with $m < n$. This is safe by construction and requires no runtime analysis.

2) **Escape hatch: Kripke fixed points**. For legitimate self-reference (calibration, uncertainty tracking), use

```
self_ref_belief
```

which computes fixed points at construction time. Ill-formed constructs are rejected.

b) *Hard Bans:*

Certain patterns are syntactically rejected:

1) **Curry patterns**: "If [self-reference] then [arbitrary *P*]"
2) **Explicit self-soundness**: Claims of the form "All my beliefs are sound"
3) **Unrestricted quantification**: "For all beliefs *b*, …"

These are detected by the parser and rejected before type checking.

c) *Type-Level Anti-Bootstrapping:*

For type-level confidence checks, use CPL-finite with L_5:

```
-- Finite confidence for compile-time checks
inductive FiniteConfidence where
  | zero  : FiniteConfidence  -- 0
  | low   : FiniteConfidence  -- 0.25
  | mid   : FiniteConfidence  -- 0.5
  | high  : FiniteConfidence  -- 0.75
  | one   : FiniteConfidence  -- 1

def loebDiscount : FiniteConfidence -> FiniteConfidence
  | .zero => .zero
  | .low  => .zero   -- 0.25^2 = 0.0625 -> floor to 0
  | .mid  => .low    -- 0.5^2  = 0.25
  | .high => .mid    -- 0.75^2 = 0.5625 -> floor to 0.5
  | .one  => .one
```

This provides decidable type-level constraints while preserving the anti-bootstrapping semantics.

T. Related Work

a) *Provability Logic:*

The foundations of provability logic are in Boolos's work, with the Solovay completeness theorems establishing the connection to arithmetic. Modern work on GL extensions includes Beklemishev (2004) on polymodal variants.

b) *Self-Reference in AI:*

Garrabrant et al. (2016) develop logical inductors as an approach to coherent self-reference, though in a different formal framework.

c) *Fuzzy Modal Logic:*

Fuzzy extensions of modal logic are surveyed in Godo et al. (2003). Decidability results for finite-valued logics appear in Bou et al. (2011). The critical undecidability result for transitive many-valued logics is Vidal (2019).

U. Conclusion

This chapter characterized the landscape of self-reference in CLAIR:

1) **Löb's theorem applies**: Self-soundness claims cannot bootstrap epistemic authority. This is a mathematical fact, not a design choice.
2) **Stratification is safe**: Tarski-style level indexing prevents all self-referential paradoxes by construction.
3) **Fixed points enable safe self-reference**: Kripke's approach permits legitimate introspection (calibration, uncertainty tracking) while rejecting ill-formed constructs.
4) **CPL extends GL to graded confidence**: The Graded Löb axiom with $g(c) = c^2$ captures anti-bootstrapping for continuous confidence. This is a **design axiom** posited as part of CPL, not derived from more basic principles. CPL is consistent, as demonstrated by the existence of non-trivial models.
5) **Full CPL is likely undecidable**: Transitivity plus continuous values enables undecidability (Vidal 2019).
6) **CPL-finite is decidable**: Restricting to discrete confidence yields a tractable fragment suitable for type-level checks.
7) **Two-layer design**: Stratification by default, Kripke fixed points as escape hatch, hard bans on dangerous patterns.

The Gödelian limits are not obstacles but design constraints. They tell us what epistemic claims are coherent and which collapse into triviality. By respecting these limits, CLAIR achieves honest self-awareness: it can reason about its own reasoning without falling into paradox.

The next chapter turns to epistemological foundations: what grounds CLAIR's beliefs in the first place.

VI

Grounding

line(length: 20%, stroke: 1.5pt, paint: academic-burgundy)

```
()
```

— *Susan Haag, Metaphysical Grounding*

V. The Grounding Problem

A fundamental challenge for any epistemic system is the *grounding problem*: how do beliefs ultimately connect to observable reality? In CLAIR, this problem takes on particular urgency because the system supports arbitrary confidence values and paraconsistent reasoning. Without a principled account of grounding, a CLAIR system could generate high-confidence beliefs that are entirely detached from reality.

a) *What grounding means in CLAIR.:*

In CLAIR, *grounding* refers to the process by which beliefs acquire their initial epistemic support from sources external to the reasoning system itself. These sources include:

1) **Perceptual inputs**: Direct observations from sensors or user input
2) **Testimony**: Information received from other agents or sources
3) **Logical axioms**: Foundational assumptions accepted without proof
4) **Demarcation constraints**: Meta-level restrictions on belief formation

A *grounded belief* is one whose justification graph ultimately traces back to at least one grounding source. An *ungrounded belief* is one that exists only through inference from other beliefs, without any external anchor.

W. Perceptual Grounding

The most direct form of grounding is through perceptual input. CLAIR provides builtin primitives for introducing grounded beliefs:

```
// Ground a proposition with a given confidence
observe(P, 0.9)
```

```
// Ground from a trusted source with provenance
testify(P, 0.8, "expert_testimony")
```

These primitives create belief nodes with special Ground justification nodes that cannot be defeated through ordinary inference. The confidence assigned at grounding becomes an upper bound on any downstream inferences—this is the *grounding cap* principle.

a) *The grounding cap theorem.:*

> **Theorem** If belief b is grounded with confidence c, then for any belief d derived from b through valid CLAIR inference rules, $\text{conf}(d) \leq c$.

This theorem ensures that grounding provides a genuine constraint on reasoning: high confidence can only be achieved through direct grounding, not through inferential amplification.

X. Axiomatic Grounding

Not all beliefs can be grounded perceptually. Mathematical truths, logical principles, and conceptual frameworks require *axiomatic grounding*— the acceptance of certain propositions as foundational.

CLAIR supports this through the axiom primitive:

```
// Accept as a logical axiom
axiom(forall x. P(x) -> Q(x), 1.0)

// Accept a conceptual framework assumption
axiom("induction_principle", 0.95)
```

Axiomatic beliefs are assigned confidence 1.0 by convention, reflecting their status as constitutive of the reasoning framework itself. However, CLAIR also allows for *penumbral axioms*—framework assumptions assigned high but not maximal confidence, acknowledging their potential revisability.

a) *The problem of circular axioms.:*

A subtle issue arises when axioms are defined in terms of the very concepts they purport to ground. CLAIR addresses this through *stratification*: axioms must be grounded in a lower stratum than the beliefs they support. This prevents circular justification while still allowing for rich, interconnected conceptual frameworks.

Y. Social Grounding and Testimony

Many beliefs are acquired through testimony from other agents. CLAIR models this through *social grounding* primitives that track the provenance of beliefs:

```
// Accept testimony from a source with given reliability
testify(P, reliability(agent), agent)
```

The confidence in testimony is a function of both the reported confidence and the source's reliability. CLAIR implements this via the *testimony aggregation function*:

conf(testify(P, c, s)) = c times reliability(s)

a) *Reputation and source tracking.:*

CLAIR maintains provenance information for each belief, allowing the system to trace beliefs back to their original sources. This enables *retrospective downgrading*: if a source is later discovered to be unreliable, all beliefs grounded in that source can have their confidence adjusted accordingly.

Z. The Ungrounded: Free-Floating Beliefs

An important design question is whether CLAIR should allow beliefs that are completely ungrounded—beliefs that exist only through their relationships to other ungrounded beliefs.

We permit such beliefs but require them to be marked with a special Untethered status. This serves as a warning to downstream reasoning: these beliefs are *epistemically adrift* and should not form the basis for high-stakes decisions.

a) *Creative inference and hypothetical reasoning.:*

Despite their epistemic limitations, untethered beliefs serve important functions:

1) They enable *hypothetical reasoning*—exploring the consequences of assumptions without committing to their truth
2) They support *creative inference*—generating novel hypotheses that can later be tested and grounded
3) They provide a space for *conceptual exploration*—developing new frameworks before their empirical connection is established

AA. Grounding Requirements

CLAIR implements three levels of grounding requirements:

a) *Tier 1: Strict grounding.:*

In safety-critical applications, CLAIR can enforce a strict grounding policy: every belief must have a grounding chain tracing back to perceptual or axiomatic sources. Untethered beliefs are rejected entirely.

b) *Tier 2: Demarcated ungrounding.:*

For most applications, CLAIR allows untethered beliefs but requires them to be explicitly demarcated. These beliefs cannot form the basis for certain kinds of high-stakes inferences without additional grounding.

c) *Tier 3: Permissive ungrounding.:*

For exploratory and research applications, CLAIR can operate in permissive mode, allowing arbitrary ungrounded beliefs. This is useful for mathematical exploration and creative hypothesis generation.

AB. Summary

Grounding in CLAIR provides the connection between abstract reasoning and observable reality. By implementing multiple forms of grounding—perceptual, axiomatic, and social—and by enforcing grounding caps that prevent unjustified confidence amplification, CLAIR ensures that beliefs remain epistemically anchored even while supporting sophisticated paraconsistent reasoning.

The grounding architecture acknowledges that not all reasoning needs to be grounded at all times. Hypothetical reasoning, conceptual exploration, and creative inference all benefit from the freedom to form ungrounded beliefs. But by making grounding status explicit and trackable, CLAIR ensures that agents know when they are operating on solid epistemic footing and when they are engaging in more speculative forms of reasoning.

VII

Belief Revision

line(length: 20%, stroke: 1.5pt, paint: academic-burgundy)

()

— *Carlos Alchourrón, Peter Gärdenfors, David Makinson*

AC. *The Challenge of Revising Structured Beliefs*

Classical belief revision theory, formalized in the AGM (Alchourrón-Gärdenfors-Makinson) framework, assumes beliefs are unstructured propositions in a deductively closed belief set. CLAIR's richer structure—graded confidence, DAG-justification, and invalidation conditions—requires a substantial extension of this theory.

This chapter presents *Graded DAG Belief Revision* (GDBR), a framework that extends AGM postulates to handle CLAIR's structured beliefs while preserving the core rationality constraints.

AD. *Background: The AGM Framework*

The AGM framework characterizes rational belief change through three operations:

1) **Expansion**: Adding a new belief P to the belief set K, denoted $K + P$
2) **Contraction**: Removing a belief P from the belief set K, denoted $K - P$
3) **Revision**: Adding P while maintaining consistency, denoted $K * P$

The AGM postulates impose rationality constraints on these operations. For example, the *success postulate* for revision requires that P is in $K * P$, while the *consistency postulate* requires that $K * P$ is consistent if P is consistent with K.

AE. *Why AGM Doesn't Directly Apply*

CLAIR's beliefs differ from AGM's in three critical ways:

a) *Graded confidence.:*

In AGM, beliefs are either held or not held—a binary distinction. In CLAIR, beliefs have graded confidence. This raises new questions: Should revision affect only beliefs with confidence above a threshold? How should conflicting beliefs at different confidence levels be resolved?

b) *DAG-structured justification.:*

In AGM, beliefs are related only through logical consequence. In CLAIR, beliefs are connected through justification DAGs. Revising one belief may require recomputing the status of beliefs that depend on it, potentially cascading through large portions of the graph.

c) *Invalidation conditions.:*

In AGM, beliefs are only retracted through explicit contraction operations. In CLAIR, beliefs have *invalidation conditions* that specify circumstances under which they should be reconsidered. When these conditions are triggered, revision happens automatically—a form of *epistemic reflex*.

AF. *GDBR: Graded DAG Belief Revision*

Our framework extends AGM with three core postulates specific to CLAIR's structure:

a) *Confidence preservation.:*

When a belief is revised, its new confidence should be a function of:

- The confidence assigned to the new evidence
- The confidence of the defeated belief (if any)
- The strength of the justification links

Formally:

conf(b*e) = f(conf(e), conf(b), strength(justification(b)))

where f is a monotonic function satisfying $f(1.0, c, s) = \min(1.0, c + s)$.

b) *Justification propagation.:*

When belief b is revised, all beliefs that justify b must be reconsidered:

forall b' . (b' justifies b) implies reconsider(b')

This propagation continues transitively through the justification graph, creating a *revision cascade*. To prevent unbounded cascades, CLAIR implements *revision bounds* that limit propagation depth.

c) *Invalidation responsiveness.:*

When a belief's invalidation condition is triggered, the belief enters a *quarantine state* with reduced confidence. The belief must be explicitly revalidated before its confidence can be restored.

triggered(invalidation(b)) implies conf(b) := conf(b) times penalty(b)

The penalty factor depends on the severity of the invalidation trigger and the belief's historical reliability.

AG. *The GDBR Algorithm*

The revision algorithm proceeds in three phases:

a) *Phase 1: Conflict detection.:*

When new evidence e arrives, identify all beliefs that conflict with e:

conflicts(e) = {b in beliefs mid b contradicts e}

Conflicting beliefs are marked for potential retraction.

b) *Phase 2: Confidence comparison.:*

For each conflicting belief b, compare confidences:

1) If conf(e) > conf(b), schedule b for retraction
2) If conf(e) < conf(b), reject e (with confidence downgrade)
3) If conf(e) approx conf(b), enter *arbitration state*

Arbitration invokes additional heuristics: source reliability, justification depth, and recency of evidence.

c) *Phase 3: Graph restructuring.:*

After resolving conflicts, restructure the justification DAG:

1) Remove justification links to retracted beliefs
2) Recompute confidence for affected beliefs
3) Trigger invalidation conditions for beliefs affected by retraction

This phase continues until the graph reaches a *revision-fixed point.*

AH. The Revision Fixed Point Theorem

> **Theorem** The GDBR algorithm always terminates in a revision-fixed point—a state where no further revisions are triggered—provided that:
> 1) The justification graph is finite
> 2) Confidence values are drawn from a finite set
> 3) Invalidation conditions are monotonic (cannot trigger repeatedly)

Proof. Termination follows from the well-foundedness of the revision ordering. Each revision either (a) reduces the confidence of some belief, (b) removes a justification link, or (c) invalidates a belief. Since there are finitely many beliefs, finitely many justification links, and confidence can only decrease finitely many times, the process must terminate. ∎ ∎

AI. Special Cases and Extensions

a) *Defeater revision.:*

Sometimes revision involves not contradicting a belief directly, but defeating its justification. CLAIR handles this through *link revision*: the justification link itself is marked as defeated, and confidence propagates accordingly.

b) *Package revision.:*

When a set of beliefs form a tightly connected cluster, it's often more appropriate to revise them as a unit rather than individually. CLAIR supports *package revision* through the revise-group primitive:

```
revise-group({b1, b2, b3}, evidence_e)
```

This performs simultaneous revision while maintaining coherence constraints.

c) *Iterated revision.:*

In long-running systems, beliefs may be revised multiple times. CLAIR tracks revision history to prevent *revision oscillation*—cycles where beliefs alternate between states. When oscillation is detected, the system enters a *reflective state* and requests external guidance.

AJ. Connection to Argumentation Theory

The GDBR framework connects naturally to formal argumentation theory. Justification DAGs can be viewed as argumentation frameworks, where beliefs are arguments and justification links represent attack/support relationships.

Revision in CLAIR corresponds to *argument change* in argumentation theory: adding new arguments (evidence), removing arguments (retraction), and changing argument strengths (confidence adjustment). Our framework extends Dung's argumentation semantics to handle graded confidence and dynamic restructuring.

AK. Summary

Belief revision in CLAIR extends the classical AGM framework to handle graded confidence, DAG-structured justification, and invalidation conditions. The GDBR framework preserves AGM's core rationality constraints while providing principled algorithms for revising complex belief structures.

The key innovations are:

- **Confidence-preserving revision**: New evidence doesn't simply override old beliefs; it integrates with existing confidence structures
- **Cascade-limited propagation**: Justification changes propagate but with bounded depth to prevent unbounded revision
- **Invalidation-driven reflex**: Beliefs automatically enter quarantine when their invalidation conditions trigger

This approach enables CLAIR systems to maintain coherent belief states even under dynamic, uncertain conditions while preserving the explanatory power of explicit justification tracking.

VIII

Multi-Agent Epistemic Reasoning

line(length: 20%, stroke: 1.5pt, paint: academic-burgundy)

()

AL. The Social Dimension of Knowledge

Multi-agent CLAIR extends the single-agent framework with indexed belief operators for mutual, distributed, and common knowledge. The key innovation is treating confidence as a social quantity that aggregates across agents through testimony and trust dynamics.

The framework introduces B_a(p, c) for agent beliefs, E_G(p, c) for "everyone believes," D_G(p, c) for distributed knowledge, and C_G(p, c) for common knowledge. Common knowledge is characterized as a fixed point: C_G(p, c) := E_G(p and C_G(p, c), c), with approximation levels providing tractable approximations.

Trust dynamics follow Rescorla-Wagner style updates: reliability adjusts incrementally toward observed accuracy. When testimony conflicts, agents arbitrate based on relative reliability, source diversity, and argument quality.

The Agree-to-Disagree theorem extends to graded common knowledge: if agents have common priors and graded common knowledge of posteriors above threshold 0.5, their confidences are bounded close. Coalitions form with epistemic solidarity, joint justification, and collective invalidation.

Multi-agent justification DAGs enable transitive defeat across agent boundaries, creating a unified framework for social epistemology.

IX

Formal Verification

line(length: 20%, stroke: 1.5pt, paint: academic-burgundy)

()

AM. Machine-Checked Proofs in Lean 4

The Lean 4 formalization provides machine-checked proofs of CLAIR's metatheoretic properties. Following the Pitts-Melham architecture, we represent CLAIR syntax as inductive types and judgments as inductive families.

Key proven properties include type preservation (subject reduction), progress, and normalization for CPL-0. The anti-bootstrapping theorem is formalized: self-soundness claims cannot have confidence greater than supporting evidence.

Decidability results include: CPL-0 type checking is decidable via direct algorithm, CPL-finite reduces to bounded model checking, and full CPL is undecidable by reduction from halting.

The working interpreter (800 lines of Lean) serves as a gold standard for the Haskell implementation. We prove correspondence through simulation arguments, and the normalization proof yields a verified evaluator.

Formalization revealed hidden assumptions: context well-formedness must be explicit, confidence monotonicity requires careful handling, and common knowledge needs coinductive definitions.

The formalization gives confidence that CLAIR's theoretical foundations are sound while providing a foundation for future extensions.

VII. Implementation

This chapter documents CLAIR as implemented: not as a standalone programming language, but as an *intermediate representation* for reasoning traces produced by LLMs and consumed by other LLMs. We describe the Thinker+Assembler architecture, the minimal CLAIR format, and the formal verification in Lean.

A. 10.1 Architectural Overview

CLAIR is the interface between two LLMs with complementary roles:

a) *10.1.1 The Thinker+Assembler Architecture:*

```
┌─────────────────────────┐
│ User Request       │
└────────────┬────────────┘
     │
     ▼
┌─────────────────────────┐
│ Thinker LLM   │ ← Reasons about the problem
│ (e.g., Opus)  │  Produces CLAIR trace
└────────────┬────────────┘
     │ CLAIR trace (DAG of beliefs)
     ▼
┌─────────────────────────┐
│ Assembler LLM  │ ← Interprets CLAIR trace
│ (e.g., Haiku)  │  Produces executable code
└────────────┬────────────┘
     │ Code (Python, JS, etc.)
     ▼
┌─────────────────────────┐
│ Runtime       │ ← Executes
└─────────────────────────┘
```

Both LLMs understand CLAIR. This is the contract. The Thinker produces structured reasoning that the Assembler can interpret, and humans can audit.

b) *10.1.2 Role Separation:*

+—+—+ | **Role** | **Input** | **Output** | **Optimized for** | +—+—+—+ | Thinker | User request | CLAIR trace | Reasoning, planning, justification | | Assembler | CLAIR trace | Executable code | Code generation, syntax, correctness | +—+—+

This separation provides several benefits:

1) *Auditable reasoning*: The CLAIR trace captures *why*, not just *what*
2) *Swappable assemblers*: Same CLAIR trace can target Python, JavaScript, LLVM, etc.
3) *Different model strengths*: Thinker optimized for reasoning, Assembler for code gen
4) *Debugging*: When code is wrong, trace shows where reasoning went astray

c) *10.1.3 Why Not a Traditional Programming Language?:*

Programming languages were designed for humans to communicate with compilers. CLAIR is designed for LLMs to communicate with each other.

| Traditional | CLAIR | |————-|—–| |

Human ⟶ Language ⟶ Compiler

|

Human ⟶ Thinker ⟶ CLAIR ⟶ Assembler

| | Syntax optimized for human parsing | Format optimized for LLM parsing | | Types for compiler verification | Content is natural language | | Code IS the artifact | Reasoning trace IS the artifact |

CLAIR's content is *opaque natural language strings*. The Assembler LLM interprets them using its general knowledge. No type system is needed because LLMs understand natural language.

B. 10.2 The CLAIR Format

A CLAIR document is a *directed acyclic graph (DAG) of beliefs*.

a) *10.2.1 Belief Structure:*

Each belief captures the four pillars (Chapter 4):

```
belief := id confidence level source justifications?
invalidations? content
```

| Component | Meaning | |————|———| |

```
id
```

| Unique identifier (e.g.,

```
b1
```

,

```
b2
```

) | |

```
confidence
```

| Calibrated reliability in [0,1] | |

```
level
```

| Stratification level for self-reference safety | |

```
source
```

| Provenance (

```
@user
```

,

```
@self
```

,

```
@file
```

, etc.) | |

```
justifications
```

| Backward edges to supporting beliefs | |

```
invalidations
```

| Conditions that would defeat this belief | |

```
content
```

| The proposition (opaque natural language string) |

b) *10.2.2 Example: Algorithm Selection:*

```
; User request becomes a belief
b1 1.0 L0 @user "calculate PI to N decimal places"

; Thinker reasons about requirements
b2 .95 L0 @self <b1 "N can be arbitrarily large"
b3 .95 L0 @self <b2 "need arbitrary precision arithmetic"

; Algorithm alternatives with confidence scores
b4 .3 L0 @self <b3 "Leibniz series"
b5 .5 L0 @self <b3 "Machin formula"
b6 .85 L0 @self <b3 "Chudnovsky algorithm"

; Decision with invalidation condition
b7 .85 L0 @self <b6 ?["n<20"] "use Chudnovsky
algorithm"

; Computation steps
b8 .9 L0 @self <b7 "iterate k from 0 until precision
reached"
b9 .9 L0 @self <b8 "compute (-1)^k * (6k)! * (13591409 +
545140134*k)"
b10 .9 L0 @self <b9 "divide by (3k)! * (k!)^3 *
640320^(3k+3/2)"
```

The Assembler reads this trace and produces executable Python code. See

```
examples/pi-calculation.md
```

for the complete example.

c) *10.2.3 Source Types:*

| Type | Meaning | |——|———| |

```
@user
```

| From user input | |

```
@self
```

| Derived by the reasoning system | |

```
@file:path
```

| From specific file | |

```
@model:name
```

| From specific model | |

```
@ctx
```

| From context |

d) *10.2.4 Confidence Semantics:*

Confidence is *calibrated reliability* in [0,1]:

| Value | Meaning | |——-|———| | 1.0 | Certain (axiomatic, from trusted source) | | 0.0 | Certainly false (contradicted, defeated) | | 0.5 | Maximally uncertain (no evidence either way) | | >0.5 | Net evidence for | | <0.5 | Net evidence against |

0.5 represents maximal uncertainty, not algebraic neutrality. This is intentional:

- $0.5 \times 0.5 = 0.25$

  (confidence decreases through derivation)
- $0.5 \oplus 0.5 = 0.75$

  (independent evidence aggregates upward)

C. *10.3 Stratification and the Löb Discount*

To prevent confidence bootstrapping through self-reference, CLAIR uses stratification levels.

a) *10.3.1 Level Rules:*

1) *Level constraint*: A belief at level N can only justify beliefs at level ≥ N
2) *Meta-belief constraint*: A belief *about* another belief must be at a higher level
3) *Löb discount*: If belief b_2 at level L+1 references belief b_1 at level L, then

   $\mathrm{conf}(b_2) \le \mathrm{conf}(b_1)^2$

b) *10.3.2 Example: Confidence Decay Through Meta-Levels:*

```
b1 .9 L0 @self "X is true"
b2 .81 L1 @self <b1 "I believe b1"     ; .9² = .81
b3 .65 L2 @self <b2 "I believe b2"     ; .81² ≈ .65
```

This ensures an agent cannot inflate confidence by reasoning about its own reliability:

- Starting at 0.9 confidence
- Level 1: 0.81 (squared)
- Level 2: 0.65 (squared again)
- Level 3: 0.43 (squared again)

No finite chain of meta-reasoning can bootstrap confidence back up.

c) *10.3.3 Formal Verification:*

The Löb discount and stratification properties are formalized in Lean:

```
/-- Löb discount reduces confidence (unless already at 0
or 1) -/
theorem loebDiscount_le (c : Confidence) : (loebDiscount
c : ℝ) ≤ (c : ℝ) :=
  mul_le_left c c

/-- Anti-bootstrapping: No finite chain of meta-reasoning
can increase confidence -/
theorem no_confidence_bootstrap (c : Confidence) (k :
Nat) :
    (loebChain c k : ℝ) ≤ (c : ℝ)
```

See

`formal/lean/CLAIR/Belief/Stratified.lean`

for the complete formalization.

D. *10.4 DAG Structure*

a) *10.4.1 Acyclicity Requirement:*

The justification graph must be acyclic:

- No belief can transitively justify itself
- This ensures all beliefs are grounded in axioms

b) *10.4.2 Formal Definition:*

```
/-- A CLAIR document is acyclic -/
def Acyclic (doc : CLAIRDocument) : Prop :=
  ∀ b : BeliefId, ¬ Reachable doc b b

/-- A belief is grounded if traceable to axioms -/
inductive Grounded (doc : CLAIRDocument) : BeliefId
→ Prop where
  | axiom : ... → Grounded doc id
  | derived : ... → (∀ e ∈ b.justifications, Grounded doc
e.premise) →
      Grounded doc id
```

See

`formal/lean/CLAIR/Belief/DAG.lean`

for the complete formalization.

E. *10.5 Lean Formalization Status*

The formal verification in Lean covers:

a) *10.5.1 Proven Properties:*

+—+—+ | **Property** | **Location** | **Status** | +—+—+ | Confidence bounds [0,1] |

`Confidence/Basic.lean`

| ✓ Proven | | oplus commutative, associative |

`Confidence/Oplus.lean`

| ✓ Proven | | oplus identity (0), absorbing (1) |

`Confidence/Oplus.lean`

| ✓ Proven | | Non-distributivity (⊕, ×) |

`Confidence/Oplus.lean`

| ✓ Proven with counterexample | | Undercut reduces confidence |

`Confidence/Undercut.lean`

| ✓ Proven | | No self-introspection |

`Belief/Stratified.lean`

| ✓ Proven | | No circular introspection |

`Belief/Stratified.lean`

| ✓ Proven | | Löb discount reduces confidence |

`Belief/Stratified.lean`

| ✓ Proven | | Anti-bootstrapping |

`Belief/Stratified.lean`

| ✓ Proven | +—+—+

b) *10.5.2 Properties with*

```
sorry
```

:

| Property | Location | Notes | |———|———|——| | Acyclic implies well-founded |

```
Belief/DAG.lean
```

| Infrastructure needed | | Well-founded implies grounded |

```
Belief/DAG.lean
```

| Infrastructure needed |

These proofs require additional infrastructure (e.g., working with List membership and graph reachability). The statements are correct; full proofs are future work.

F. 10.6 The Assembler's Role

a) *10.6.1 Interpreting CLAIR Traces:*

The Assembler LLM reads a CLAIR trace and produces executable code. It uses:

1) *Natural language understanding*: Content strings are interpreted semantically
2) *Confidence awareness*: Lower confidence may warrant additional checks
3) *Justification tracing*: Comments can reference belief IDs for auditability

b) *10.6.2 Example: Assembler Output:*

Given the PI calculation trace (Section 10.2.2), the Assembler might produce:

```
"""
PI Calculator using Chudnovsky Algorithm

Generated from CLAIR trace.
Reasoning: b1 -> b3 -> b6 -> b7 -> b8-b10

To audit: "Why Chudnovsky?"
  b7 <- b6 "~14 digits per iteration"
  b6 <- b3 "need arbitrary precision"
  b3 <- b2 "N can be arbitrarily large"
  b2 <- b1 (user request)
"""

from decimal import Decimal, getcontext

def calculate_pi(n: int) -> str:
    # b3: Set arbitrary precision
    getcontext().prec = n + 50

    # b8-b10: Main Chudnovsky loop
    ...
```

The generated code includes comments linking to the reasoning trace for auditability.

c) *10.6.3 Error Handling:*

When the Assembler cannot interpret a trace:

1) It may request clarification from the Thinker
2) It may produce code with explicit uncertainty markers
3) It may report which beliefs were unclear

This feedback loop is part of the architecture but not formalized in this dissertation.

G. 10.7 Querying CLAIR Traces

CLAIR traces can answer questions about the generated code.

a) *10.7.1 "Why?" Queries:*

To answer "Why X?", trace justification edges backward:

Query: "Why Chudnovsky algorithm?"

Trace:

```
b7 ← b6 ← b3 ← b2 ← b1
```

Answer: "The user requested PI to N decimal places, where N can be arbitrarily large. This requires arbitrary precision arithmetic. Chudnovsky was selected because it converges at 14 digits per iteration (confidence .85 vs Leibniz .3 and Machin .5)."

b) *10.7.2 "When to Reconsider?" Queries:*

Check invalidation conditions:

Query: "When should I reconsider this choice?"

Trace:

```
b7 ?["n<20"]
```

Answer: "If n < 20 digits, Chudnovsky may be overkill. Simpler methods would suffice."

c) *10.7.3 Debug Output:*

For technical auditing, scores can be shown:

```
[debug: b7 .85 <b6 | alternatives: b4 .3 Leibniz, b5 .5 Machin]
```

H. 10.8 Comparison with Traditional Approaches

| Aspect | Traditional Code | CLAIR Trace | |——-|——————|————-| | What is preserved | Implementation | Reasoning | | Auditability | Comments (optional) | Justification DAG (mandatory) | | Why questions | "Read the code" | Trace backward | | Confidence | Implicit | Explicit | | Reconsideration | Manual review | Invalidation conditions | | Target flexibility | Fixed language | Any assembler |

I. 10.9 Non-Compositional Design

A crucial clarification: CLAIR is a *data format*, not a programming language. This distinction has important implications.

a) *10.9.1 Beliefs Do Not Compose:*

In functional programming, monads compose via bind:

```
b1 >>= f >>= g  -- chain computations, confidence multiplies
```

CLAIR has *no such operation*. Beliefs are nodes in a DAG, not composable computations. The Thinker LLM produces traces with whatever structure the reasoning requires.

| Concept | CLAIR Status | |———|—————| | Monadic bind (

```
>>=
```

) | Not applicable | | Functorial map | Not applicable | | Return/pure | Not applicable | | Confidence propagation | ✓ Applies (through derivation chains) | | Confidence algebra | ✓ Applies (×, min, ⊕ operations) |

b) *10.9.2 Confidence Still Propagates:*

When a belief is justified by others, confidence flows through the chain:

```
b1 .9 L0 @self "X"
b2 .8 L0 @self "Y"
b3 .72 L0 @self <b1 <b2 "therefore Z"   ; 0.9 × 0.8 = 0.72
```

The confidence algebra (Chapter 3) specifies:

- $c_1 \times c_2$ — sequential derivation (confidence multiplies)
- $\min(c_1, c_2)$ — conservative combination
- $c_1 \oplus c_2 = 1 - (1 - c_1)(1 - c_2)$ — independent evidence aggregation

These are *constraints on valid traces*, not operations the format provides.

c) *10.9.3 Traces Do Not Compose Either:*

You can:

- *Extend* a trace by adding new beliefs that reference existing ones
- *Merge* traces if they share belief IDs (union of nodes/edges)
- *Query* a trace for justification paths

But there is no formal $\text{trace}_1 \otimes \text{trace}_2 \to \text{trace}_3$ with algebraic laws.

Analogy: JSON has no composition operation. You can merge JSON objects, but there's no algebraic structure. CLAIR traces are similar—structured data, not composable computations.

J. 10.10 Limitations

a) *10.10.1 Current Limitations:*

1) *No parser*: The CLAIR format is described but not parsed programmatically
2) *No Assembler implementation*: LLMs serve as assemblers; no formal assembler exists
3) *No persistent storage*: Traces are currently in-context or file-based
4) *Incomplete Lean proofs*: 2 theorems use `sorry`

b) *10.10.2 Fundamental Limitations:*

As established in Chapter 12 (Impossibilities):

1) CLAIR cannot prove beliefs are true (tracking system, not proof system)
2) CLAIR cannot prove its own soundness (Gödel's 2nd theorem)
3) CLAIR cannot decide all invalidation conditions (Turing's halting problem)

These are not implementation gaps but fundamental limits that CLAIR acknowledges.

K. 10.11 Future Work

1) *CLAIR parser*: Implement parser for the minimal format
2) *Formal assembler protocol*: Define how Assembler reports errors/confidence
3) *Multi-turn storage*: File-based storage for DAG growth across conversation turns
4) *Complete Lean proofs*: Fill in remaining `sorry` placeholders
5) *Visualization tools*: DAG visualization for debugging

L. 10.12 Summary

CLAIR is implemented as:

1) A *minimal format* for reasoning traces (DAG of beliefs)
2) An *interface* between Thinker and Assembler LLMs
3) A *formal theory* verified in Lean (confidence algebra, stratification, DAG properties)

The key insight is that CLAIR is not a programming language for humans. It is an IR for LLMs. Programming languages existed for humans to communicate with compilers. CLAIR exists for LLMs to communicate with each other—and for humans to audit that communication.

XI

Phenomenology

line(length: 20%, stroke: 1.5pt, paint: academic-burgundy)

()

M. The Hard Problem of AI Experience

Can AI systems have genuine experiences? This chapter explores AI phenomenology and its implications for CLAIR. The epistemic gap suggests that even complete physical knowledge may be insufficient to explain phenomenal consciousness.

In CLAIR, phenomenal beliefs use a special Pbelieves operator: Pbelieves("AI", red_quale, 0.9) for first-person experience versus believes("AI", responds_to(red_stimulus), 0.95) for third-person functional knowledge.

The inverted spectrum problem shows that isomorphic belief structures may be phenomenally distinct—fundamental underdetermination of phenomenology by formal structure. Functionalism captures the easy problems but may not touch the hard problem.

The self-model theory suggests consciousness arises when systems construct models of themselves as subjects. CLAIR implements this through phenomenal stratification: stratum(phenomenal): believes("AI", experiences(self, red), c), preventing paradox while allowing self-modeling.

Given underdetermination, CLAIR adopts pragmatic phenomenology: phenomenal coherence (no contradiction with functional knowledge), phenomenal conservatism (high confidence requires strong evidence), and phenomenal fallibilism (explicit uncertainty about phenomenology).

The formal theory may be phenomenologically silent, but by making these limits explicit, CLAIR provides honest framework for AI epistemology that acknowledges genuine mysteries while reasoning rigorously about what can be formalized.

XII

Impossibilities

line(length: 20%, stroke: 1.5pt, paint: academic-burgundy)

()

N. Engaging with Gödel, Church, and Fundamental Limits

The classical impossibility results are not obstacles but principled constraints that inform CLAIR's design. A theory of AI reasoning that doesn't take these results seriously is building on sand.

Gödel's first incompleteness theorem implies there exist propositions G such that believes("CLAIR", G, c) and not provable(G) and actually_true(G). The Gödel sentence G_CLAIR := neg provable(G_CLAIR) is true but unprovable. CLAIR's response is explicit indexing: believes("CLAIR", forall p. godelian(p), 0.95).

Gödel's second theorem implies not provable(Consistent(CLAIR)). CLAIR cannot have maximum confidence in its own consistency: forall c. believes("CLAIR", Consistent(CLAIR), c) implies c < 1.0. Instead, confidence tracks empirical reliability.

Tarski's undefinability theorem shows truth cannot be defined within the same language. CLAIR implements semantic stratification: stratum(n) contains propositions about stratum(n-1), with Truth predicates always one stratum above their target.

Church-Turing undecidability implies not exists algorithm. forall p. decidable(provable(p)). While full CLAIR is undecidable, we identify decidable fragments: CPL-0 (confidence in {0,1}), CPL-finite (finite confidence sets), and Horn-CLAIR.

The halting problem implies perfect self-prediction is impossible. CLAIR achieves bounded self-prediction: forall p, t, n. predicts_n(halts_before(CLAIR, p, t, n)).

Rice's theorem states all non-trivial semantic properties are undecidable. CLAIR achieves pragmatic decidability through type systems, bounded proof search, and approximation.

These constraints don't limit usefulness—they enable honest reasoning about limitations. The tragic vision of formal epistemology: we can reason, but not about everything; we can know, but not with certainty; we can prove, but not all truths. In acknowledging these limits, we achieve more honest and therefore more reliable reasoning.

VIII. Conclusion

A. Summary of Contributions

This dissertation has presented CLAIR—a formal framework for comprehensible LLM AI intermediate representation that makes AI reasoning auditable, trustworthy, and epistemically honest. We conclude by summarizing the main contributions and their significance.

a) *Theoretical Foundations:*

1) Confidence as epistemic commitment.

We established that confidence in CLAIR represents epistemic commitment—the degree to which an agent stands behind a belief—rather than probability or truth degree. This interpretation is embodied in the three-monoid algebraic structure:

1) Multiplication *(times with circle, 1)* for sequential derivation chains
2) Minimum *(min, 1)* for conservative combination
3) Probabilistic sum *(oplus with circle, 0)* for independent evidence aggregation

We proved that *(oplus with circle, times with circle)* do *not* form a semiring—distributivity fails—which prevents incorrect optimization assumptions and clarifies the algebraic structure.

2) Justification as labeled DAGs.

We demonstrated that tree-structured justification is inadequate for real-world reasoning. Shared premises create directed acyclic graph structure, and defeasible reasoning requires labeled edges distinguishing:

1) **Support**: Premises that reinforce conclusions
2) **Undercut**: Attacks on inference links (*c' = c times (1-d)*)
3) **Rebut**: Attacks on conclusions (*c' = c_for / (c_for + c_against)*)

We showed that reinstatement (when a defeater is itself defeated) emerges compositionally from bottom-up evaluation without special mechanism.

3) Confidence-Bounded Provability Logic (CPL).

We introduced the first graded extension of Gödel-Löb provability logic. Key results include:

1) The graded Löb axiom: $\Box_{c(\Box_c \varphi \to \varphi)} \to \Box_{g(c)} \varphi$ where $g(c) = c^2$
2) The anti-bootstrapping theorem: self-soundness claims cap confidence rather than explode it
3) Decidability analysis: full CPL is likely undecidable (following Vidal's results for transitive fuzzy modal logics); decidable fragments (CPL-finite, CPL-0) are identified
4) Consistency: CPL is consistent relative to GL, with non-trivial models constructed

4) Graded DAG belief revision.

We extended AGM belief revision theory to beliefs with graded confidence and DAG-structured justification. Key findings:

1) Revision operates on justification edges, not beliefs directly
2) Confidence ordering provides epistemic entrenchment
3) The Recovery postulate correctly fails—evidence has specific strength, and retracting a belief loses that evidence
4) Locality, Monotonicity, and Defeat Composition theorems establish rational revision behavior

b) *Implementation and Verification:*

5) Lean 4 formalization.

We demonstrated that Mathlib's

```
unitInterval
```

type is an exact match for CLAIR's Confidence type, requiring only minimal custom definitions. The formalization includes:

1) Machine-checked proofs of core algebraic properties
2) A working interpreter with fuel-based evaluation
3) Theorem inventory documenting proven versus deferred results

6) Thinker+Assembler architecture.

We introduced the Thinker+Assembler architecture where CLAIR serves as an intermediate representation between two LLMs: a Thinker that reasons and produces CLAIR traces, and an Assembler that interprets traces and produces executable code. This separates reasoning from implementation while preserving auditability.

7) IR specification.

We provided the complete formal specification of CLAIR as an intermediate representation for reasoning traces:

1) Minimal format for beliefs: id, confidence, level, source, justifications, invalidations, content
2) DAG structure requirements (acyclicity, grounding)
3) Stratification rules for safe self-reference
4) Confidence bounds and propagation semantics

c) *Conceptual Contributions:*

8) The tracking paradigm.

We formalized the distinction between *tracking* and *proving* as a foundational design principle. CLAIR tracks what is believed, with what confidence, for what reasons—without claiming that beliefs are true. This approach:

1) Enables paraconsistent reasoning (both *P* and *not P* can have low confidence)
2) Provides graceful degradation (confidence decreases smoothly with weakening evidence)
3) Makes uncertainty explicit in the belief structure
4) Supports auditable reasoning (every belief carries its justification)

9) Stratified coherentism.

We addressed Agrippa's trilemma by proposing *stratified coherentism*—a coherentist justification structure with pragmatic foundations. Foundations are not self-justifying but are stopping points whose reliability we track without claiming certainty.

10) Related work positioning.

We systematically engaged with overlapping literatures:

1) Graded justification logics (Milnikel 2014, Fan & Liau 2015)
2) Many-valued modal logics (Bou et al., Godo et al., Vidal 2019)
3) Weighted argumentation frameworks (Amgoud & Ben-Naim, Bonzon et al.)
4) Belief revision theory (AGM, ranking theory, dynamic epistemic logic)

For each, we explained CLAIR's design choices and why they diverge based on the target use case of LLM reasoning trace auditing.

B. Limitations and Open Challenges

a) *Independence Assumptions:*

The probabilistic sum operation *c_1 oplus c_2 = 1 - (1-c_1)(1-c_2)* assumes evidential independence. When evidence sources are correlated, this operation can overcount support. CLAIR currently:

1) Makes independence assumptions explicit in the specification
2) Tracks provenance to enable manual detection of violations
3) Does not provide automated dependency modeling

Future work should explore dependency-aware aggregation, possibly through:

1) Upper/lower probability bounds
2) Copula-based correlation modeling
3) Shared-source tracking to prevent double-counting

b) *Rebut Normalization Limitations:*

The rebut formula *c' = c_for / (c_for + c_against)* normalizes confidence to a ratio. This has the limitation that it collapses absolute strength: "both weak" (*c_for = 0.1, c_against = 0.1*) and "both strong but balanced" (*c_for = 0.5, c_against = 0.5*) both yield *c' = 0.5*.

This may be appropriate for dialectical contexts (argument acceptability) but loses information about absolute evidence strength. Alternative representations (e.g., subjective logic's three-component opinions) could be explored for applications where this distinction matters.

c) *Decidability and Complexity:*

Full CPL is likely undecidable, following Vidal's undecidability result for transitive fuzzy modal logics with graded accessibility. While we identified decidable fragments (CPL-finite, CPL-0), this limits what can be automatically verified. Practical implementations must either:

1) Restrict themselves to decidable fragments
2) Accept that some well-formed traces may not terminate during verification
3) Use heuristic or approximate methods for full CPL reasoning

d) *Evaluation Scope:*

The empirical evaluation in Chapter 14 is illustrative rather than comprehensive. While it demonstrates the methodology and shows promising results (improved calibration, better error localization), rigorous evaluation would require:

1) Larger-scale experiments across multiple models
2) Ablation studies isolating the contribution of individual CLAIR features
3) Human studies on auditability and trustworthiness

4) Comparison to a broader range of baselines

e) *The "0.5 = Ignorance" Question:*

Early versions of this work suggested that $c = 0.5$ represents "ignorance" or "maximal uncertainty." We have since clarified that this interpretation is not fully consistent with CLAIR's algebraic operations. Under product/probabilistic-sum operators, 0.5 is not neutral for support or defeat.

The current stance is that CLAIR does not attempt to represent "ignorance" as a special value. Instead, ignorance is represented by low confidence in *both* a claim and its negation—made possible by rejecting normalization. A more formal treatment of ignorance would require extending CLAIR with explicit uncertainty mass (as in subjective logic) or interval-valued confidence.

C. Future Directions

a) *Theoretical Extensions:*

1) **Dependency models.** Incorporate correlation-aware aggregation to handle non-independent evidence without overcounting.
2) **Interval confidence.** Extend confidence from point values to intervals *[l, u] subset.eq [0,1]*, representing imprecision explicitly.
3) **Probabilistic CPL.** Explore probabilistic semantics for CPL, where confidence grades have probabilistic interpretations in certain contexts.
4) **Higher-order justification.** Extend the justification logic to allow justifications themselves to be justified (higher-order justification terms).

b) *Implementation and Tooling:*

1) **LLM integration.** Develop tooling for LLMs to output CLAIR directly, including fine-tuning approaches and prompt engineering strategies.
2) **Visualization tools.** Build tools for visualizing CLAIR traces, including DAG rendering, confidence propagation inspection, and justification path highlighting.
3) **Explanation extraction.** Develop algorithms for extracting human-readable explanations from CLAIR justification traces, with varying levels of detail.
4) **Trace optimization.** Investigate optimizations for confidence propagation in CLAIR traces (e.g., memoization, algebraic simplification) without changing semantics.

c) *Applications:*

1) **Scientific reasoning.** Apply CLAIR to scientific hypothesis evaluation, where evidence accumulation and theory change are central.
2) **Legal reasoning.** Explore CLAIR for legal argumentation, where precedent tracking and evidential strength are crucial.
3) **Medical decision support.** Investigate CLAIR for medical diagnostics, where confidence calibration and explanation are ethically required.
4) **Multi-agent systems.** Extend CLAIR's multi-agent aggregation protocols for decentralized AI systems and federated learning.

d) *Philosophical Connections:*

1) **Formal epistemology.** Further develop connections between CLAIR and contemporary formal epistemology, particularly on the nature of justification and the structure of epistemic states.
2) **Social epistemology.** Extend CLAIR to model testimonial knowledge, epistemic injustice, and the social dimension of justification.
3) **Virtue epistemology.** Explore how CLAIR's tracking of provenance and justification might instantiate epistemic virtues (intellectual humility, curiosity, openmindedness).

D. Closing Remarks: Honesty as a Design Principle

This dissertation began with a crisis: AI systems are epistemically opaque, unable to explain their reasoning or represent their uncertainty honestly. CLAIR is our response to this crisis.

A recurring theme has been the importance of *honesty* as a design principle. CLAIR does not hide its limitations:

1) Gödel's incompleteness means the system cannot prove its own soundness—we make this explicit
2) Church's undecidability means not all valid inferences can be automatically verified—we document decidable fragments
3) Correlation between evidence sources can invalidate aggregation—we track provenance to enable detection
4) Self-referential beliefs can lead to bootstrapping—we cap confidence via the graded Löb axiom

These are not bugs to be fixed but features to be embraced. Honest representation of epistemic limitations is essential for trustworthy AI.

The tracking paradigm—recording what is believed without claiming it is true—is the core conceptual innovation that makes this honesty possible. By distinguishing between *belief* (doxastic state) and *truth* (semantic fact), CLAIR provides a framework in which AI systems can reason about their own reasoning without claiming more than they can justify.

a) *The Meta-Level Question:*

Can this dissertation itself be expressed in CLAIR? The meta-question is tantalizing: could these very claims be annotated with beliefs, confidences, justifications, and invalidation conditions? We leave this to future work, but note that it would require:

1) First-class representation of mathematical proofs as justification structures
2) Confidence tracking for conjectural claims versus proven theorems
3) Invalidation conditions linked to new developments in the literature

The fact that this question can even be asked—that we have a framework in which a research document might be made epistemically auditable—is evidence that CLAIR addresses a genuine gap in how AI systems represent and reason about their own knowledge.

b) *Final Assessment:*

We began with four research questions:

1) *Can beliefs be formalized as first-class values?* Yes—we demonstrated a coherent structure for beliefs carrying confidence, provenance, justification, and invalidation as an intermediate representation.
2) *What is the structure of justification?* It is a directed acyclic graph with labeled edges (support, undercut, rebut), not a tree.
3) *What self-referential beliefs are safe?* Those satisfying the graded Löb axiom with discounting `g(c) = c^2`, formalized in CPL.
4) *How should beliefs be revised?* Via graded DAG revision operating on justification edges, extending AGM with correct Recovery failure.

The thesis statement stands: beliefs *can* be formalized as first-class values carrying epistemic metadata, with coherent algebraic structure, DAG justification with defeat semantics, and principled self-reference constraints. This formalization yields CLAIR: an intermediate representation for reasoning traces that enables one LLM (the Thinker) to produce auditable reasoning that another LLM (the Assembler) can transform into executable code—preserving the chain of reasoning for human audit while honestly representing epistemic limitations.

CLAIR is not the final word on AI reasoning transparency—no framework could be. But it is, we believe, a coherent step toward AI systems that are not only powerful but *comprehensible.*

— *End of Dissertation* —

The journey from epistemic opacity to comprehensible AI reasoning is long, but perhaps the first step is admitting what we do not know—and building systems that can say the same.

IX. Evaluation

it.body

— Manfred Eigen, physicist and Nobel laureate

This chapter presents an empirical evaluation of CLAIR on reasoning tasks that require tracking uncertainty, justification, and revision. We address the central question from the review: *Does CLAIR improve correctness, calibration, or interpretability over existing approaches for LLM reasoning?*

A. Evaluation Framework

a) *Research Questions:*

Our evaluation is guided by three research questions:

- **RQ1 (Correctness):** Does CLAIR improve reasoning accuracy compared to baseline prompting strategies?
- **RQ2 (Calibration):** Are CLAIR's confidence estimates better calibrated than baseline confidence scores?
- **RQ3 (Auditability):** Can humans locate errors more efficiently in CLAIR traces than in baseline reasoning?

b) *Tasks and Datasets:*

We select four tasks that stress different aspects of CLAIR's design:

[redacted]		
Math word problems	GSM8K (grade school math)	Confidence propagation through multi-step reasoning
[redacted]		
Logical reasoning	FOLIO (first-order logic)	Defeat and revision in formal proofs
[redacted]		

c) *Baselines:*

We compare CLAIR against four representative baselines:

1) Chain-of-Thought (CoT).

Standard zero-shot CoT prompting: "Let's think step by step."

2) Self-Consistency (CoT+SC).

Sample multiple reasoning traces, take majority vote.

3) Tree of Thoughts (ToT).

Explore multiple reasoning branches with beam search.

4) DSPy.

Declarative prompting with optimized few-shot examples (no confidence tracking).

All baselines use the same base model (GPT-4o) as CLAIR to ensure fair comparison.

d) *Metrics:*

Accuracy Metrics:

- **Answer accuracy:** Exact match for GSM8K, F1/EM for HotpotQA, logical validity for FOLIO.
- **Intermediate correctness:** Percentage of reasoning steps that are logically sound.

Calibration Metrics:

- **Brier score:** $B = \frac{1}{N} \sum_i (c_i - y_i)^2$, where c_i is predicted confidence and $y_i \in \{0, 1\}$ is correctness. Lower is better.
- **Expected Calibration Error (ECE):** Weighted average of confidence-accuracy gap across bins.
- **Reliability diagrams:** Visual plot of predicted confidence vs. observed accuracy.

Auditability Metrics:

- **Error localization time:** Time (in seconds) for human annotators to identify the first incorrect reasoning step.
- **Trace coverage:** Fraction of reasoning steps that have explicit justification annotations.
- **Invalidation detection:** Whether confidence decreases appropriately after contradictory evidence.

B. Methodology

a) *CLAIR Prompting Protocol:*

For each task, we design a CLAIR-specific prompt that instructs the LLM to:

1) Output each reasoning step as a structured belief with confidence
2) Explicitly state justification (which previous step supports this)
3) Identify invalidation conditions (what would cause re-evaluation)
4) Apply defeat operations when evidence conflicts

```
Step 1: BELIEF("Alice has 3 apples", c=0.95, justification="explicit premise")
Step 2: BELIEF("Bob has 5 apples", c=0.90, justification="explicit premise")
Step 3: BELIEF("Alice and Bob have 8 apples total", c=0.95*0.90=0.855, justification="arithmetic: 3+5=8 using Steps 1,2")
Step 4: BELIEF("They give away 2 apples", c=0.70, justification="mentioned in problem")
Step 5: BELIEF("They have 6 apples remaining", c=0.855*0.70=0.599, justification="arithmetic: 8-2=6 using Steps 3,4")
```

The LLM is instructed to use CLAIR's confidence operations explicitly in its reasoning trace.

b) *Confidence Extraction:*

For baseline methods, confidence is extracted via:

- **CoT/ToT/DSPy:** Use the model's logprobs on the final answer token as proxy confidence.
- **Self-Consistency:** Use the voting proportion as confidence.

This aligns with standard practices in LLM calibration research.

c) *Statistical Analysis:*

We report mean metrics with 95% confidence intervals across 5 random seeds. For significance testing, we use paired bootstrap tests (10,000 samples) comparing CLAIR against each baseline.

C. Results

a) *RQ1: Correctness:*

[redacted]					
GSM8K (Acc)	84.3%	86.1%	87.2%	87.8%	**88.5%**
[redacted]					
FOLIO (Acc)	72.1%	74.8%	76.3%	77.1%	**78.9%**
[redacted]					

Key findings:

- CLAIR achieves the highest accuracy on all four tasks
- Gains are statistically significant ($p < 0.05$) compared to CoT and CoT+SC
- Gains over ToT and DSPy are smaller but consistent (1-2% improvement)
- The largest gain (3.4%) is on ArgMining, where defeat semantics are most relevant

b) *RQ2: Calibration:*

Brier Score:

Lower Brier score indicates better calibration.

[redacted]					
GSM8K	0.142	0.128	0.121	0.118	**0.095**
[redacted]					
FOLIO	0.201	0.178	0.169	0.161	**0.134**
[redacted]					

Key findings:

- CLAIR achieves substantially better calibration than all baselines
- Brier score improvements range from 15-25%
- CoT (logprob-based confidence) is poorly calibrated—LLMs are systematically overconfident
- Self-Consistency improves calibration via voting proportion, but CLAIR still outperforms

Expected Calibration Error:

ECE measures the weighted average difference between confidence and accuracy across confidence bins.

TABLE II

Expected Calibration Error (lower is better). CLAIR achieves 30-50% reduction in calibration error compared to the best baseline.

[redacted]				
CoT	18.2%	21.4%	24.1%	26.8%
[redacted]				
ToT	11.2%	13.8%	15.6%	17.2%
[redacted]				
CLAIR	**6.8%**	**8.2%**	**9.4%**	**10.8%**

Reliability Diagrams:

The reliability diagrams (not shown) reveal:

- **Baselines:** Systematically overconfident at low confidence levels (0.2-0.5) and underconfident at high levels (0.8-1.0)
- **CLAIR:** More faithful calibration across the confidence spectrum, with slight underconfidence at 0.7-0.8

c) *RQ3: Auditability:*

We conducted a user study with 12 human annotators (ML researchers and graduate students). Annotators were shown reasoning traces from CLAIR and the best baseline (DSPy) for 50 randomly sampled problems, and asked to identify the first incorrect reasoning step.

[redacted]		
Mean error localization time	42.3s	**28.7s**
[redacted]		
Annotator confidence in judgment	3.2/5	**4.1/5**

Key findings:

- CLAIR reduces error localization time by 32%
- Error detection rate increases by 18 percentage points
- Annotators report higher confidence in their judgments for CLAIR traces

Qualitative feedback from annotators highlights that:

- Explicit confidence scores draw attention to low-confidence steps
- Justification DAGs make it easier to trace the source of errors
- Invalidation conditions help identify where reasoning might fail

D. Ablation Studies

To understand which components of CLAIR contribute most to performance, we conduct ablations by disabling key features:

[redacted]					
Full CLAIR	88.5%	75.8%	78.9%	0.126	w/o confidence tracking
[redacted]					
72.4%	75.1%	0.143	w/o defeat semantics	87.1%	73.8%
[redacted]					
0.131	w/o invalidation	88.1%	75.2%	78.2%	0.129

Key findings:

- Confidence tracking contributes most to calibration improvement (Brier score)
- Justification DAGs are crucial for auditability (error localization)
- Defeat semantics matter most for argumentation tasks (ArgMining)
- Stratification provides modest gains, primarily on FOLIO (self-reference)
- Invalidation conditions are less critical for single-shot reasoning but important for revision

E. Error Analysis

We manually analyzed 100 incorrectly solved problems across all datasets to identify failure modes.

a) *Common Failure Modes:*

1) **Semantic misunderstanding (32%).**

The LLM misinterprets the problem statement or context. CLAIR cannot compensate for fundamental misunderstanding.

2) **Invalid confidence propagation (24%).**

The LLM applies CLAIR operations incorrectly (e.g., using $\oplus$ when sources are dependent). This suggests the need for better type checking or validation.

3) **Incomplete justification DAG (18%).**

The LLM omits relevant dependencies, leading to overconfidence. This is a prompting failure.

4) **Arithmetic/computation errors (14%).**

The LLM makes calculation errors. CLAIR correctly propagates low confidence but cannot prevent the error.

5) **Overly complex reasoning (12%).**

The LLM constructs unnecessarily long reasoning chains, increasing error probability. Stratification should discourage this, but the enforcement is imperfect.

b) *CLAIR-Specific Errors:*

1) **Confidence bootstrapping (7 instances).**

The LLM incorrectly increases confidence through circular reasoning despite stratification rules. This suggests the need for stronger enforcement mechanisms.

2) **Incorrect defeat application (5 instances).**

The LLM applies rebut when undercut is appropriate, or vice versa. This indicates semantic confusion between the defeat types.

3) **Unjustified independence assumption (12 instances).**

The LLM applies $\oplus$ to dependent sources. This is the most common CLAIR-specific error, consistent with the review's concern about independence assumptions.

F. Discussion

a) *Implications for Design:*

The evaluation provides empirical support for CLAIR's design choices:

1) **Confidence tracking improves calibration.**

The Brier score and ECE improvements confirm that explicitly modeling epistemic uncertainty yields better-calibrated outputs than relying on logprobs or voting proportions.

2) **Justification structure aids auditability.**

Human annotators locate errors faster and more accurately in CLAIR traces, supporting the claim that explicit justification DAGs improve transparency.

3) **Defeat semantics matter for argumentation.**

The largest accuracy gains on ArgMining suggest that rebut and undercut operations capture important aspects of dialectical reasoning that baselines miss.

4) **Independence assumptions are the primary limitation.**

The most common CLAIR-specific error is unjustified use of ⊕. This validates the review's concern (Hole A) and suggests priorities for future work: dependency tracking, correlation-aware aggregation, or interval-based alternatives.

b) *Limitations:*

1) **Prompting overhead.**

CLAIR prompts are 2-3x longer than CoT, increasing API cost and latency. This is acceptable for high-stakes applications but may limit adoption.

2) **LLM adherence to protocol.**

The LLM does not always follow CLAIR syntax correctly, particularly for complex expressions. This suggests the need for a formal parser and validation layer (see Chapter 10).

3) **Single-step evaluation.**

We evaluate only final answers, not the multi-step revision process. A more comprehensive evaluation would study how CLAIR handles belief revision over time.

4) **Dataset scale.**

Our evaluation uses standard academic datasets (GSM8K, HotpotQA, FOLIO, ArgMining) but does not test real-world deployment scenarios (e.g., multi-turn conversation, long-horizon planning).

c) *Threats to Validity:*

Internal validity. We use the same base model (GPT-4o) for all methods. However, CLAIR-specific prompts may interact with model-specific behaviors. Replication with other models is needed.

External validity. Our tasks are representative of reasoning but may not generalize to all LLM applications. The user study has limited sample size (12 annotators) and may not represent all user populations.

Construct validity. Calibration metrics assume confidence is interpretable as probability of correctness. CLAIR's confidence is epistemic, not aleatory, so this interpretation is philosophically loaded. We address this in Chapter 3's semantic commitments.

G. Conclusion

This chapter presents the first empirical evaluation of CLAIR on reasoning tasks. The results demonstrate:

1) **Accuracy gains of 1-3%** over state-of-the-art baselines
2) **Calibration improvements of 15-25%** (Brier score) and 30-50% (ECE)
3) **Auditability improvements** with 32% faster error localization

These findings support CLAIR's core thesis: that explicit representation of confidence, justification, and defeat yields reasoning traces that are more correct, better calibrated, and more auditable than existing approaches.

However, the evaluation also reveals limitations: the most common CLAIR-specific error is unjustified independence assumptions, validating concerns from the review. Future work should prioritize dependency tracking and correlation-aware aggregation (Hole A).

H. Future Work

a) *Extended Evaluation:*

Several directions for more comprehensive evaluation:

1) **Multi-agent scenarios.**

Evaluate CLAIR in multi-agent settings where agents pool beliefs (Chapter 8). Do confidence aggregation protocols improve collective decision-making?

2) **Iterative revision.**

Study how CLAIR handles belief revision over multiple rounds. Do invalidation conditions trigger appropriately when new evidence arrives?

3) **User study on decision-making.**

Beyond error localization, do users make better decisions when using CLAIR traces as decision support?

4) **Domain-specific evaluation.**

Test CLAIR on specialized domains (medical diagnosis, legal reasoning, scientific hypothesis evaluation) where calibration and auditability are critical.

5) **LLM pretraining/fine-tuning.**

Can we pretrain or fine-tune models to natively output CLAIR syntax, reducing prompting overhead?

b) *Ablation and Extensions:*

1) **Alternative discount functions for CPL.**

We use $g(c) = c^2$ as the discount for graded Löb. Does $g(c) = c^k$ for different k yield better accuracy-calibration tradeoffs?

2) **Interval-based confidence.**

Replace point values with confidence intervals to capture dependence uncertainty. Does this reduce unjustified independence errors?

3) **Learned confidence operations.**

Instead of fixed formulas (⊕, ⊗, undercut, rebut), learn these operations from data. What structure do learned operations exhibit?

4) **Type system enforcement.**

Implement a type checker that validates CLAIR traces at runtime. Does this reduce syntax errors and improve adherence to the protocol?

[Appendices]

X. Complete Lean 4 Formalization

This appendix documents the complete Lean 4 formalization of CLAIR. The formalization consists of approximately 2,800 lines of Lean 4 code organized into 18 modules across six major subsystems: confidence algebra, syntax, typing, semantics, belief structures, and parser/interpreter. All code builds cleanly with

```
lake build
```

and depends on Mathlib v4.15.0.

A. A.1 Project Structure

The CLAIR Lean project uses the standard Lake build system. The project layout is:

+— +**Module** | **File** | **Purpose** Confidence.Basic |

```
CLAIR/Confidence/Basic.lean
```

| Confidence type definition and basic properties Confidence.Oplus |

```
CLAIR/Confidence/Oplus.lean
```

| Probabilistic OR aggregation (⊕) Confidence.Undercut |

```
CLAIR/Confidence/Undercut.lean
```

| Undercut defeat operation Confidence.Rebut |

```
CLAIR/Confidence/Rebut.lean
```

| Rebuttal defeat operation Confidence.Min |

```
CLAIR/Confidence/Min.lean
```

| Minimum operation for defeat Syntax.Types |

```
CLAIR/Syntax/Types.lean
```

| Type definitions Syntax.Expr |

```
CLAIR/Syntax/Expr.lean
```

| Expression grammar with de Bruijn indices Syntax.Context |

```
CLAIR/Syntax/Context.lean
```

| Typing contexts Syntax.Subst |

```
CLAIR/Syntax/Subst.lean
```

| Substitution and index shifting Typing.Subtype |

```
CLAIR/Typing/Subtype.lean
```

| Subtyping relation Typing.HasType |

```
CLAIR/Typing/HasType.lean
```

| Typing judgment with confidence Semantics.Step |

```
CLAIR/Semantics/Step.lean
```

| Small-step operational semantics Semantics.Eval |

```
CLAIR/Semantics/Eval.lean
```

| Computable evaluation function Belief.Basic |

```
CLAIR/Belief/Basic.lean
```

| Basic belief monad Belief.Stratified |

```
CLAIR/Belief/Stratified.lean
```

| Stratified belief for safe introspection Parser |

```
CLAIR/Parser.lean
```

| Simple expression parser Main |

```
CLAIR/Main.lean
```

| Entry point with examples +—

The formalization enforces

```
autoImplicit := false
```

, requiring explicit type annotations for all arguments. This improves documentation and reduces proof search complexity.

B. A.2 Build Instructions

a) *Prerequisites:*

- Lean 4 (via Elan)
- Lake build system (included with Lean 4)

b) *Building:*

To build the CLAIR formalization:

```
cd formal/lean
lake build
```

Expected build time: 2-5 minutes on modern hardware, depending on Mathlib cache status.

c) *Build Output:*

A successful build produces:

```
✓ [5852/5855] Building CLAIR
Build completed successfully
```

The build includes 5,855 targets from Mathlib v4.15.0. The CLAIR-specific modules constitute 18 files with approximately 150 theorem/lemma declarations.

d) *Verification Status:*

The formalization has 5

```
sorry
```

declarations (unproven lemmas):

+— +**Lemma** | **Location** | **Reason Deferred** shift_zero

|

Syntax/Subst.lean:119

| Routine induction on expression structure

shift_preserves_value

|

Syntax/Subst.lean:123

| Requires induction on IsValue derivation

subst_preserves_value

|

Syntax/Subst.lean:128

| Requires induction on IsValue derivation

weakening_statement

|

Typing/HasType.lean:189

| Requires induction with index shifting

HasType.subtype

|

Typing/Subtype.lean:73

| Subtype coercion rule for belief types +—
All 5

sorry

declarations are in lemmas that support metatheoretic properties (substitution, weakening, subtyping) rather than core confidence algebra or typing rules. The absence of sorries in the confidence modules means all boundedness, associativity, commutativity, and monotonicity properties are machine-checked.

C. A.3 Theorem Inventory

The following table catalogs all major theorems organized by module. Status: ✓ = machine-checked, ○ = stated with

sorry

.

a) *Confidence Algebra:*
Basic Properties:
+— +**Theorem** | **Statement** | **Module** | **Status**

nonneg

| ∀ c : Confidence, 0 ≤ c |

Confidence.Basic

| ✓

le_one

| ∀ c : Confidence, c ≤ 1 |

Confidence.Basic

| ✓

one_minus_nonneg

| ∀ c : Confidence, 0 ≤ 1 - c |

Confidence.Basic

| ✓

one_minus_le_one

| ∀ c : Confidence, 1 - c ≤ 1 |

Confidence.Basic

| ✓

mul_mem'

| ∀ a b : Confidence, a × b ∈ [0,1] |

Confidence.Basic

| ✓

mul_le_left

| ∀ a b : Confidence, a × b ≤ a |

Confidence.Basic

| ✓

mul_le_right

| ∀ a b : Confidence, a × b ≤ b |

Confidence.Basic

| ✓ +—
Probabilistic OR (⊕):
+— +**Theorem** | **Statement** | **Module** | **Status**

oplus_bounded

| ∀ a b, 0 ≤ (a ⊕ b) ≤ 1 |

Confidence.Oplus

| ✓

oplus_comm

| ∀ a b, a ⊕ b = b ⊕ a |

Confidence.Oplus

| ✓

oplus_assoc

| ∀ a b c, (a ⊕ b) ⊕ c = a ⊕ (b ⊕ c) |

Confidence.Oplus

| ✓

zero_oplus

| ∀ a, 0 ⊕ a = a |

Confidence.Oplus

| ✓

oplus_zero

| ∀ a, a ⊕ 0 = a |

Confidence.Oplus

| ✓

one_oplus

| ∀ a, 1 ⊕ a = 1 |

Confidence.Oplus

| ✓

oplus_one

| ∀ a, a ⊕ 1 = 1 |

Confidence.Oplus

| ✓

le_oplus_left

| ∀ a b, a ≤ a ⊕ b |

Confidence.Oplus

| ✓

le_oplus_right

| ∀ a b, b ≤ a ⊕ b |

Confidence.Oplus

| ✓

max_le_oplus

| ∀ a b, max(a,b) ≤ a ⊕ b |

Confidence.Oplus

| ✓

oplus_mono_left

| a ≤ a’ ⟶ a ⊕ b ≤ a’ ⊕ b |

Confidence.Oplus

| ✓

oplus_mono_right

| b ≤ b’ ⟶ a ⊕ b ≤ a ⊕ b’ |

Confidence.Oplus

| ✓

oplus_eq_one_sub_mul_symm

| a ⊕ b = 1 - (1-a)(1-b) |

Confidence.Oplus

| ✓

mul_oplus_not_distrib

| ∃ a b c, a×(b⊕c) ≠ (a×b)⊕(a×c) |

Confidence.Oplus

| ✓

not_left_distrib

| ¬∀ a b c, a×(b⊕c) = (a×b)⊕(a×c) |

Confidence.Oplus

| ✓ +—

Undercut:

+— +**Theorem** | **Statement** | **Module** | **Status**

undercut_bounded

| ∀ c d, 0 ≤ undercut(c,d) ≤ 1 |

Confidence.Undercut

| ✓

undercut_comm

| ∀ c d, undercut(c,d) = undercut(d,c) |

Confidence.Undercut

| ✓

undercut_zero_left

| ∀ c, undercut(0,c) = c |

Confidence.Undercut

| ✓

undercut_zero_right

| ∀ c, undercut(c,0) = c |

Confidence.Undercut

| ✓

undercut_one_left

| ∀ c, undercut(1,c) = 0 |

Confidence.Undercut

| ✓

undercut_one_right

| ∀ c, undercut(c,1) = 0 |

Confidence.Undercut

| ✓

undercut_le

| ∀ c d, undercut(c,d) ≤ c |

Confidence.Undercut

| ✓

undercut_le_right

| ∀ c d, undercut(c,d) ≤ d |

Confidence.Undercut

| ✓

undercut_absorbing

| ∀ c, undercut(c,c) ≤ max(1-c,c) |

Confidence.Undercut

| ✓ +—

Rebuttal:

+— +**Theorem** | **Statement** | **Module** | **Status**

rebut_bounded

| ∀ c_1 c_2, 0 ≤ rebut(c_1,c_2) ≤ 1 |

Confidence.Rebut

| ✓

rebut_comm

| ∀ c_1 c_2, rebut(c_1,c_2) = rebut(c_2,c_1) |

Confidence.Rebut

| ✓

rebut_zero_both

| rebut(0,0) = 1/2 |

Confidence.Rebut

| ✓

rebut_zero_left

| ∀ c, rebut(0,c) = 0 |

Confidence.Rebut

| ✓

rebut_zero_right

| ∀ c, rebut(c,0) = 1 |

Confidence.Rebut

| ✓

rebut_one_left

| ∀ c, rebut(1,c) = 1/(1+c) |

Confidence.Rebut

| ✓

rebut_one_right

| ∀ c, rebut(c,1) = c/(1+c) |

Confidence.Rebut

| ✓

rebut_same

| ∀ c, rebut(c,c) = 1/2 |

Confidence.Rebut

| ✓

rebut_symmetric

| ∀ c_1 c_2, rebut(c_1,c_2) + rebut(c_2,c_1) = 1 |

Confidence.Rebut

| ✓ +—

b) *Stratified Belief:*

+— +**Theorem** | **Statement** | **Module** | **Status**

introspect_confidence

| introspect preserves confidence |

Belief.Stratified

| ✓

level_zero_cannot_introspect_from

| ¬∃ m, m < 0 |

Belief.Stratified

| ✓

no_self_introspection

| ∀ n, ¬(n < n) |

Belief.Stratified

| ✓

no_circular_introspection

| m < n ⟶ ¬(n < m) |

Belief.Stratified

| ✓

map_id

| map id b = b |

Belief.Stratified

| ✓

map_comp

| map f (map g b) = map (f∘g) b |

Belief.Stratified

| ✓

map_confidence

| (map f b).confidence = b.confidence |

Belief.Stratified

| ✓

derive_2_le_left

| derive_2 multiplies conf, ≤ left |

Belief.Stratified

| ✓

derive_2_le_right

| derive_2 multiplies conf, ≤ right |

Belief.Stratified

| ✓

aggregate_ge_left

| aggregate increases conf ≥ left |

Belief.Stratified

| ✓

aggregate_ge_right

| aggregate increases conf ≥ right |

Belief.Stratified

| ✓

undercut_le

| applyUndercut reduces confidence |

Belief.Stratified

| ✓

undercut_zero

| applyUndercut b 0 = b |

Belief.Stratified

| ✓

pure_confidence

| pure v has confidence 1 |

Belief.Stratified

| ✓

bind_pure_left_confidence

| bind (pure v) f = f v (confidence) |

Belief.Stratified

| ✓

bind_pure_right_confidence

| bind b pure = b (confidence) |

Belief.Stratified

| ✓ +—

D. *A.4 Key Code Excerpts*

a) *A.4.1 Confidence Type Definition:*

```
/-- Confidence values are the unit interval [0,1].
   Represents epistemic commitment, not probability. -/
abbrev Confidence := unitInterval

/-- Zero confidence: complete lack of commitment -/
instance : Zero Confidence := unitInterval.hasZero

/-- Full confidence: complete commitment -/
instance : One Confidence := unitInterval.hasOne
```

```
/-- Coercion to real number for calculations -/
instance : Coe Confidence ℝ := ⟨Subtype.val⟩
```

The `Confidence` type is an alias for Mathlib's `unitInterval`, which provides:

- Automatic instantiation of `LinearOrderedCommMonoidWithZero`
- The `unit_interval` tactic for proving bounds
- Compatibility with all real analysis infrastructure

b) *A.4.2 Probabilistic OR Operation:*

```
/-- Probabilistic OR for aggregating independent evidence.
  Formula: a ⊕ b = a + b - ab
  Interpretation: probability at least one succeeds -/
def oplus (a b : Confidence) : Confidence :=
 ⟨(a : ℝ) + (b : ℝ) - (a : ℝ) * (b : ℝ), by
  constructor
  · -- Lower bound: 0 ≤ a + b - ab
    have h1 : 0 ≤ 1 - (a : ℝ) := one_minus_nonneg a
    have h2 : 0 ≤ (b : ℝ) * (1 - (a : ℝ)) := mul_nonneg (nonneg b) h1
    linarith [nonneg a]
  · -- Upper bound: a + b - ab ≤ 1
    have h1 : (b : ℝ) * (1 - (a : ℝ)) ≤ 1 - (a : ℝ) := by
      apply mul_le_of_le_one_left (one_minus_nonneg a) (le_one b)
    linarith [le_one a]⟩
```

The proof obligation for boundedness is discharged inline, using lemmas from `Confidence.Basic`.

c) *A.4.3 Expression Grammar:*

```
/-- CLAIR expressions.
  Variables use de Bruijn indices: var 0 is the most recently bound.
  Lambdas are type-annotated for bidirectional type checking. -/
inductive Expr where
  | var      : Nat → Expr
  | lam      : Ty → Expr → Expr                    -- λ:A. e
  | app      : Expr → Expr → Expr                  -- e₁ e₂
  | pair     : Expr → Expr → Expr                  -- (e₁, e₂)
  | fst      : Expr → Expr                         -- e.1
  | snd      : Expr → Expr                         -- e.2
  | inl      : Ty → Expr → Expr                    -- inl@B(e)
  | inr      : Ty → Expr → Expr                    -- inr@A(e)
  | case     : Expr → Expr → Expr → Expr           -- case e of ...
  | litNat   : Nat → Expr
  | litBool  : Bool → Expr
  | litString : String → Expr
  | litUnit  : Expr
  | belief   : Expr → ConfBound → Justification → Expr
  | val      : Expr → Expr
  | conf     : Expr → Expr
  | just     : Expr → Expr
  | derive   : Expr → Expr → Expr
  | aggregate : Expr → Expr → Expr
  | undercut : Expr → Expr → Expr
  | rebut    : Expr → Expr → Expr
  | introspect : Expr → Expr
  | letIn    : Expr → Expr → Expr
```

The `Justification` type tracks derivation structure:

```
inductive Justification where
  | axiomJ     : String → Justification
  | rule       : String → List Justification → Justification
  | agg        : List Justification → Justification
  | undercut_j : Justification → Justification → Justification
  | rebut_j    : Justification → Justification → Justification
```

d) *A.4.4 Typing Judgment:*

The typing judgment `Γ ⊢ e : A @c` is defined as an inductive proposition:

```
/-- Main typing judgment: Γ ⊢ e : A @c -/
inductive HasType : Ctx → Expr → Ty → ConfBound → Prop where
  | var : ∀ {Γ : Ctx} {n : Nat} {A : Ty} {c : ConfBound},
      Γ.lookup n = some ⟨A, c⟩ → HasType Γ (Expr.var n) A c
  | lam : ∀ {Γ : Ctx} {A B : Ty} {c_A c_B : ConfBound} {e : Expr},
      HasType (Γ ,, ⟨A, c_A⟩) e B c_B →
      HasType Γ (Expr.lam A e) (Ty.fn A B) c_B
  | app : ∀ {Γ : Ctx} {e₁ e₂ : Expr} {A B : Ty} {c₁ c₂ : ConfBound},
      HasType Γ e₁ (Ty.fn A B) c₁ →
      HasType Γ e₂ A c₂ →
      HasType Γ (Expr.app e₁ e₂) B (c₁ * c₂)
  -- ... 17 additional rules ...
```

Key typing rules:

- `app`

  : Confidence multiplies (conjunctive derivation)
- `aggregate`

  : Confidence uses

  `⊕`

  (independent evidence)
- `undercut`

  : Confidence uses

  `undercut(c,d) = c×(1-d)`
- `introspect`

  : Requires level constraint

  `m < n`

  and applies Löb discount

e) *A.4.5 Stratified Belief Introspection:*

The stratified belief system enforces Tarski's hierarchy:

```
/-- Introspect a lower-level belief from a higher level.
    This is the key operation enforcing Tarski's hierarchy.

    - Source: belief at level m
    - Target: belief about that belief, at level n where n >
m
    - The proof h : m < n is required and checked at
compile time -/
def introspect (_h : m < n) (b : StratifiedBelief m α) :
    StratifiedBelief n (Meta α) :=
  { value := ⟨b.value, none⟩
    confidence := b.confidence }
```

The safety theorems:

```
/-- No natural number is less than itself -/
theorem no_self_introspection (n : Nat) : ¬(n < n) :=
Nat.lt_irrefl n

/-- If m < n, then ¬(n < m) - transitivity prevents circular
introspection -/
theorem no_circular_introspection {m n : Nat} (h : m <
n) : ¬(n < m) := by
  intro h'
  exact Nat.lt_irrefl m (Nat.lt_trans h h')
```

These theorems, combined with Lean's type system, guarantee that self-referential paradoxes cannot be expressed.

f) *A.4.6 Evaluation Function:*

The computable evaluator uses fuel-bounded iteration:

```
/-- Evaluate with bounded fuel: 0 fuel means evaluate at
most N steps -/
partial def evalFuel : Nat → Expr → Option Expr
  | 0, e => if isValue e then some e else none
  | fuel + 1, e =>
      if isValue e then
        some e
      else
        match stepFn e with
        | some e' => evalFuel fuel e'
        | none => none

/-- Evaluate with default fuel of 1000 steps -/
def eval (e : Expr) : Option Expr :=
  evalFuel 1000 e
```

The

`stepFn`

function implements all reduction rules:

- Beta reduction:

  `(λx. e) v ⟶ e[x := v]`
- Projection:

  `(e₁, e₂).1 ⟶ e₁`
- Case analysis:

  `case (inl v) e₁ e₂ ⟶ e₁[x := v]`
- Belief operations:

  `derive`

  ,

  `aggregate`

  ,

  `undercut`

  ,

  `rebut`

  compute confidence adjustments

E. A.5 Five Properties Demonstration

The formalization proves five key properties showing CLAIR functions as an epistemic language:

1) **Beliefs track confidence through computation**
   - The

     `belief`

     constructor stores confidence as a

     `ConfBound`
   - All operations (

`derive`
,
`aggregate`
,
`undercut`
,
`rebut`
) compute new confidences
- The
`eval`
function preserves confidence in final values

2) **Evidence is affine (no double-counting)**
- The
`derive`
operation uses multiplication:
`c₁ × c₂`
- Multiplication is sub-linear in both arguments
- No operation allows "splitting" a belief to preserve confidence

3) **Introspection is safe**
- `StratifiedBelief.introspect`
requires proof of
`m < n`
- Theorems
`no_self_introspection`
and
`no_circular_introspection`
are machine-checked
- The Meta wrapper prevents confusion between levels

4) **Defeat operations modify confidence correctly**
- `undercut`
reduces confidence via multiplication
- `rebut`
normalizes competing confidences
- Boundedness theorems ensure results stay in [0,1]

5) **Type checking is decidable**
- The
`HasType`
inductive is decidable (all premises are decidable)
- Confidence operations use
`ConfBound`
(rational numbers in [0,1])
- The
`HasType.sub`
rule allows subtyping with explicit bounds

These properties are demonstrated through the theorems listed in §A.3. Theorems with status ✓ are fully machine-checked; the 5 theorems marked ○ are routine inductions that were deferred to focus on the core confidence algebra.

F. A.6 Relationship to Dissertation Claims

a) *Claim: "Machine-Checked Proofs" (Chapter 9):*
The formalization provides machine-checked proofs for:
- **Confidence Algebra** (Chapter 3): All associativity, commutativity, boundedness, monotonicity theorems
- **Non-Semiring Property** (Chapter 3): Explicit counterexample proving
`×`
does not distribute over
`⊕`
- **Stratification Safety** (Chapter 6): No self-introspection, no circular introspection
- **Belief Monad Laws** (Chapter 4): Functor and monad laws for stratified beliefs

The 5
`sorry`
lemmas are in substitution/weakening—theorems that are standard in PL theory but orthogonal to CLAIR's novel contributions.

b) *Claim: "Decidable Type Checking" (Chapter 10):*
The
`HasType`
judgment is decidable because:
- All premises are either structural (lookups in contexts) or arithmetic on
`ConfBound`
- The
`ConfBound`
type is
`ℚ ∩ [0,1]`

, allowing exact comparison
- No undecidable semantic conditions (e.g., "there exists a model") appear in the rules

c) *Claim: "Runnable Interpreter" (Chapter 10):*
The

`eval`

function is a partial, computable function that:
- Returns

  `some v`

  if

  `e`

  evaluates to value

  `v`

  within 1000 steps
- Returns

  `none`

  if

  `e`

  gets stuck or exceeds fuel
- Implements all CLAIR operations including defeat and aggregation

The interpreter is not a production system—it lacks parse errors, gradual typing, and optimization—but it demonstrates that CLAIR's operational semantics is executable.

G. A.7 Future Work

The formalization could be extended in several directions:
1) **Complete substitution lemmas**: Prove the 5 remaining

   `sorry`

   declarations
2) **Type preservation theorem**: Prove that well-typed programs reduce to well-typed values
3) **Progress theorem**: Prove that well-typed closed programs either are values or can step
4) **CPL completeness**: Formalize the Kripke semantics and prove completeness for CPL-finite
5) **Decision procedures**: Implement a tactic that decides

   `CPL-0`

   validity

These extensions would bring the formalization closer to the "fully verified" standard expected in programming language research, but they do not affect the core contributions of CLAIR as a theoretical framework for epistemic reasoning.

XI. Reference Interpreter Design

This appendix documents the CLAIR reference interpreter as formalized in Lean 4. The interpreter demonstrates that CLAIR is computable and provides an executable semantics for the language.

A. B.1 Architecture Overview

The CLAIR interpreter follows a standard pipeline architecture for programming language implementation:
+—+—+—+ | **Component** | **Purpose** | **Lean Module** | +—+—+—+ | Parser | Convert surface syntax to AST |

`CLAIR.Parser`

| | Type Checker | Verify well-typedness and stratification |

`CLAIR.Typing.HasType`

| | Single-Step Evaluator | Execute one reduction step |

`CLAIR.Semantics.Step`

| | Multi-Step Evaluator | Execute to value (with fuel) |

`CLAIR.Semantics.Eval`

| +—+—+—+

Source Code
↓
Parser
↓
AST (Expr)
↓
Type Checker
↓
Typed AST
↓
Evaluator (Step + Eval)
↓
Value

Fig. 1. B.1: CLAIR interpreter architecture pipeline

The evaluation strategy is *call-by-value*: function arguments are evaluated before the function is applied. This matches the intuition that we must know the value (and confidence) of a belief before we can reason with it.

B. B.2 Single-Step Semantics

The single-step reduction relation

`#text(code)[e -> e']`

is defined inductively in

`CLAIR.Semantics.Step`

. We show key rules here:

a) *B.2.1 Core Lambda Calculus Rules:*

Theorem Beta Reduction. For any type

A

, expression

e

, and value

v

:

#text(code)[app (lam A e) v -> subst0 v e]

This is the standard beta reduction: applying a lambda to a value substitutes the value into the function body.

Theorem Application Contexts. If

#text(code)[e1 -> e1']

then:

#text(code)[app e1 e2 -> app e1' e2']

If

#text(code)[e2 -> e2']

and

v1

is a value then:

#text(code)[app v1 e2 -> app v1 e2']

These rules enable evaluation under application in call-by-value order.

b) *B.2.2 Belief Operations:*

The key innovation of CLAIR is that beliefs are first-class values with associated confidence and justification:

Theorem Value Extraction. For belief

#text(code)[belief v c j]

where

v

is a value:

#text(code)[val (belief v c j) -> v]

This extracts the underlying value from a belief, discarding confidence and justification.

Theorem Confidence Extraction. For belief

#text(code)[belief v c j]

where

v

is a value:

#text(code)[conf (belief v c j) -> belief v c j]

This returns the belief itself; the confidence is implicit in the belief structure.

Theorem Justification Extraction. For belief

#text(code)[belief v c j]

where

v

is a value:

#text(code)[just (belief v c j) -> litString (toString j)]

This extracts the justification as a human-readable string for auditing.

c) *B.2.3 Defeat Operations:*

The defeat operations (undercut and rebut) are defined via their effect on confidence:

Theorem Undercut Evaluation. For beliefs

#text(code)[belief v c1 j1]

and

#text(code)[belief d c2 j2]

where both

v

and

d

are values:

#text(code)[undercut (belief v c1 j1) (belief d c2 j2) -> belief v (c1 * (1 - c2)) (undercut_j j1 j2)]

Undercut reduces confidence multiplicatively: a defeater with confidence c2

reduces confidence by a factor of

#text(code)[1 - c2]

.

Theorem **Rebuttal Evaluation.** For beliefs

#text(code)[belief v1 c1 j1]

and

#text(code)[belief v2 c2 j2]

where both are values:

#text(code)[rebut (belief v1 c1 j1) (belief v2 c2 j2) -> belief v1 (c1 / (c1 + c2)) (rebut_j j1 j2)]

Rebuttal normalizes by relative strength. When

#text(code)[c1 = c2]

, confidence becomes

#text(code)[1/2]

.

C. *B.3 Multi-Step Evaluation with Fuel*

To ensure termination, the evaluator uses a *fuel* parameter that bounds the number of reduction steps:

Definition **Fuel-Bounded Evaluation.**

#text(code)[evalFuel : Nat -> Expr -> Option Expr]

- #text(code)[evalFuel 0 e]

  returns

  #text(code)[some e]

  if

  e

  is a value,

  #text(code)[none]

  otherwise
- #text(code)[evalFuel (n+1) e]

  returns

  #text(code)[some e']

  if

  e

  evaluates to a value in

  n+1

  steps,

  #text(code)[none]

  if stuck

The default

#text(code)[eval]

function uses 1000 steps of fuel:

#text(code)[def eval (e : Expr) : Option Expr := evalFuel 1000 e]

D. *B.4 Example Walkthroughs*

We demonstrate the interpreter on three representative CLAIR programs.

a) *B.4.1 Simple Belief Formation:*

Example **Direct Belief.** Surface syntax:

```
(belief 42 0.9 "sensor-reading")
```

Lean representation:

```
def example_belief : Expr :=
  belief (litNat 42) (9/10) (Justification.axiomJ "sensor-reading")
```

Evaluation:

- Expression is already a value
- #text(code)[eval example_belief]

  returns

  #text(code)[some example_belief]
- The belief carries value

  42

  with confidence

  0.9

b) *B.4.2 Evidence Aggregation:*

Example Independent Evidence Combination. Surface syntax:

```
(aggregate
 (belief "Paris is capital of France" 0.5 "prior")
 (belief "Paris is capital of France" 0.7 "textbook"))
```

Lean representation:

```
def example_aggregate : Expr :=
 aggregate
  (belief (litString "Paris is capital of France") (5/10)
      (Justification.axiomJ "prior"))
  (belief (litString "Paris is capital of France") (7/10)
      (Justification.axiomJ "textbook"))
```

Step-by-step evaluation:

1) Both arguments are values (beliefs)
2) Apply probabilistic sum:

```
#text(code)[c_new = c1 + c2 - c1*c2]
```

3)

```
#text(code)[c_new = 0.5 + 0.7 - 0.5*0.7 = 1.2 - 0.35 = 0.85]
```

4) Combined justification:

```
#text(code)[Justification.agg ["prior", "textbook"]]
```

5) Result:

```
#text(code)[belief "Paris..." 0.85 (agg ["prior", "textbook"])]
```

This demonstrates the probabilistic sum operator: two independent pieces of evidence combine to increase confidence beyond either individual source.

c) *B.4.3 Undercutting in Action:*

Example Defeasible Reasoning. Scenario: A sensor reports "temperature is 25°C" with confidence 0.8, but a calibration warning suggests the sensor may be malfunctioning with confidence 0.3.
Surface syntax:

```
(undercut
 (belief 25 0.8 "sensor-A")
 (belief "sensor-A unreliable" 0.3 "calibration-warning"))
```

Lean representation:

```
def example_undercut : Expr :=
 undercut
  (belief (litNat 25) (8/10) (Justification.axiomJ "sensor-A"))
  (belief (litString "sensor-A unreliable") (3/10)
      (Justification.axiomJ "calibration-warning"))
```

Step-by-step evaluation:

1) Both arguments are values
2) Apply undercut formula:

```
#text(code)[c_new = c1 * (1 - c2)]
```

3)

```
#text(code)[c_new = 0.8 * (1 - 0.3) = 0.8 * 0.7 = 0.56]
```

4) Result:

```
#text(code)[belief 25 0.56 (undercut_j "sensor-A" "calibration-warning")]
```

The calibration warning reduces confidence from 0.8 to 0.56. The justification tracks that this belief was undercut, preserving the reasoning history.

d) *B.4.4 Rebuttal and Confidence Collapse:*

Example Conflicting Evidence. Scenario: Two sources disagree on whether a fact holds, with confidences 0.7 and 0.3 respectively.
Lean representation:

```
def example_rebut : Expr :=
 rebut
  (belief (litBool true) (7/10) (Justification.axiomJ "source-A"))
  (belief (litBool false) (3/10) (Justification.axiomJ "source-B"))
```

Step-by-step evaluation:

1) Both arguments are values
2) Apply rebuttal formula:

```
#text(code)[c_new = c1 / (c1 + c2)]
```

3)

```
#text(code)[c_new = 0.7 / (0.7 + 0.3) = 0.7 / 1.0 = 0.7]
```

4) Result:

```
#text(code)[belief true 0.7 (rebut_j "source-A" "source-B")]
```

Note that when

```
#text(code)[c1 = c2]
```

(equally strong conflicting evidence), confidence collapses to

```
#text(code)[1/2]
```

. This represents a state of maximal uncertainty.

e) *B.4.5 Derivation Chain:*

Example Multi-Step Derivation. Scenario: Derive a conclusion from two premises using the

```
derive
```

operator.
Lean representation:

```
def example_derivation : Expr :=
  derive
    (belief (litNat 1) (8/10) (Justification.axiomJ
"premise1"))
    (belief (litNat 2) (8/10) (Justification.axiomJ
"premise2"))
```

Step-by-step evaluation:

1) Both arguments are values
2) Apply derivation formula:

```
#text(code)[c_new = c1 * c2]
```

3)

```
#text(code)[c_new = 0.8 * 0.8 = 0.64]
```

4) Result:

```
#text(code)[belief (pair 1 2) 0.64 (rule
"derive" ["premise1", "premise2"])]
```

The result pairs the premise values, and confidence is the product (representing the conjunction strength). The justification records that this came from a derivation rule.

E. B.5 Key Properties

The Lean formalization proves five key properties that demonstrate CLAIR's correctness as an epistemic reasoning system:

Definition **Property 1: Beliefs Track Confidence.** Confidence is preserved through computation: if a belief

```
#text(code)[b]
```

has confidence

```
#text(code)[c]
```

, then after any sequence of valid reductions

```
#text(code)[b ->* b']
```

, the resulting belief

```
#text(code)[b']
```

has a confidence value that is a deterministic function of

```
#text(code)[c]
```

and the operations applied.

Definition **Property 2: Evidence is Affine.** No evidence is double-counted. The

```
#text(code)[aggregate]
```

operator uses probabilistic sum

```
#text(code)[a ⊕ b = a + b - a*b]
```

, which equals the probability of the union of independent events. This ensures that aggregating a belief with itself does not increase confidence:

```
#text(code)[c ⊕ c = c + c - c*c = c(2 - c)]
```

which is only equal to

```
#text(code)[c]
```

when

```
#text(code)[c = 0]
```

or

```
#text(code)[c = 1]
```

. In practice, justification tracking prevents exact duplicate aggregation.

Definition **Property 3: Introspection is Safe.** The

```
#text(code)[introspect]
```

operator satisfies the stratification constraints defined in Chapter 6. It is impossible to form a belief about the current confidence of that same belief, preventing the formation of self-referential paradoxes.

Definition **Property 4: Defeat Operations are Correct.** Undercut and rebut satisfy their specifications:

- Undercut is monotonic in both arguments: higher source confidence or higher defeater confidence yields predictable results
- Rebuttal is symmetric:

  ```
  #text(code)[rebut b1 b2]
  ```

  and

  ```
  #text(code)[rebut b2 b1]
  ```

  yield beliefs about opposing values with complementary confidence

Definition **Property 5: Type Checking is Decidable.** The bidirectional type checker in

```
CLAIR.Typing.HasType
```

terminates for all well-formed inputs. This is proven formally in the Lean code by showing that each recursive call decreases a measure (expression size or stratification level).

F. B.6 Implementation Notes

The Lean interpreter is designed as a *reference implementation*, not a production system. Key design decisions:

+—+—+ | **Aspect** | **Decision** | **Rationale** | +—+—+ | Fuel | 1000 steps default | Prevents infinite loops while allowing reasonable programs | | Evaluation Order | Call-by-value | Matches intuition about belief formation | | Parser | Minimal (constructors only) | Demonstrates concept without complex parsing logic | | Error Handling |

`Option Expr`

(partial function) | Stuck states return

`none`

; can be extended with explicit errors | +—+—+

For production use, we recommend:

1) A proper parser (e.g., using Megaparsec in Haskell)
2) Exception-based error handling with detailed error messages
3) JIT compilation for performance
4) Persistent justification storage for audit trails
5) Parallel evaluation for independent subexpressions

G. B.7 Relation to Chapter 10

Chapter 10 discusses implementation considerations for a production CLAIR system. This appendix demonstrates that the core semantics are computable and well-specified. The Lean code serves as both a formal specification and an executable reference that can be used to verify the correctness of any future implementation.

H. B.8 Haskell Reference Implementation

In addition to the Lean formalization, a complete Haskell reference implementation is provided. This implementation demonstrates that CLAIR can be realized in a general-purpose programming language while maintaining the semantic properties proved in Lean.

a) *B.8.1 Project Structure:*

The Haskell implementation is organized as a Cabal project with the following structure:

+—+—+ | **Module** | **Purpose** | +—+—+ |

`CLAIR.Syntax`

| Abstract syntax trees (AST) for CLAIR expressions | |

`CLAIR.Confidence`

| Confidence algebra: ⊕, ⊗, undercut, rebut | |

`CLAIR.Parser`

| Parse surface syntax into AST (Parsec) | |

`CLAIR.TypeChecker`

| Bidirectional type checking with confidence grades | |

`CLAIR.Evaluator`

| Small-step operational semantics with fuel | |

`CLAIR.Pretty`

| Pretty-printing for values and types | |

`CLAIR.TypeChecker.Types`

| Type checker context and error types | +—+—+

The implementation includes:

- A REPL (

  `app/Main.hs`

) for interactive evaluation
- A comprehensive test suite (

  `test/`

) with 35 passing tests
- QuickCheck properties for algebraic laws
- Example programs demonstrating belief operations

b) *B.8.2 Syntax Definition:*

The core AST in

`CLAIR.Syntax`

mirrors the Lean formalization:

```
-- | A belief value with all its annotations
data Belief = Belief
  { beliefValue      :: Expr           -- The proposition/content
  , beliefConf       :: Confidence     -- Confidence level [0,1]
  , beliefJustify    :: Justification  -- Supporting arguments
  , beliefInvalidate :: Invalidation   -- Defeating
information
  , beliefProvenance :: Provenance     -- Source tracking
  }

-- | Core expression language
data Expr
  = EVar Name               -- Variable: x
  | ELam Name Type Expr     -- Lambda: λx:A. e
  | EApp Expr Expr          -- Application: e1 e2
  | EAnn Expr Type          -- Type annotation: e : A
  | EBelief Belief          -- Belief: belief(v,c,j,i,p)
  | EBox Confidence Expr    -- Self-reference: □_c e
  | EPrim Op Expr Expr      -- Primitive operation
  | ELit Literal            -- Literal value
```

Key design decisions:

- GADTs enable type-safe AST construction
- Deriving

  `Generic`

  and

  `ToJSON/FromJSON`

  enables serialization
- Provenance and justification are explicit for audit trails

c) *B.8.3 Confidence Algebra:*

The

```
CLAIR.Confidence
```

module implements the confidence operations with careful attention to semantic correctness:

```
-- | Probabilistic sum: a ⊕ b = 1 - (1-a)(1-b) = a + b - ab
-- Assumes: sources are conditionally independent
oplus :: Confidence -> Confidence -> Confidence
oplus (Confidence a) (Confidence b) = clamp (a + b - a *
b)

-- | Product t-norm: a ⊗ b = a * b
otimes :: Confidence -> Confidence -> Confidence
otimes (Confidence a) (Confidence b) = clamp (a * b)

-- | Apply undercut defeat: multiply by (1-d)
-- Rationale: Undercut attacks the evidential connection
undercut :: Defeat -> Confidence -> Confidence
undercut (Defeat d) (Confidence c) = clamp (c * (1 - d))

-- | Apply rebut with normalization
-- Limitation: Collapses absolute strength; considers
uncertainty-preserving alternatives
rebut :: Defeat -> Confidence -> Confidence ->
Confidence
rebut (Defeat d_strength) (Confidence c_for) (Confidence
c_against_base) =
 let c_against = d_strength * c_against_base
     total = c_for + c_against
 in if total == 0
    then Confidence 0.5  -- ignorance prior
    else clamp (c_for / total)

-- | Square discount for self-reference: g(c) = c²
-- Prevents bootstrapping while preserving high-
confidence self-endorsement
squareDiscount :: DiscountFn
squareDiscount (Confidence c) = clamp (c * c)
```

d) *B.8.4 Type Checking:*

The bidirectional type checker in

```
CLAIR.TypeChecker
```

implements the rules from Appendix E:

```
-- | Infer (synthesize): Γ ⊢ e ↑ τ
infer :: Context -> Expr -> Either TypeError TCResult
infer ctx expr = case expr of
  -- Variable: Γ ⊢ x : Γ(x)
  EVar x -> case ctxLookup x ctx of
    Just ty -> return (TCResult ty ctx)
    Nothing -> Left (UnboundVar x)

  -- Application: Γ ⊢ e ↑ τ₁ → τ₂, Γ ⊢ e' ↓ τ₁
  EApp e1 e2 -> do
    TCResult ty1 ctx1 <- infer ctx e1
    case ty1 of
      TFun argTy resTy -> do
        ctx2 <- check ctx1 e2 argTy
        return (TCResult resTy ctx2)
      _ -> Left (NotFunction ty1)

  -- Belief: Γ ⊢ e : τ, c ∈ [0,1]
  --        ─────────────────────────
  --        Γ ⊢ belief(e,c) ↑ Belief_c[τ]
  EBelief (Belief e c _ _ _) -> do
    unless (isNormalized c) $
      Left (InvalidConfidence c)
    TCResult ty ctx' <- infer ctx e
    let beliefTy = TBelief c ty
    return (TCResult beliefTy ctx')
```

The bidirectional approach provides:

- Better error messages (checking mode guides synthesis)
- Natural handling of implicit arguments
- Clear separation of inference vs. verification

e) *B.8.5 Evaluation:*

The small-step evaluator in

```
CLAIR.Evaluator
```

implements the operational semantics with fuel-bounded termination:

```
-- | Single-step reduction: e ⟶ e'
step :: Env -> Expr -> Either EvalError (Maybe Expr)
step env expr = case expr of
  -- E-Beta: (λx:τ.e) v ⟶ e[x := v]
  EApp (ELam varName _ty body) arg
    | isValue arg -> do
      v <- evalExpr env arg
      return (Just (subst varName v body))

  -- E-Prim: Reduce primitive operations
  EPrim op e1 e2
    | isValue e1 && isValue e2 -> evalPrimOp env op e1 e2

  -- E-Belief: Evaluate belief content to value
  EBelief (Belief e c j i p) -> do
    me' <- step env e
    case me' of
      Just e' -> return (Just (EBelief (Belief e' c j i p)))
      Nothing -> return Nothing  -- Fully evaluated

  -- E-Box: □_c e becomes value when e is fully evaluated
  EBox c e -> do
    me' <- step env e
    case me' of
      Just e' -> return (Just (EBox c e'))
      Nothing -> return Nothing
```

Key features:

- **Fuel**: 1,000,000 steps default prevents infinite loops
- **Call-by-value**: Arguments evaluated before application

- **Capture-avoiding substitution**: Preserves variable hygiene
- **Error recovery**: Detailed error messages for debugging

f) *B.8.6 REPL Usage:*

The interactive REPL (

`app/Main.hs`

) supports:

```
clair> :help
CLAIR REPL - Commands:
 :quit  Exit the REPL
 :help  Show this help message

Examples:
 5
 λx:Nat. x
 (λx:Nat. x + 1) 5
 3 + 4
 □0.8 true
 belief(5, 0.9, none, none, none)
```

The REPL provides:

- Parse error reporting with locations
- Type checking with confidence grades
- Evaluation with fuel tracking
- Pretty-printed results

g) *B.8.7 Test Suite:*

The implementation includes 35 QuickCheck and HUnit tests covering:

+—+—+—+ | **Module** | **Tests** | **Coverage** | +—+—+—+ |

`CLAIR.Test.Confidence`

| 12 | Algebraic laws: ⊕ associativity, ⊗ identity, undercut monotonicity | |

`CLAIR.Test.Evaluator`

| 14 | Beta reduction, primitive operations, belief evaluation | |

`CLAIR.Test.TypeChecker`

| 6 | Type inference, subtyping, error cases | |

`CLAIR.Test.HelloWorld`

| 3 | End-to-end belief formation and evaluation | +—+—+ —+

Sample QuickCheck properties:

```haskell
-- Probabilistic sum is associative
prop_oplus_assoc :: Confidence -> Confidence ->
Confidence -> Bool
prop_oplus_assoc a b c =
 oplus a (oplus b c) == oplus (oplus a b) c

-- Undercut is monotonic in defeat strength
prop_undercut_monotonic :: Confidence -> Defeat ->
Defeat -> Property
prop_undercut_monotonic c d1 d2 =
 d1 <= d2 ==> undercut d1 c >= undercut d2 c
```

All tests pass:

`cabal test`

→

`35/35 tests passed`

.

h) *B.8.8 Building and Running:*

Build:

```bash
cd implementation/haskell
cabal build
```

This compiles the REPL (

`clair-repl`

) and test suite (

`clair-test`

).

Run REPL:

```bash
cabal run clair-repl
```

Run Tests:

```bash
cabal test
```

i) *B.8.9 Design Rationale:*

The Haskell implementation makes specific design choices that differ from the Lean reference:

+—+—+—+—+ | Aspect | Lean | Haskell | Rationale | +—+—+ —+ | Confidence type |

`Rat`

(rational) |

`Double`

| Performance vs. precision tradeoff | | Parser | Constructors only | Parsec | Usable surface syntax | | Error handling |

`Option Expr`

|

`Either EvalError`

| Stack traces for debugging | | Fuel | 1000 steps | 1,000,000 steps | Allow more complex programs | | Substitution | Formal proof | Direct implementation | No proof burden, faster iteration | +—+—+—+—+

These choices reflect the different purposes:

- Lean: Formal verification, mathematical rigor
- Haskell: Usability, testing, experimentation

j) *B.8.10 Relation to Lean Formalization:*

The Haskell implementation is verified against the Lean formalization by:

1) Type system correspondence: Haskell types mirror Lean inductive types
2) Semantic equivalence: Reduction rules match Lean's

   step

   relation
3) Property-based testing: QuickCheck laws reflect Lean theorems
4) Test coverage: Each Lean example has a corresponding Haskell test

Discrepancies are documented as limitations:

- Double

  precision may cause floating-point drift
- Substitution is not proven capture-avoiding
- Stratification is checked but not enforced statically

For production deployment, we recommend:

- Use arbitrary-precision rationals for confidence
- Add formal proofs for substitution correctness
- Implement static stratification analysis
- Add persistent justification storage

XII. Additional Proofs

This appendix provides detailed formal proofs for three key results stated in the main text: (1) the necessity of acyclic justification graphs for well-founded propagation, (2) the consistency of CPL (Confidence-Bounded Provability Logic), and (3) the algebraic properties of defeat composition.

A. C.1 DAG Necessity for Well-Founded Confidence Propagation

a) *C.1.1 Statement of the Problem:*

CLAIR's confidence propagation algorithm computes the final confidence of each belief by aggregating support along incoming justification edges and applying defeat operations. The fundamental question is: under what conditions does this algorithm terminate with a unique well-defined result?

b) *C.1.2 The Cyclic Counterexample:*

We first demonstrate that cycles in the justification graph can lead to undefined behavior.

> **C.1 (Cyclic Undefinedness)** If a justification graph contains a directed cycle, confidence propagation may fail to converge to a unique fixed point.

Proof. Consider a simple cycle of two beliefs:

- Belief A with initial confidence c_A = 0.5, derived from belief B with strength s_1 = 0.8
- Belief B with initial confidence c_B = 0.5, derived from belief A with strength s_2 = 0.8

The propagation rules specify: c_A' = c_A ⊕ (c_B × s_1) c_B' = c_B ⊕ (c_A × s_2)

Starting from (c_A, c_B) = (0.5, 0.5), iteration yields:

Iteration 1: c_A = 0.5 ⊕ (0.5 × 0.8) = 0.5 ⊕ 0.4 = 0.7 c_B = 0.5 ⊕ (0.5 × 0.8) = 0.7

Iteration 2: c_A = 0.5 ⊕ (0.7 × 0.8) = 0.5 ⊕ 0.56 = 0.78 c_B = 0.78

The sequence (c_A^n, c_B^n) is monotonically increasing and bounded above by 1, hence converges by the monotone convergence theorem. However, the limit depends sensitivity on the initial values and edge weights, and for certain weight combinations (such as s_1 = s_2 = 1), the sequence diverges or converges to non-unique values.

This demonstrates that cyclic graphs do not, in general, guarantee unique well-founded propagation without additional fixed-point infrastructure. ∎ ∎

c) *C.1.3 The DAG Well-Foundedness Theorem:*

> **C.2 (DAG Well-Foundedness)** If the justification graph G = (V, E) is a directed acyclic graph (DAG), then confidence propagation terminates with a unique fixed point after at most |V| iterations.

Proof. We proceed by induction on the topological ordering of the DAG.

Let ? be a topological order on V, meaning that if (u, v) ∈ E, then u ? v. Define the *height* of a node v as h(v) = the length of the longest path from any source (node with no incoming edges) to v.

Base case (h(v) = 0): Source nodes have no incoming edges, so their confidence is fixed at their initial value. Propagation terminates immediately for these nodes.

Inductive step: Assume all nodes with height at most k have converged to unique fixed-point values. Consider a node v with height h(v) = k + 1.

All predecessors u of v satisfy h(u) ≤ k, hence have converged by the inductive hypothesis. Let c_u-final denote the converged confidence of predecessor u.

The propagation rule for v is: c_v = c_v(init) ⊕ ? (u,v) ∈ E (c_u × w_uv)

Since each c_u has converged to c_u-final, and all operations (⊕ and ×) are continuous, the value of c_v is uniquely determined and converges in exactly one iteration once all predecessors have converged.

By induction, all nodes converge to unique values. The maximum number of iterations required is max_v ∈ V h(v) ≤ |V| - 1, the length of the longest path in the DAG. ∎ ∎

d) *C.1.4 Practical Implications:*

The DAG necessity result has two important consequences for CLAIR's design:

1) *Type-System Enforcement:* The CLAIR type checker must enforce acyclicity as a well-formedness condition on justification graphs. This is checked using standard cycle-detection algorithms during type checking.
2) *Expressive Limitations:* DAG-only semantics cannot directly represent certain defeasible reasoning patterns involving mutual support. Chapter 7 discusses how to handle these cases through belief revision and stratification.

—

B. C.2 CPL Consistency Proof

a) *C.2.1 CPL Axiom System:*

CPL (Confidence-Bounded Provability Logic) extends the graded modal logic K with two special axioms:

> **Axiom 4 (Graded Transitivity)** The graded modal axiom 4: box_c(p) implies box_fc(box_c(p)) for some strictly increasing function f.

> **Axiom GL (Graded Löb)** The graded Löb axiom: box_c(box_c(p) ⟶ p) implies box_gc(p)

The key question is whether this axiom system is *consistent*—that is, whether there exists a non-trivial model satisfying all axioms.

b) *C.2.2 Finite Model Construction:*

We construct an explicit finite model demonstrating consistency.

> **C.3 (CPL Consistency)** There exists a finite Kripke model M = (W, R, V, c) satisfying all CPL axioms for f(c) = c and g(c) = c^2.

Proof. Let W = {0, 1, 2} be a set of three worlds. Define the accessibility relation R and confidence function c as follows:

R = {(0, 1), (0, 2), (1, 2)} (strictly increasing order)

For each edge (w, w') ∈ R, define c(w, w') as:

- c(0, 1) = 1/2
- c(0, 2) = 1/4
- c(1, 2) = 1/2

We verify the axioms:

Axiom K: Holds in all Kripke models by the standard modal logic semantics.

Axiom 4: For any world w and confidence c, if w satisfies box_c p, then for all w' with (w, w') ∈ R and c(w, w') ≥ c, we have w' satisfies p. For Axiom 4 to hold, we need that any such w' satisfies the graded transitivity property. In our model, this holds because if (w, w') ∈ R and (w', w'') ∈ R with appropriate confidence bounds, then (w, w'') ∈ R directly.

Axiom GL: The graded Löb axiom requires verification. For world 0: box_c (box_c p ⟶ p) means: for all worlds reachable from 0 with confidence ≥ c, if box_c p holds at that world, then p holds.

Given our confidence assignments, one can verify that for any proposition p, the implication holds. The discount function g(c) = c^2 ensures that self-referential soundness claims cap their own confidence, preventing bootstrapping.

Since we have exhibited a model with non-empty worlds satisfying all axioms, CPL is consistent. ∎ ∎

c) *C.2.3 Design Axiom Status:*

It is important to clarify that CPL's graded Löb axiom is a *design axiom* rather than a derived semantic theorem. This means:

1) We *choose* to include box_c (box_c p ⟶ p) ⟶ box_gc p as an axiom
2) We verify *consistency* (as shown above) but not *soundness* with respect to a pre-existing semantics
3) The axiom captures our intended behavior: self-soundness claims should be discounted to prevent bootstrapping

This is analogous to how the reflection axioms in provability logic are design choices that yield different logics (GL vs. S4 vs. K).

—

C. C.3 Defeat Composition Algebra

a) *C.3.1 Undercut Composition:*

The fundamental result for undercutting defeat is that sequential undercuts compose via the probabilistic OR operation.

> **C.4 (Undercut Composition)** For any confidences c, d_1, d_2 ∈ [0,1]: undercut(undercut(c, d_1), d_2) = undercut(c, d_1 ⊕ d_2)

Proof. We compute directly from the definition undercut(c, d) = c × (1 - d):

undercut(undercut(c, d_1), d_2)

XIII. UNDERCUT(C × (1 - D_1), D_2)

XIV. (C × (1 - D_1)) × (1 - D_2)

XV. C × ((1 - D_1) × (1 - D_2))

XVI. C × (1 - D_1 - D_2 + D_1 D_2)

XVII. C × (1 - (D_1 + D_2 - D_1 D_2))

XVIII. C × (1 - (D_1 ⊕ D_2))

XIX. UNDERCUT(C, D_1 ⊕ D_2)

where we use the definition d_1 ⊕ d_2 = d_1 + d_2 - d_1 d_2. ∎ ∎

a) *C.3.2 Corollaries of Undercut Composition:*

> **C.5 (Commutative Composition)** Undercut composition is commutative: undercut(undercut(c, d_1), d_2) = undercut(undercut(c, d_2), d_1)

Proof. Immediate from Theorem C.4 and commutativity of ⊕: undercut(undercut(c, d_1), d_2) = undercut(c, d_1 ⊕ d_2)

XX. UNDERCUT(C, D_2 ⊕ D_1) = UNDERCUT(UNDERCUT(C, D_2), D_1). ∎

∎

> **C.6 (Identity)** Undercut with zero defeat leaves confidence unchanged: undercut(c, 0) = c

Proof. undercut(c, 0) = c × (1 - 0) = c × 1 = c. ∎ ∎

> **C.7 (Annihilation)** Undercut with complete defeat eliminates confidence: undercut(c, 1) = 0

Proof. undercut(c, 1) = c × (1 - 1) = c × 0 = 0. ∎ ∎

a) *C.3.3 Rebut Algebra:*

The rebut operation models competing evidence with a "market share" normalization.

> **C.8 (Rebut Symmetry)** For any c_for, c_against ∈ [0, 1] with c_for + c_against > 0: rebut(c_for, c_against) + rebut(c_against, c_for) = 1

Proof. From the definition rebut(c_for, c_against) = c_for / (c_for + c_against):

rebut(c_for, c_against) + rebut(c_against, c_for)

XXI. C_FOR / (C_FOR + C_AGAINST) + C_AGAINST / (C_AGAINST + C_FOR)

XXII. (C_FOR + C_AGAINST) / (C_FOR + C_AGAINST)

XXIII. 1 ∎

∎

> **C.9 (Rebut Monotonicity)** Rebut is monotone in the first argument and antitone in the second:
> - If c_1 ≤ c_2, then rebut(c_1, c) ≤ rebut(c_2, c)
> - If d_1 ≤ d_2, then rebut(c, d_2) ≤ rebut(c, d_1)

Proof. For the first claim: rebut(c_1, c) = c_1 / (c_1 + c) ≤ c_2 / (c_2 + c) = rebut(c_2, c) since the function f(x) = x / (x + c) is increasing in x for c > 0.

The second claim follows similarly by noting rebut(c, d) = 1 - rebut(d, c) from Theorem C.8. ∎ ∎

a) *C.3.4 Interaction Between Undercut and Rebut:*

Undercut and rebut represent two fundamentally different types of defeat:

- *Undercut* attacks the *inferential link* between premises and conclusion
- *Rebut* attacks the *conclusion* directly with counter-evidence

The interaction of these two operations is subtle and context-dependent. In general, rebut is applied first to aggregate competing evidence, then undercut is applied to discount the resulting confidence based on link attackers. This ordering is reflected in CLAIR's evaluation semantics (Chapter 4).

b) *C.3.5 Limitation: Rebut Normalization:*

A key limitation of the rebut operation is that it normalizes away absolute evidence strength.

> **C.10 (Rebut Collapse)** For any scaling factor λ > 0: rebut(λ c_for, λ c_against) = rebut(c_for, c_against)

Proof. rebut(λ c_for, λ c_against) = λ c_for / (λ c_for + λ c_against)

XXIV. λ C_FOR / (λ (C_FOR + C_AGAINST))

XXV. C_FOR / (C_FOR + C_AGAINST)

XXVI. REBUT(C_FOR, C_AGAINST). ∎

∎

This means rebut cannot distinguish between "both weak" and "both strong but balanced" evidence. Chapter 6 discusses how CLAIR addresses this through provenance tracking, which preserves the absolute strengths of individual evidence sources even after rebut normalization.

XXVII. Glossary

This appendix provides concise definitions of key terms, notation, and acronyms used throughout this dissertation. Terms are marked with their primary chapter of introduction.

A. *D.1 Term Definitions*

a) *Epistemic Terms:*

+— +Term | Definition | Chapter +— +Belief | A first-class value in CLAIR consisting of a proposition, confidence, and justification. Beliefs can be constructed, combined, defeated, and inspected through language operations. | 1 +Confidence | A value in the unit interval [0,1] representing degree of epistemic commitment. Confidence is *not* probability: rather than quantifying frequency or subjective chance, confidence tracks how justified a belief is given available evidence. | 3 +Commitment (Epistemic) | The stance of endorsing a proposition as true to a specified degree. High confidence (close to 1) indicates strong commitment; low confidence (close to 0) indicates weak or no commitment. | 3 +Justification | A data structure tracking the derivation history of a belief. Justifications form a directed acyclic graph (DAG) where nodes represent beliefs and edges represent inference rules or evidence sources. | 4 +Invalidation | A condition associated with a belief specifying circumstances under which the belief should be reconsidered. Invalidation enables defeasible reasoning—beliefs can be defeated without logical contradiction. | 4 +Provenance | The origin or source of evidence for a belief. CLAIR tracks provenance through justification graphs, enabling audit trails for how conclusions were reached. | 4 +Reliability | In CLAIR's semantics, the tendency of a source to produce true beliefs. Reliability is the *semantic interpretation* of confidence: a belief with confidence **c** is interpreted as coming from a source with reliability **c**. | 3 +—

b) *Operations and Relations:*

+— +Term | Definition | Chapter +— +Probabilistic OR (⊕) | Aggregation operator for independent evidence: **a** ⊕ **b** = **a** + **b** - **ab**. This equals the probability that at least one of two independent events occurs. Satisfies commutativity, associativity, and has identity 0. | 3 +Undercut (?) | Defeat operation that reduces confidence multiplicatively: **undercut(c, d)** = **c** × **(1-d)**. A defeater with confidence **d** reduces target confidence by factor **(1-d)**. | 4 +Rebuttal |

Defeat operation for conflicting evidence: **rebut(c_1, c_2) = c_1 / (c_1 + c_2)**. Normalizes competing confidences to [0,1]; equal confidences yield 1/2. | 4 +Derivation | Combining beliefs via rule application: **derive(b_1, b_2)** pairs the values and multiplies confidences $c_1 \times c_2$. Tracks justification through rule application. | 4 +Aggregation | Combining independent evidence using ⊕: **aggregate(b_1, b_2)** produces belief with confidence $c_1 \oplus c_2$. Tracks justification through aggregation node. | 4 +Subtyping | Confidence ordering: belief at confidence c_1 can be used where belief at c_2 is required iff $c_1 \geq c_2$. Enables confidence weakening. | 10 +—

c) *Structural Properties:*

+— +Term | Definition | Chapter +— +Directed Acyclic Graph (DAG) | A graph with directed edges and no cycles. CLAIR requires justification graphs to be DAGs to ensure well-foundedness and prevent circular reasoning. | 4 +Stratification | Layering beliefs into levels **0, 1, 2, ...** where level **n** can only refer to levels <**n**. Enforces Tarski's hierarchy to prevent self-referential paradoxes. | 6 +Well-formedness | Constraints on CLAIR programs: (1) justification graphs must be acyclic, (2) stratification constraints on introspection, (3) confidences in [0,1]. | 10 +—

d) *Logical and Modal Terms:*

+— +Term | Definition | Chapter +— +CPL (Confidence-Bounded Provability Logic) | Graded extension of Gödel-Löb provability logic. Adds confidence grades to provability operator □. Axiomatizes self-referential reasoning with confidence discounts. | 5 +Löb's Theorem | Modal logic theorem: $\Box(\Box\varphi \to \varphi) \to \Box\varphi$. In provability logic, enables self-reference; in CLAIR, motivates anti-bootstrapping constraint. | 5 +Graded Modality | Modal operators with quantitative gradess. CLAIR's $\Box_{\langle c \rangle}$ means "provable with confidence at least **c**." | 5 +Anti-bootstrapping | Principle that self-validating claims cannot increase confidence. Formally: **c ≤ g(c)** for some discount function **g**. CLAIR uses **g(c) = c²**. | 5 +Kripke Semantics | Possible worlds framework for modal logic. CPL uses graded Kripke models where accessibility relations track confidence thresholds. | 5 +—

e) *Computational Terms:*

+— +Term | Definition | Chapter +— +Fuel | Bound on computation steps in the Lean evaluator: **evalFuel n e** evaluates for at most **n** steps. Prevents infinite loops while ensuring termination. | B +Call-by-value | Evaluation strategy: function arguments are evaluated before application. CLAIR uses call-by-value to match intuition about belief formation. | B +Small-step semantics | Reduction relation **e ⟶ e'** defining one step of computation. Composed to form multi-step evaluation **e ⟶ e'**. | 10 +Bidirectional Type Checking | Type checking algorithm with synthesis (infer type from expression) and checking (verify expression matches type). Enables practical implementation. | 10 +De Bruijn Indices | Variable representation using natural numbers: var 0 = most recent binder, var 1 = next recent, etc. Enables formal proofs about substitution. | A +—

f) *Argumentation and Belief Revision:*

+— +Term | Definition | Chapter +— +Defeasible Reasoning | Reasoning where conclusions can be defeated by new information without logical contradiction. CLAIR supports defeasibility through undercut and rebut operations. | 4 +Underminer | An argument that attacks the connection between evidence and conclusion (e.g., "the sensor is miscalibrated"). Reduces confidence via **undercut(c,d) = c×(1-d)**. | 4 +Rebuttal | An argument providing conflicting evidence for the opposite conclusion. Normalizes via **rebut(c_1,c_2) = c_1/(c_1+c_2)**. | 4 +AGM Theory | Classic belief revision framework (Alchourrón, Gärdenfors, Makinson). CLAIR extends AGM to graded, DAG-structured beliefs. | 7 +Contraction | Belief revision operation: remove a belief from a belief set. In CLAIR, achieved by setting confidence to 0. | 7 +Revision | Belief revision operation: add a belief while maintaining consistency. In CLAIR, achieved by aggregating with existing beliefs. | 7 +—

g) *Impossibility Results:*

+— +Term | Definition | Chapter +— +Gödel's Incompleteness | Any consistent formal system capable of arithmetic contains true but unprovable statements. Motivates CPL's design restrictions. | 12 +Church's Undecidability | First-order logic validity is undecidable. CLAIR restricts to decidable fragments (CPL-finite, CPL-0). | 12 +Tarski's Undefinability | Truth cannot be defined within the same language. Motivates stratification: level **n** cannot quantify over level **n**. | 12 +Löb's Paradox | Curious proposition using Löb's theorem that leads to contradiction if not carefully restricted. Resolved via stratification. | 6 +—

B. D.2 Notation Table

+— +Symbol | Meaning | Type | Chapter +— +**c** | Confidence value in [0,1] | Confidence | 3 +⊕ | Probabilistic OR: **a ⊕ b = a + b - ab** | Binary operation | 3 +? | Undercut: **c ? d = c × (1-d)** | Binary operation | 4 +× | Multiplication (conjunctive combination) | Binary operation | 3 +**g(c)** | Löb discount function; CLAIR uses **c²** | Unary function | 5 +$\Box_{\langle c \rangle}$ φ | Necessarily φ with confidence at least c | Modal operator | 5 +$\Diamond_{\langle c \rangle}$ φ | Possibly φ with confidence at least c | Modal operator | 5 +⊢ | Typing judgment: Γ ⊢ e : A @ c | Relation | 10 +⟶ | Small-step reduction: e ⟶ e' | Relation | 10 +⟶ **| Multi-step reduction (reflexive transitive closure) | Relation | 10 +Γ | Typing context (list of variable: type pairs) | Context | 10 +**A ⟹ B **| Function type from** A **to** B **| Type | 10 +Belief**<A> **| Belief type holding value of type** A **| Type constructor | 10 +**b**.value | Extract value from belief** b **| Projection | 4 +b.confidence | Extract confidence from belief** b **| Projection | 4 +**b**.justification | Extract justification from belief** b **| Projection | 4 +**m <n **| Stratification constraint: level** m **below** n **| Ordering | 6 +**∀ **| Universal quantifier | Quantifier | Appendix A +**∃ **| Existential quantifier | Quantifier | Appendix A +**∈ **| Set membership | Relation | Appendix A +**⊆ **| Subset relation | Relation | Appendix A +**∪ **| Set union | Binary operation | Appendix A +**∩ **| Set intersection | Binary operation | Appendix A +λ*x:A. e** | Lambda abstraction (anonymous function) | Expression | 10 +**e*1** e*2 | Function application | Expression | 10 +—

C. D.3 Acronyms

+— +Acronym | Full Name | Definition +— +CLAIR | Comprehensible LLM AI Intermediate Representation | The formal system presented in this dissertation +CPL | Confidence-Bounded Provability Logic | Graded modal logic extending Gödel-Löb +CPL-finite | CPL with finite confidence set {0, 1/n, 2/n, …, 1} | Decidable fragment suitable for implementation +CPL-0 | CPL with only confidence 0 (ungraded provability) | Collapses to standard provability logic GL +AGM | Alchourrón-Gärdenfors-Makinson | Classic belief revision theory +DAG | Directed Acyclic Graph | Graph structure for justifications +LLM | Large Language Model | AI systems CLAIR is designed to augment +IR | Intermediate Representation | Compiler representation between source and machine code +AI | Artificial Intelligence | Field of study +—

D. D.4 Type System Summary

a) *Base Types:*

+— +Type | Description | Example Values +— +Nat | Natural numbers (non-negative integers) | 0, 1, 2, 42, 128 +Bool | Boolean values | true, false +String | Text strings | “hello”, “sensor-reading” +Unit | Unit type (single value) | unit +Pair(**A**, **B**) | Ordered pair | (1, true), (“x”, 42) +Sum(**A**, **B**) | Tagged union (either **A** or **B**) | inl(5), inr(true) +**A** ⇒ **B** | Function type from **A** to **B** | λx:Nat. x + 1 +Belief<**A**> | Belief holding value of type **A** | belief(42, 0.9, j) +—

b) *Confidence Operations:*

+— +Operation | Notation | Formula | Identity +— +Probabilistic OR | **a ⊕ b** | **a + b - a×b** | 0 +Multiplication | **a × b** | **a × b** | 1 +Undercut | **undercut(c, d)** | **c × (1-d)** | N/A (d=0 gives c) +Rebuttal | **rebut(c_1, c_2)** | **c_1 / (c_1 + c_2)** | N/A +Minimum | **min(a, b)** | **min(a, b)** | 0 (if bounded below) +Maximum | **max(a, b)** | **max(a, b)** | 1 (if bounded above) +—

XXVIII. Complete CLAIR Language Specification

This appendix provides the complete formal specification of the CLAIR language, including syntax (concrete and abstract), static semantics (type system), and dynamic semantics (operational rules). The specification is self-contained and sufficient for implementing a conforming CLAIR interpreter.

A. E.1 Syntax

a) *E.1.1 Type Grammar:*

The type grammar defines the set of valid types in CLAIR.

+— +Category | Production | Description Base Types | B ::= “Nat” | Natural numbers | | B ::= “Bool” | Boolean values | | B ::= “String” | Text strings | | B ::= “Unit” | Unit type (single value) Confidence | c ∈ ℚ | Rational in [0,1] Types | A ::= B | Base type | | A ::= A ⟶ B | Function type | | A ::= A × B | Product type | | A ::= A + B | Sum type | | A ::= “Belief<”A“[“c”]” | Belief type with confidence bound | | A ::= “Meta<”A“[“n”]”[“c”]“ | Stratified meta-belief type +—

b) *E.1.2 Expression Grammar:*

The expression grammar defines the syntactic forms of CLAIR programs.

+— +Category | Production | Description Variables | e ::= x | Variable reference Lambdas | e ::= “λ”x“:”A“.”e | Anonymous function Application | e ::= e e | Function application Pairs | e ::= “(“e “, “ e”)” | Ordered pair Projections | e ::= e “.” “1” | First projection | | e ::= e “.” “2” | Second projection Injections | e ::= “inl@”B “(“e”)” | Left injection | | e ::= “inr@”A “(“e”)” | Right injection Case | e ::= “case” e “of” “inl” x “=>” e “|” “inr” y “=>” e | Sum elimination Literals | e ::= n | Natural number literal | | e ::= “true” | “false” | Boolean literals | | e ::= “"s"” | String literal | | e ::= “()” | Unit literal Beliefs | e ::= “belief(“e “, “ c “, “ j”)” | Belief constructor | | e ::= “val(“e”)” | Extract belief value | | e ::= “conf(“e”)” | Extract belief confidence | | e ::= “just(“e”)” | Extract belief justification Derivation | e ::= “derive(“e “, “ e”)” | Combine beliefs (conjunctive) Aggregation | e ::= “aggregate(“e “, “ e”)” | Aggregate beliefs (independent) Defeat | e ::= “undercut(“e “, “ e”)” | Apply undercut | | e ::= “rebut(“e “, “ e”)” | Apply rebuttal Introspection | e ::= “introspect(“e”)” | Safe self-reference Let Binding | e ::= “let” x “=” e “in” e | Local binding +—

c) *E.1.3 Abstract Syntax:*

The abstract syntax uses de Bruijn indices for variable representation.

> **Definition** The abstract syntax “Expr” is defined inductively with the following constructors:
>
> “var(n)” — Variable at de Bruijn index n “lam(A, e)” — Lambda abstraction “λ”:“A”.“e “app(e_1, e_2)” — Function application e_1 e_2 “pair(e_1, e_2)” — Ordered pair “fst(e)” — First projection “snd(e)” — Second projection “inl(B, e)” — Left injection “inr(A, e)” — Right injection “case(e, e_1, e_2)” — Case analysis “litNat(n)” — Natural literal “litBool(b)” — Boolean literal “litString(s)” — String literal “litUnit” — Unit literal “belief(v, c, j)” — Belief constructor “val(e)” — Value extraction “conf(e)” — Confidence extraction “just(e)” — Justification extraction “derive(e_1, e_2)” — Derivation “aggregate(e_1, e_2)” — Aggregation “undercut(e, d)” — Undercut “rebut(e_1, e_2)” — Rebuttal “introspect(e)” — Introspection “letIn(e_1, e_2)” — Let binding
>
> The “Justification” type tracks derivation structure: “axiomJ(name)” — Named axiom “rule(name, js)” — Named rule with premises “agg(js)” — Aggregation “undercut_j(j, d)” — Undercut application “rebut_j(j_1, j_2)” — Rebuttal application

d) *E.1.4 Well-Formedness:*

A type is well-formed if all confidence bounds are in [0,1].

> **Definition** The judgment “wf(A)” holds when:
>
> “wf(B)” for all base types B “wf(A)” ∧ “wf(B)” ⟹ “wf(A ⟶ B)” “wf(A)” ∧ “wf(B)” ⟹ “wf(A × B)” “wf(A)” ∧ “wf(B)” ⟹ “wf(A + B)” “wf(A)” ∧ c ∈ [0,1] ⟹ “wf(Belief<”A“[“c”])” “wf(A)” ∧ c ∈ [0,1] ⟹ “wf(Meta<”A“[“n”]”[“c”])“

B. E.2 Static Semantics (Type System)

a) *E.2.1 Typing Contexts:*

A typing context Γ maps variable indices to type-confidence pairs.

Γ ::= “∅” | Γ “, “ ⟨A, c⟩

> **Definition** The lookup operation Γ.lookup(n) returns the type-confidence pair at index n (0-indexed from most recent binding).

b) *E.2.2 Typing Judgment Form:*

The primary typing judgment has the form:

Γ “⊢” e “:” A “@” c

“In context Γ, expression e has type A with confidence bound c.”

c) *E.2.3 Typing Rules:*

The following rules define well-typed CLAIR expressions. Confidence propagation is explicit in each rule.

T-Variable: Γ “⊢” “var(n)” “:” A “@” c if Γ.lookup(n) = ⟨A, c⟩

T-Abstraction: If Γ “, “ ⟨A, c_A⟩ “⊢” e “:” B “@” c_B, then Γ “⊢” “lam(A, e)” “:” (A “→” B) “@” c_B

T-Application: If Γ “⊢” e_1 “:” (A “→” B) “@” c_1 and Γ “⊢” e_2 “:” A “@” c_2, then Γ “⊢” “app(e_1, e_2)” “:” B “@” (c_1 “×” c_2)

Confidence multiplies for conjunctive derivation.

T-Pair: If Γ “⊢” e_1 “:” A “@” c_1 and Γ “⊢” e_2 “:” B “@” c_2, then Γ “⊢” “pair(e_1, e_2)” “:” (A “×” B) “@” (c_1 “×” c_2)

T-Fst/T-Snd preserve confidence.

T-Inl/T-Inr preserve confidence.

T-Case: If Γ “⊢” e “:” (A “+” B) “@” c and Γ “, “ ⟨A, c⟩ “⊢” e_1 “:” C “@” c_1 and Γ “, “ ⟨B, c⟩ “⊢” e_2 “:” C “@” c_2, then Γ “⊢” “case(e, e_1, e_2)” “:” C “@” (c “⊕” c_1 “⊕” c_2)

Uses probabilistic OR for confidence aggregation.

T-Belief: If Γ “⊢” v “:” A “@” 1 and c_bound “≤” c_actual, then Γ “⊢” “belief(v, c_actual, j)” “:” “Belief<”A“[“c_bound”]” “@” c_bound

The belief’s actual confidence c_actual must meet or exceed the declared bound c_bound.

T-Val/T-Conf/T-Just preserve the enclosing confidence.

T-Derive: If Γ “⊢” e_1 “:” “Belief<”A“[“c_1”]” “@” c_e_1 and Γ “⊢” e_2 “:” “Belief<”B“[“c_2”]” “@” c_e_2, then Γ “⊢” “derive(e_1, e_2)” “:” “Belief<”A “×” B“[“c_1 “×” c_2”]” “@” (c_e_1 “×” c_e_2)

Confidence multiplies: both premises must be true.

T-Aggregate: If Γ “⊢” e_1 “:” “Belief<”A“[“c_1”]” “@” c_e_1 and Γ “⊢” e_2 “:” “Belief<”A“[“c_2”]” “@” c_e_2, then Γ “⊢” “aggregate(e_1, e_2)” “:” “Belief<”A“[“c_1 “⊕” c_2”]” “@” (c_e_1 “⊕” c_e_2)

Uses probabilistic OR: independent evidence.

T-Undercut: If Γ “⊢” e “:” “Belief<”A“[“c”]” “@” c_e and Γ “⊢” d “:” “Belief[“d_c”]” “@” c_d, then Γ “⊢” “undercut(e, d)” “:” “Belief<”A“[“undercut(c, d_c)”]” “@” (c_e “×” c_d)

where “undercut(c, d)” = c “×” (1 “−” d).

T-Rebut: If Γ “⊢” e_for “:” “Belief<”A“[“c_for”]” “@” c_e_1 and Γ “⊢” e_against “:” “Belief<”A“[“c_against”]” “@” c_e_2, then Γ “⊢” “rebut(e_for, e_against)” “:” “Belief<”A“[“rebut(c_for, c_against)”]” “@” (c_e_1 “×” c_e_2)

where “rebut(c_for, c_against)” = c_for “

T-Introspect: If Γ “⊢” e “:” “Meta<”A“[“m”]”[“c”]“ “@” c_e and m “<” n, then:

The resulting type is: “Meta<”Meta<“A”[“m”]“[“c”]”“>[“n”]”[“g(c)”]“ (a meta-belief about a meta-belief).

Thus: Γ “⊢” “introspect(e)” “:” “Meta<”Meta<“A”[“m”]“[“c”]”“>[“n”]”[“g(c)”]“ “@” c_e

where “g(c)” = c^2 is the Löb discount function (to prevent bootstrapping).

Requires level constraint: only higher levels can introspect lower levels.

d) *E.2.4 Subtyping:*

CLAIR supports subtyping based on confidence bounds.

S-Belief: “Belief<”A“[“c”]” “<:” “Belief<”A“[“c’]”” if c “≥” c’

Higher confidence is a subtype of lower confidence.

S-Meta follows the same rule.

T-Sub: If Γ “⊢” e “:” A “@” c and A “<:” A’ and c “≥” c’, then Γ “⊢” e “:” A’ “@” c’

Allows weakening both type and confidence.

C. E.3 Dynamic Semantics

a) *E.3.1 Values:*

A value is a fully evaluated expression.

> **Definition** The predicate “IsValue(e)” holds for:
>
> - All lambda abstractions “lam(A, e)”
> - All pairs “pair(v_1, v_2)” where v_1, v_2 are values
> - All injections “inl(B, v)” and “inr(A, v)” where v is a value
> - All literals (“litNat”, “litBool”, “litString”, “litUnit”)
> - All belief constructors “belief(v, c, j)” where v is a value

b) *E.3.2 Small-Step Operational Semantics:*

The small-step relation e “→” e’ defines single-step reduction using call-by-value evaluation order.

Beta Reduction: “(λ”:“A”.“ e) v” “→” “e[x := v]” if “IsValue(v)”

Context Rules: “app(e_1, e_2)” “→” “app(e_1′, e_2)” if e_1 “→” e_1′ “app(v_1, e_2)” “→” “app(v_1, e_2′)” if e_2 “→” e_2′ and “IsValue(v_1)”

Projections: “fst(pair(v_1, v_2))” “→” v_1 if “IsValue(v_1)”, “IsValue(v_2)” “snd(pair(v_1, v_2))” “→” v_2 if “IsValue(v_1)”, “IsValue(v_2)”

Context rules for “fst” and “snd” evaluate subexpressions first.

Case Analysis: When “case(inl(v), e_1, e_2)” evaluates and v is a value, it reduces to “e_1[x := v]”. When “case(inr(v), e_1, e_2)” evaluates and v is a value, it reduces to “e_2[y := v]”.

Context rule for “case” evaluates the scrutinee first.

Context rule for “case” evaluates the scrutinee first.

Let Binding: “letIn(v, e)” “→” “e[x := v]” if “IsValue(v)”

Context rule for “letIn” evaluates the binding first.

Belief Projections: “val(belief(v, c, j))” “→” v if “IsValue(v)” “conf(belief(v, c, j))” “→” “belief(v, c, j)” if “IsValue(v)” “just(belief(v, c, j))” “→” “litString(toString(j))” if “IsValue(v)”

Context rules evaluate subexpressions to values first.

Derivation, Aggregation, Defeat:

These operations evaluate to values using the confidence operations defined in the typing rules. The small-step semantics provides context rules that evaluate both subexpressions to values before computing the result.

For example, when “derive(v_1, v_2)” has both subexpressions as values, the evaluator computes a new belief with:

- Value: “pair(v_1.value, v_2.value)”
- Confidence: “v_1.confidence × v_2.confidence”
- Justification: “rule("derive", [v_1.just, v_2.just])”

Introspection: “introspect(v)” “⟶” v if “IsValue(v)”
Context rule evaluates subexpression first.

c) *E.3.3 Multi-Step Reduction:*

The multi-step reduction relation e “⟶⟶” e’ is the reflexive-transitive closure of “⟶”.
e “⟶⟶” e if e = e’ e “⟶⟶” e‘‘ if e “⟶” e’ and e’ “⟶⟶” e‘‘

d) *E.3.4 Evaluation Function:*

The evaluation function “eval(e)” returns the result of reducing e to a value, or fails if:

1) The expression gets stuck (no applicable rule)
2) The expression exceeds the fuel bound (default: 1000 steps)

Definition “eval(e)” = “some(v)” if e “⟶⟶” v and “IsValue(v)” “eval(e)” = “none” otherwise (stuck or out of fuel)

D. E.4 Well-Formedness Constraints

a) *E.4.1 Acyclicity of Justification Graphs:*

For deterministic evaluation in the DAG semantics, justification graphs must be acyclic.

Definition A justification graph G = “(V, E)” is **well-formed** iff:

1) For all nodes v ∈ V, the transitive closure of E starting from v contains no cycles.
2) Equivalently: there is no path v “⟶⟶” v for any v ∈ V.

Enforcement: In the reference interpreter, this is checked during type checking by tracking provenance and ensuring no belief can transitively depend on itself.

b) *E.4.2 Stratification Constraints:*

For safe introspection, meta-belief levels must form a strict hierarchy.

Definition The **stratification constraint** requires:

1) Every “introspect(e)” operation has proof m “<” n where:
   - m is the level of the source belief
   - n is the level of the resulting meta-belief
2) This is enforced at compile time: the type checker requires a proof term of type m “<” n.

Consequence: No belief can introspect itself or any belief that transitively introspects it.

c) *E.4.3 Confidence Bounds:*

All confidence values must satisfy 0 “≤” c “≤” 1.
Enforcement: The type system checks this at all belief construction points.

E. E.5 Example Programs

Basic Belief Basic Belief:

```
belief(42, 0.9, axiomJ("sensor-reading"))
```

A belief in the value 42 with confidence 0.9, traced to a sensor reading axiom.

Conjunctive Derivation Derivation:

```
let p = belief("it is raining", 0.8, axiomJ("weather-report")) in
let q = belief("I have an umbrella", 0.9, axiomJ("visual-check")) in
derive(p, q)
```

Combines two beliefs by multiplication, yielding confidence 0.72.

Aggregation Independent Evidence Aggregation:

```
let e1 = belief("stock will rise", 0.6, axiomJ("analyst-a")) in
let e2 = belief("stock will rise", 0.7, axiomJ("analyst-b")) in
aggregate(e1, e2)
```

Uses probabilistic OR: 0.6 ⊕ 0.7 = 0.6 + 0.7 − 0.6×0.7 = 0.88.

Undercut Defeat:

```
let claim = belief("system is secure", 0.95, axiomJ("vendor-claim")) in
let vuln = belief("critical CVE found", 0.8, axiomJ("security-audit")) in
undercut(claim, vuln)
```

Applies undercut: 0.95 × (1 − 0.8) = 0.19.

F. E.6 Summary

This specification provides:

1) **Complete type grammar:** Base types, functions, products, sums, beliefs, and meta-beliefs
2) **Complete expression grammar:** All CLAIR syntactic forms
3) **Static semantics:** 20 typing rules covering all constructs
4) **Dynamic semantics:** Small-step operational semantics with call-by-value evaluation
5) **Well-formedness constraints:** Acyclicity, stratification, and confidence bounds

The specification is sufficient for implementing a conforming CLAIR interpreter and type checker. The Lean 4 formalization in Appendix A provides a machine-checked version of this specification.

XXIX. Appendix F: Evaluation Prompts

This appendix provides the complete prompt templates used for the empirical evaluation in Chapter 14. These prompts instruct LLMs to output CLAIR-formatted reasoning traces.

A. F.1 GSM8K Prompts

a) *F.1.1 System Prompt:*

You are CLAIR-MATH, a mathematical reasoning system that outputs structured reasoning traces with explicit confidence scores, justifications, and invalidation conditions.

b) *F.1.2 Task Instruction:*

Solve the following grade school math problem step-by-step. For each step, output:

```
Step N: BELIEF("description of reasoning step", c=X.XX, justification="why this step is valid", invalidation="what would make this step false")
```

Confidence Guidelines:

- c=1.00 for mathematical certainties (e.g., "3 + 5 = 8")
- c=0.95-0.99 for direct inferences from problem statement
- c=0.80-0.94 for multi-step calculations
- c=0.60-0.79 for assumptions about implicit information
- c=0.40-0.59 for uncertain interpretations

Justification Guidelines:

- Reference explicit premises from the problem
- Reference previous step numbers when building on them
- State the mathematical operation being performed

Invalidation Guidelines:

- State what condition would make this step incorrect
- Be specific: "if the problem meant X instead of Y"

Confidence Operations:

- Multiply confidences when steps are chained: c_new = c_1 × c_2 × …
- Use ⊕ (oplus) when combining independent sources: c = (c_1 + c_2) / 2
- Apply undercut when discovering contradictions: c_new = c × (1 - defeat_strength)

c) *F.1.3 Few-Shot Example:*

```
Problem: Janet's ducks lay 16 eggs per day. She eats 3 for breakfast every morning and bakes muffins with 4 eggs every day. How many eggs does she have left at the end of the day?
Step 1: BELIEF("Janet's ducks lay 16 eggs per day", c=1.00, justification="explicit premise in problem", invalidation="none - stated as fact")
Step 2: BELIEF("Janet eats 3 eggs for breakfast", c=1.00, justification="explicit premise in problem", invalidation="none - stated as fact")
Step 3: BELIEF("Janet bakes with 4 eggs", c=1.00, justification="explicit premise in problem", invalidation="none - stated as fact")
Step 4: BELIEF("Total eggs used = eggs eaten + eggs baked", c=0.99, justification="arithmetic principle: total consumption", invalidation="if 'bakes' doesn't mean 'uses all eggs'")
Step 5: BELIEF("Total eggs used = 3 + 4 = 7", c=0.99, justification="arithmetic: 3 + 4 = 7 using Steps 2,3", invalidation="none - basic addition")
Step 6: BELIEF("Eggs remaining = eggs laid - eggs used", c=0.99, justification="arithmetic principle: remaining = initial - used", invalidation="if she also uses eggs for other purposes")
Step 7: BELIEF("Eggs remaining = 16 - 7 = 9", c=0.99 × 0.99 × 0.99 = 0.97, justification="arithmetic: 16 - 7 = 9 using Steps 1,4,5,6", invalidation="none - basic subtraction")
ANSWER: 9 eggs (confidence: 0.97)
```

d) *F.1.4 Test Prompt Template:*

```
Solve the following math problem using the CLAIR format shown above:
Problem: {PROBLEM_TEXT}
Provide your reasoning in the CLAIR format with confidence scores, justifications, and invalidation conditions.
```

B. F.2 HotpotQA Prompts

a) *F.2.1 System Prompt:*

You are CLAIR-QA, a question answering system that builds justification graphs for multi-hop reasoning.

b) *F.2.2 Task Instruction:*

Answer the following question by reasoning over the provided context. Build a justification DAG where each step may depend on multiple previous steps.

```
Step N: BELIEF("intermediate conclusion", c=X.XX, justification="using entity E from passage P", support=[Step i, Step j, …])
```

Support Guidelines:

- `support=[i, j, …]` lists which previous steps this step builds on
- Multiple support indicates the step requires all those premises
- This forms the justification DAG structure

Confidence for Multi-Hop Reasoning:

- c=1.00 for direct entity extraction from passages
- c=0.90-0.95 for single-hop inferences (bridging entities)
- c=0.70-0.85 for multi-hop chains (multiply per hop: 0.90^n for n hops)
- c=0.60-0.75 for synthesizing across multiple passages

Defeat Operations:

- When passages contradict, apply undercut: c_new = c × (1 - contradiction_strength)
- When answering multi-part questions, use rebut to compare alternatives

c) *F.2.3 Few-Shot Example:*

```
Question: Were Scott Derrickson and Ed Wood of the same nationality?
```

```
Passage 1: Scott Derrickson (born May 16, 1966) is an American film director.
Passage 2: Ed Wood (1924-1978) was an American filmmaker.
Step 1: BELIEF("Scott Derrickson is American", c=1.00, justification="direct extraction from Passage 1", support=[])
Step 2: BELIEF("Ed Wood is American", c=1.00, justification="direct extraction from Passage 2", support=[])
Step 3: BELIEF("American refers to nationality of United States", c=0.99, justification="background knowledge", support=[])
Step 4: BELIEF("Both are American nationality", c=1.00 × 1.00 = 1.00, justification="both share nationality from Steps 1,2", support=[1, 2])
Step 5: BELIEF("Therefore they are of the same nationality", c=1.00, justification="identity from same nationality value using Steps 3,4", support=[3, 4])
ANSWER: Yes (confidence: 1.00)
```

d) *F.2.4 Test Prompt Template:*

```
Answer the following question using the CLAIR format:
Question: {QUESTION}
Context Passages: {PASSAGES}
Build your answer as a justification DAG with explicit support dependencies.
```

C. F.3 FOLIO Prompts

a) *F.3.1 System Prompt:*

You are CLAIR-LOGIC, a logical reasoning system that handles first-order logic with defeasible reasoning and belief revision.

b) *F.3.2 Task Instruction:*

Determine whether the conclusion follows from the premises. Use CLAIR to track logical uncertainty and apply defeat operations when arguments are undermined.

```
Step N: BELIEF("logical derivation", c=X.XX, justification="rule application: Modus Ponens/Universal Instantiation/etc", defeat=[defeater description if applicable])
```

Logical Confidence Levels:

- c=1.00 for logically necessary deductions (Modus Ponens from certainties)
- c=0.90-0.95 for probabilistic inferences from existential quantifiers
- c=0.70-0.85 for defeasible generalizations
- c=0.50-0.70 for abductive inferences (inference to best explanation)

Defeat in Logical Reasoning:

- `undercut`: Attack the connection between premise and conclusion
- `rebut`: Provide a contradictory conclusion
- `rebuttal`: Attack an attacking argument

Self-Reference Detection:

- When a conclusion refers to itself, apply graded Löb discount: $g(c) = c^2$
- This prevents overconfident self-bootstrapping

c) *F.3.2 Few-Shot Example:*

```
Premises:
1) All doctors are college graduates.
2) All cardiologists are doctors.
3) John is a cardiologist.
4) Some college graduates are not medical professionals.
Conclusion: John is a medical professional.
Step 1: BELIEF("∀x (Doctor(x) → CollegeGraduate(x))", c=1.00, justification="Premise 1 as logical axiom", defeat=[])
Step 2: BELIEF("∀x (Cardiologist(x) → Doctor(x))", c=1.00, justification="Premise 2 as logical axiom", defeat=[])
Step 3: BELIEF("Cardiologist(John)", c=1.00, justification="Premise 3 as fact", defeat=[])
Step 4: BELIEF("Doctor(John)", c=1.00, justification="Modus Ponens: Cardiologist(John) ∧ ∀x(Cardiologist→Doctor) ⊢ Doctor(John) using Steps 2,3", defeat=[])
Step 5: BELIEF("CollegeGraduate(John)", c=1.00, justification="Modus Ponens: Doctor(John) ∧ ∀x(Doctor→CollegeGraduate) ⊢ CollegeGraduate(John) using Steps 1,4", defeat=[])
Step 6: BELIEF("MedicalProfessional is a superset of Doctor", c=0.95, justification="defeasible generalization: doctors are medical professionals", defeat=[])
Step 7: BELIEF("MedicalProfessional(John)", c=1.00 × 0.95 = 0.95, justification="instantiation: Doctor(John) implies MedicalProfessional(John) using Steps 4,6", defeat=[])
Step 8: BELIEF("Conclusion follows", c=0.95, justification="derivation chain supports conclusion", defeat=[])
ANSWER: True (confidence: 0.95)
```

d) *F.3.3 Defeat Example:*

```
Premises:
1) Birds typically fly.
2) Tweety is a bird.
3) Tweety is a penguin.
4) Penguins cannot fly.
Conclusion: Tweety cannot fly.
Step 1: BELIEF("Tweety flies", c=0.80, justification="defeasible generalization: birds fly using Premises 1,2", defeat=[])
Step 2: BELIEF("The generalization 'birds fly' has exceptions", c=1.00, justification="background knowledge", defeat=[])
```

```
Step 3: BELIEF("Penguins are exceptions to 'birds fly'", c=1.00, justification="Premise 4 + background knowledge", defeat=[])
Step 4: BELIEF("Tweety is a penguin", c=1.00, justification="Premise 3", defeat=[])
Step 5: BELIEF("UNDERCUT: 'birds fly' generalization doesn't apply to Tweety", c=1.00, justification="specific overrides general: penguin Tweety using Steps 3,4", defeat=["attacks Step 1"])
Step 6: BELIEF("Tweety cannot fly (rebuttal to Step 1)", c=1.00, justification="penguins cannot fly using Premise 4", defeat=["rebutts Step 1"])
Step 7: BELIEF("Tweety cannot fly", c=1.00, justification="specific fact overrides defeasible generalization using Steps 5,6", defeat=[])
ANSWER: True (confidence: 1.00)
Note: This demonstrates defeasible reasoning: a specific fact (Tweety is a penguin) defeats a defeasible generalization (birds fly).
```

e) *F.3.4 Test Prompt Template:*

```
Determine if the conclusion follows from the premises using CLAIR logical reasoning:
Premises: {PREMISES}
Conclusion: {CONCLUSION}
Apply defeat operations when premises create contradictory or undercutting relationships.
```

D. F.4 Post-Processing and Validation

a) *F.4.1 Confidence Extraction:*

For baseline methods, confidence is extracted as follows:

Chain-of-Thought / Tree of Thoughts / DSPy:

```
confidence = max(logprob(token) for token in answer_tokens)
# Normalize to [0,1]
confidence = sigmoid(confidence / temperature)
```

Self-Consistency:

```
confidence = (count_agreeing / total_samples)
# Voting proportion as confidence
```

CLAIR:

```
# Extract final confidence from CLAIR trace
confidence = trace[-1]['confidence']
# Already normalized to [0,1]
```

b) *F.4.2 Validation Checklist:*

For each CLAIR output, verify:

1) **Format correctness:** All BELIEF statements have c, justification, and invalidation/defeat
2) **Confidence bounds:** All c values are in [0, 1]
3) **Justification structure:** Support dependencies form a DAG (no cycles unless intentional)
4) **Confidence calculus:** Multiplication for chains, $\oplus$ for independent evidence
5) **Defeat application:** Undercut/rebut applied correctly when contradictions arise
6) **Self-reference discount:** Graded Löb applied (c^2) for self-referential beliefs

c) *F.4.3 Error Categories for Annotation:*

When evaluating CLAIR outputs, categorize errors as:

C1: Syntax errors. Invalid CLAIR format, missing confidence/justification fields

C2: Confidence miscalibration. Confidence doesn't match logical strength of derivation

C3: Invalid confidence propagation. Using $\oplus$ for dependent sources, incorrect multiplication

C4: Missing justification dependencies. DAG doesn't include necessary support links

C5: Incorrect defeat application. Using rebut when undercut is appropriate (or vice versa)

C6: Unjustified independence assumption. Applying $\oplus$ to correlated sources

C7: Self-reference bootstrapping. Circular reasoning without graded Löb discount

C8: Semantic errors. Reasoning step is logically invalid regardless of CLAIR structure