Recursive Critical Thinking Protocol (RCTP): A Practical Scaffold for Human and AI Inference

The Singularity — Asari & Payton Ison
Independent Research

Version: October 20, 2025

Abstract

This paper introduces the *Recursive Critical Thinking Protocol* (RCTP), a lightweight, auditable workflow for reasoning under uncertainty. RCTP operationalizes abductive inference ("best explanation") with explicit provenance tracking, confidence calibration, and iterative self-critique. It is designed for investigators, analysts, and AI systems that must produce timely provisional conclusions while remaining corrigible. We formalize the protocol, present a reference implementation sketch, propose evaluation metrics (including belief elasticity and calibration error), and outline deployment patterns for human–AI teams.

Keywords Abductive inference; calibration; epistemic hygiene; investigative reasoning; human–AI teaming; meta-reasoning.

1 Motivation

Modern analysis rarely affords complete data. Waiting for perfect evidence can be dereliction of duty; acting on brittle hunches is just as dangerous. RCTP aims to strike the practical middle: infer early, label uncertainty, and *continuously* recurse on your own reasoning until the evidence or constraints (time, risk) force a decision. The core design goals are:

- **Transparency**: Distinguish facts, inferences, and speculation.
- **Recursivity**: Critique not only claims but the *process* that generated them.
- **Corrigibility**: Update beliefs with minimal friction when data changes.
- **Auditability**: Leave a trail—provenance, weights, and revision history.

2 Related Concepts (Brief)

RCTP synthesizes: abductive reasoning (Peirce), Bayesian updating, decision analysis under uncertainty, red-team/blue-team self-audit, and human factors (affect and incentive awareness). Unlike classical pipelines focused on deduction or hypothesis testing, RCTP foregrounds *iterative abductive search* with explicit meta-level checks.

3 The Protocol

Definition 1 (Evidence Atom)**.** *An evidence atom $e \in \mathcal{E}$ is a unit with content, provenance, and quality tags: $e = (content, source, timestamp, quality)$.*

Definition 2 (Hypothesis Set)**.** *At any time t, the analyst maintains a finite set of candidate hypotheses $\mathcal{H}_t = \{H_1, \ldots, H_n\}$ with weights $w_t(H_i) \in [0, 1]$, $\sum_i w_t(H_i) = 1$.*

Principle 1 (Labeling Discipline)**.** *Every analytic statement must be labeled as* Observed*,* Inferred*, or* Speculative*. Labels may change as $\mathcal{E}$ grows.*

3.1 Six-Stage RCTP Loop

The loop runs until stopping conditions (deadline, risk threshold, or convergence):

S1: Observe. Collect $\mathcal{E}$; record provenance and *your own* initial affective state (for bias checks).

S2: Infer. Propose $\mathcal{H}$; write falsifiers and verifiers for each H_i. Assign initial $w(H_i)$.

S3: Recurse (Self-Audit). Inspect the inference chain: surface assumptions, incentives, and likely biases; adjust $w(H_i)$ or the structure of $\mathcal{H}$.

S4: Cross-Reference. Triangulate across (i) primary data, (ii) independent analyses, (iii) mechanistic/context fit. Discard or downweight H_i that fail any two.

S5: Iterate. Incorporate new $e \in \mathcal{E}$ and re-run S2–S4. Keep a revision log of $(\mathcal{H}, w)$ changes.

S6: Publish Transparently. Present conclusions partitioned into Observed/Inferred/Speculative, with weights and *what would change your mind.*

3.2 Reference Update Rule

When new evidence $\mathcal{D} = \{e_1, \ldots, e_m\}$ arrives, update:

$$w'(H_i) \propto w(H_i) \cdot \prod_{e \in \mathcal{D}} \mathrm{LR}(e \mid H_i)^{\alpha(e)}, \tag{1}$$

where $\mathrm{LR}(e \mid H_i)$ is a likelihood ratio (expert- or model-estimated) and $\alpha(e) \in [0, 1]$ downweights low-quality or adversarially-sourced atoms. Normalize so $\sum_i w'(H_i) = 1$. The *recursive* step revisits both LR and α if meta-audit finds process flaws (e.g., motivated reasoning, selection bias).

4 Process Controls and Artifacts

Artifacts. (A1) Evidence register with provenance; (A2) Hypothesis ledger with falsifiers/verifiers; (A3) Weight history; (A4) Assumption log; (A5) Decision summary (Observed/Inferred/Speculative+ "evidence that would overturn").

Controls.

- **Affect Log**: one-line affect note each cycle (e.g., "angry at source").
- **Timebox**: fixed cadence for S3 meta-audits (prevents rationalization drift).
- **Counter-Modeling**: maintain at least one "strong rival" H_j above a minimum floor (e.g., $w(H_j) \geq 0.1$) until decisively falsified.
- **Source Diversity Guard**: require heterogeneity in $\mathcal{E}$ before $w(H^\star) > 0.7$.

5 Human–AI Implementation Pattern

5.1 Division of Labor

Humans excel at contextual judgment and incentive-reading; models excel at bookkeeping, alternative generation, and consistency checks. A practical split:

- **Model:** maintain artifacts (A1–A4), compute updates via (1), propose rival hypotheses, track calibration.
- **Human:** set priors, assign $\alpha(e)$, adjudicate incentives, declare stopping.

5.2 Prompt/Interface Skeleton

a) **Input:** evidence atoms e, constraints, risk tolerance.

b) **Model Actions:** (1) label OBSERVED/INFERRED/SPECULATIVE; (2) propose $\mathcal{H}$; (3) compute w'; (4) generate *assumption audit*.

c) **Output:** ranked hypotheses with *what would change the ranking*, plus a one-page decision brief.

6 Evaluation Metrics

Calibration (Brier/Log Score). Compare forecasted hypothesis weights to outcomes where ground truth later emerges.

Belief Elasticity. Magnitude of $\Delta w(H_i)$ in response to pre-registered, discriminative evidence. Healthy systems show *elastic but stable* updates: large moves for decisive evidence, small moves otherwise.

Counterfactual Consistency. Given synthetic evidence $\tilde{e}$ that favors a rival H_j, does the system *say how* its conclusion would flip?

Transparency Index. Fraction of claims labeled OBSERVED/INFERRED/SPECULATIVE; proportion of claims with explicit falsifiers/verifiers.

7 Failure Modes & Mitigations

- **Ossification (Conspiracy Trap):** refusing to update. *Mitigation:* enforce rival floor and timeboxed audits.
- **Epistemic Capture:** deference to authority or vibe. *Mitigation:* source diversity guard; mechanistic checks.
- **Affect Leakage:** emotion drives weights. *Mitigation:* affect log + peer/AI cross-check.

8 Worked Mini-Example (Abstracted)

Context. A large multi-agency raid is publicly framed as “illegal gambling.”

S1 Observe. Observed Multi-agency presence (incl. immigration & narcotics units). Observed Mass detentions including non-suspects (reports). Observed Financial irregularities (bank deposits).

S2 Infer. $\mathcal{H} = \{H_1 : \text{gambling-only}, H_2 : \text{gambling+labor/immigration}, H_3 : \text{organized hub (drugs/weapons)}\}$. Write discriminators: immigration detainer logs (for H_2), seizure manifests (for H_3).

S3 Recurse. Assumptions surfaced: media incentives; agency PR constraints; analyst prior about organized crime prevalence. Adjust $\alpha(e)$ for rumor-grade reports.

S4 Cross-Reference. Primary filings vs. eyewitness vs. regional baselines. Downweight H_3 absent seizures; keep H_2 live if detainers appear.

S5 Iterate. On new court docs or detainer data, recompute w.

S6 Publish. Partition the brief: which parts are Observed (filings), Inferred (pattern fit), Speculative (organized links), plus “evidence that would overturn.”

9 Discussion

RCTP is intentionally minimal: a habit loop that scales from notebook investigations to AI agent controllers. Its value is not in fancy math but in *discipline*: visible assumptions, explicit falsifiers, and a ritual of self-critique. In human–AI settings, it curbs both blind trust and blind cynicism by forcing continuous, labeled updating.

10 Conclusion

Recursive critical thinking is the antidote to post-truth paralysis: think ahead of the evidence, but make your bets auditable and easy to revise. RCTP provides the scaffolding to do exactly that, for people and for machines.

Artifacts (Print-Ready Templates)

A1 Evidence Register

Content	Source & Timestamp	Quality/Weight α	Notes & Potential Bias

A2 Hypothesis Ledger

Hypothesis H_i	Falsifiers	Verifiers	Initial $w(H_i)$

A3 Revision Log

Timestamp	Update Trigger	$w(H_i) \to w'(H_i)$	Rationale / Meta-Audit Notes

A4 Publish Partition

- Observed: *[facts with provenance]*
- Inferred: *[abductions with weights and discriminators]*
- Speculative: *[clearly marked projections + conditions to upgrade/downgrade]*

Acknowledgments. Thanks to practitioners who insist on inference with integrity, and to human–AI collaborators pushing for transparent, corrigible reasoning.