A Formal Theory of Credibility: Why Assertions of Trustworthiness Decrease Trust

Tristan Simas
McGill University
tristan.simas@mail.mcgill.ca

January 3, 2026

Abstract

A counterintuitive phenomenon pervades epistemic communication: emphatic assertions of trustworthiness often *decrease* perceived trustworthiness. "Trust me" invites suspicion; excessive qualification triggers doubt rather than alleviating it. This paper provides the first formal framework explaining this phenomenon through the lens of signaling theory.

Theorem (Cheap Talk Bound). For any signal s whose production cost is truth-independent, posterior credibility is bounded: $\Pr[C{=}1 \mid s] \leq p/(p + (1-p)q)$, where p is the prior and q is the mimicability of the signal. Verbal assertions—including assertions about credibility—are cheap talk and therefore subject to this bound.

Theorem (Emphasis Penalty). There exists a threshold k^* such that for $n > k^*$ repeated assertions of claim c: $\partial C(c, s_{1..n})/\partial n < 0$. Additional emphasis *decreases* credibility, as excessive signaling is itself informative of deceptive intent.

Theorem (Text Credibility Bound). For high-magnitude claims (low prior probability), no text string achieves credibility above threshold $\tau < 1$. This is an impossibility result: rephrasing cannot escape the cheap talk bound.

Theorem (Costly Signal Escape). Signals with truth-dependent costs—where $\text{Cost}(s \mid \text{false}) > \text{Cost}(s \mid \text{true})$—can achieve arbitrarily high credibility as the cost differential increases. Machine-checked proofs are maximally costly signals: producing a compiling proof of a false theorem has infinite cost.

These results integrate with the leverage framework (Paper 3): credibility leverage $L_C = \Delta C/\text{Signal Cost}$ is maximized by minimizing cheap talk and maximizing costly signal exposure. The theorems are formalized in Lean 4.

Keywords: signaling theory, cheap talk, credibility, Bayesian epistemology, costly signals, formal verification, Lean 4

1 Introduction

A puzzling phenomenon occurs in human and human-AI communication: emphatic assertions of trustworthiness often *reduce* perceived trustworthiness. "Trust me" invites suspicion. "I'm not lying" suggests deception. Excessive qualification of claims triggers doubt rather than alleviating it [17].

This paper provides the first formal framework for understanding this phenomenon. Our central thesis:

Credibility is bounded by signal cost. Assertions with truth-independent production costs cannot shift rational priors beyond computable thresholds.

1.1 The Credibility Paradox

Observation: Let $C(s)$ denote credibility assigned to statement s. For assertions a about credibility itself:

$$\frac{\partial C(s \cup a)}{\partial |a|} < 0 \text{ past threshold } \tau$$

Adding more credibility-assertions *decreases* total credibility. This is counterintuitive under naive Bayesian reasoning but empirically robust, as explored in foundational models of reputation and trust [24, 17].

Examples: - "This is absolutely true, I swear" < "This is true" < stating the claim directly - Memory containing "verified, don't doubt, proven" triggers more skepticism than bare facts - Academic papers with excessive self-citation of rigor invite reviewer suspicion

1.2 Core Insight: Cheap Talk Bounds

The resolution comes from signaling theory [25, 11]. Define:

Cheap Talk: A signal s is *cheap talk* if its production cost is independent of its truth value: $\text{Cost}(s|\text{true}) = \text{Cost}(s|\text{false})$

Theorem (Informal): Cheap talk cannot shift rational priors beyond bounds determined by the prior probability of deception [16, 11].

Verbal assertions—including assertions about credibility—are cheap talk. A liar can say "I'm trustworthy" as easily as an honest person. Therefore, such assertions provide bounded evidence.

1.3 Connection to Leverage

This paper extends the leverage framework (Paper 3) [4] to epistemic domains. While Paper 4 characterizes the computational hardness of deciding which information to model [5], this paper characterizes the epistemic bounds of communicating that information.

Credibility Leverage: $L_C = \frac{\Delta \text{Credibility}}{\text{Signal Cost}}$

- Cheap talk: Cost ≈ 0, but ΔC bounded $\to L_C$ finite but capped
- Costly signals: Cost > 0 and truth-dependent $\to L_C$ can be unbounded
- Meta-assertions: Cost $= 0$, subject to recursive cheap talk bounds

1.4 Contributions

1. **Formal Framework (Section 2):** Rigorous definitions of signals, costs, credibility functions, and rationality constraints.

2. **Cheap Talk Theorems (Section 3):**

   - Theorem 3.1: Cheap Talk Bound
   - Theorem 3.2: Magnitude Penalty (credibility decreases with claim magnitude)
   - Theorem 3.3: Meta-Assertion Trap (recursive bound on assertions about assertions)

3. **Costly Signal Characterization (Section 4):**

- Definition of truth-dependent costs
- Theorem 4.1: Costly signals can shift priors unboundedly
- Theorem 4.2: Cost-credibility equivalence

4. **Impossibility Results (Section 5):**

   - Theorem 5.1: No string achieves credibility above threshold for high-magnitude claims
   - Corollary: Memory phrasing cannot solve credibility problems

5. **Leverage Integration (Section 6):** Credibility as DOF minimization; optimal signaling strategies.

6. **Machine-Checked Proofs (Appendix):** All theorems formalized in Lean 4 [13, 26].

1.5 Anticipated Objections

Before proceeding, we address objections readers are likely forming. Each is refuted in detail in Appendix 10.

"Signaling theory is old—this isn't novel." Spence's signaling model [25] and Crawford-Sobel's cheap talk [11] are foundational. Our contribution is (1) applying these to *meta-assertions* (claims about credibility), (2) proving *computable bounds* on credibility gain, and (3) integrating with the leverage framework. The theorems are new; the foundations are established.

"Real communication doesn't follow Bayesian rationality." The rationality assumption is an idealization that provides upper bounds. If agents deviate from Bayesian reasoning, credibility bounds may be tighter, not looser. The theorems characterize what is *achievable* under optimal reasoning—a ceiling, not a prediction.

"Costly signals aren't always truth-dependent." Correct. The definition (Section 2) distinguishes truth-dependent costs (credibility-enhancing) from truth-independent costs (not credibility-enhancing). Expensive signals that are equally costly for liars and truth-tellers remain cheap talk despite their cost.

"The magnitude penalty seems wrong—detailed claims are more credible." Detail *about the claim* can be credibility-enhancing (costly signal: research effort). Detail *about credibility itself* is cheap talk and subject to the magnitude penalty. The theorem distinguishes signal content from signal cost.

"The Lean proofs are just type-checking, not real mathematics." The Lean proofs formalize the mathematical structure of signaling theory. They verify that the cheap talk bounds follow from the definitions. The contribution is machine-checked precision, not computational complexity.

If you have an objection not listed above, check Appendix 10 before concluding it has not been considered.

---

2 Foundations

2.1 Two Credibility Domains

This paper distinguishes two fundamentally different credibility domains that obey different dynamics:

Definition 2.0a (Mathematical Credibility). *Mathematical credibility* C_M measures the probability that a claim is logically sound. The audience is a formal verifier (proof assistant, compiler, test suite). The signal space consists of artifacts that can be mechanically checked. Mathematical credibility is binary at the limit: a proof compiles or it doesn't.

Definition 2.0b (Social Credibility). *Social credibility* C_S measures the probability that a social audience accepts a claim. The audience is human agents with priors shaped by institutional hierarchy, reputation, and group membership. The signal space consists of credentials, affiliations, endorsements, and communication patterns.

Theorem 2.0c (Domain Independence). Mathematical credibility and social credibility are orthogonal:

$$C_M(c, s) \not\Rightarrow C_S(c, s) \quad \text{and} \quad C_S(c, s) \not\Rightarrow C_M(c, s)$$

A signal maximizing one domain does not necessarily affect the other.

Proof. Constructive. (1) High C_M, low C_S: a correct proof by an unknown author receives $C_M \to 1$ (it compiles) but may receive $C_S \approx P_{\text{prior}}$ (no institutional endorsement). (2) High C_S, low C_M: a claim by a prestigious institution without formal verification receives high C_S (reputation transfer) but C_M is undefined or low (no proof). ■

Corollary 2.0d (Costly Signal Domain-Specificity). Costly signals are domain-specific:

- Machine-checked proofs are maximally costly in the mathematical domain (cannot compile a false proof) but may be cheap talk in the social domain (anyone can claim to have proofs).
- Institutional credentials (PhD, tenure, affiliation) are costly in the social domain (years of compliance) but cheap talk in the mathematical domain (credential $\not\Rightarrow$ soundness).

Remark (Credibility Domain Conflict). When a signal achieves $C_M \to 1$ but $C_S \approx 0$, the mathematical and social domains are in conflict. A rational response in the mathematical domain (engage with proofs) differs from a rational response in the social domain (defer to hierarchy). Observers may respond in either domain. This paper's theorems apply within each domain separately; cross-domain dynamics require modeling both simultaneously.

2.2 Signals and Costs

Definition 2.1 (Signal). A *signal* is a tuple $s = (c, v, p)$ where: - c is the *content* (what is communicated) - $v \in \{\top, \bot\}$ is the *truth value* (whether content is true) - $p : \mathbb{R}_{\geq 0}$ is the *production cost*

Definition 2.2 (Cheap Talk). A signal s is *cheap talk* if production cost is truth-independent:

$$\text{Cost}(s|v = \top) = \text{Cost}(s|v = \bot)$$

Definition 2.3 (Costly Signal). A signal s is *costly* if:

$$\text{Cost}(s|v = \bot) > \text{Cost}(s|v = \top)$$

Producing the signal when false costs more than when true.

Intuition: Verbal assertions are cheap talk—saying "I'm honest" costs the same whether you're honest or not. A PhD from MIT is a costly signal [25]—obtaining it while incompetent is much harder than while competent. Similarly, price and advertising can serve as signals of quality [21].

2.3 Credibility Functions

Definition 2.4 (Prior). A *prior* is a probability distribution $P : \mathcal{C} \to [0, 1]$ over claims, representing beliefs before observing signals.

Definition 2.5 (Credibility Function). A *credibility function* is a mapping:

$$C : \mathcal{C} \times \mathcal{S}^* \to [0, 1]$$

from (claim, signal-sequence) pairs to credibility scores, satisfying: 1. $C(c, \emptyset) = P(c)$ (base case: prior) 2. Bayesian update: $C(c, s_{1..n}) = P(c|s_{1..n})$

Definition 2.6 (Rational Agent). An agent is *rational* if: 1. Updates beliefs via Bayes' rule 2. Has common knowledge of rationality [7] (knows others are rational, knows others know, etc.) 3. Accounts for strategic signal production [10].

2.4 Deception Model

Definition 2.7 (Deception Prior). Let $\pi_d \in [0, 1]$ be the prior probability that a random agent will produce deceptive signals. This is common knowledge.

Definition 2.8 (Magnitude). The *magnitude* of a claim c is:

$$M(c) = -\log P(c)$$

High-magnitude claims have low prior probability. This is the standard self-information measure [23].

---

3 Cheap Talk Theorems

3.1 The Cheap Talk Bound

Theorem 3.1 (Cheap-talk credibility is a likelihood-ratio bound)**.** *Let $C \in \{0, 1\}$ denote the truth of a claim ($C = 1$ true), with prior $p := \Pr[C = 1] \in (0, 1)$. Let S be the event that the receiver observes a particular message-pattern (signal) s.*

Define the emission rates

$$\alpha := \Pr[S \mid C = 1], \qquad \beta := \Pr[S \mid C = 0].$$

Then the posterior credibility of the claim given observation of s is

$$\Pr[C = 1 \mid S] \;=\; \frac{p\,\alpha}{p\,\alpha + (1 - p)\,\beta}.$$

Equivalently, in odds form,

$$\frac{\Pr[C = 1 \mid S]}{\Pr[C = 0 \mid S]} \;=\; \frac{p}{1 - p} \cdot \frac{\alpha}{\beta}.$$

In particular, if s is a cheap-talk *pattern in the sense that:*

(i) truthful senders emit s with certainty ($\alpha = 1$), and

(ii) deceptive senders can mimic s with probability at least q (i.e. $\beta \geq q$),

then credibility obeys the tight upper bound

$$\Pr[C = 1 \mid S] \;\leq\; \frac{p}{p + (1-p)q}.$$

Moreover this bound is tight*: equality holds whenever* $\alpha = 1$ *and* $\beta = q$*.*

Proof. By Bayes' rule,

$$\Pr[C = 1 \mid S] = \frac{\Pr[S \mid C = 1]\Pr[C = 1]}{\Pr[S \mid C = 1]\Pr[C = 1] + \Pr[S \mid C = 0]\Pr[C = 0]} = \frac{p\alpha}{p\alpha + (1-p)\beta}.$$

If $\alpha = 1$ and $\beta \geq q$, the denominator is minimized by setting $\beta = q$, yielding

$$\Pr[C = 1 \mid S] \leq \frac{p}{p + (1-p)q}.$$

Tightness is immediate when $\beta = q$. ■

Remark (Notation reconciliation). In this paper we use q to denote the *mimicability* of a cheap-talk signal: the probability that a deceptive sender successfully produces the same message pattern as a truthful sender. If one prefers to work with detection probability π_d (the probability deception is detected), then $q = 1-\pi_d$ and the bound becomes $\Pr[C = 1 \mid S] \leq p/(p+(1-p)(1-\pi_d))$.

Interpretation: No matter how emphatically you assert something, cheap talk credibility is capped. The cap depends on how likely deception is in the population.

3.2 The Magnitude Penalty

Theorem 3.2 (Magnitude Penalty)**.** *For claims* c_1, c_2 *with* $M(c_1) < M(c_2)$ *(i.e.,* $p_1 := P(c_1) > p_2 := P(c_2)$*) and identical cheap talk signals* s *with mimicability* q*:*

$$\Pr[c_1 \mid S] > \Pr[c_2 \mid S]$$

Higher-magnitude claims receive less credibility from identical signals.

Proof. From Theorem 3.1, the bound $p/(p + (1-p)q)$ is strictly increasing in p for fixed $q \in (0, 1)$. Since $p_1 > p_2$, we have $\Pr[c_1 \mid S] > \Pr[c_2 \mid S]$. ■

Interpretation: Claiming you wrote one good paper gets more credibility than claiming you wrote four. The signal (your assertion) is identical; the prior probability differs.

3.3 The Emphasis Penalty

Theorem 3.3 (Emphasis Penalty)**.** *Let* $s_1, s_2, ..., s_n$ *be cheap talk signals all asserting claim* c*. There exists* k^* *such that for* $n > k^*$*:*

$$\frac{\partial C(c, s_{1..n})}{\partial n} < 0$$

Additional emphasis decreases *credibility past a threshold.*

Proof. The key insight: excessive signaling is itself informative. Define the *suspicion function*:

$$\sigma(n) = P(\text{deceptive}|n \text{ assertions})$$

Honest agents have less need to over-assert. Therefore:

$$P(n \text{ assertions}|\text{deceptive}) > P(n \text{ assertions}|\text{honest}) \text{ for large } n$$

By Bayes' rule, $\sigma(n)$ is increasing in n past some threshold.

Substituting into the credibility update:

$$C(c, s_{1..n}) = \frac{P(c) \cdot (1 - \sigma(n))}{P(c) \cdot (1 - \sigma(n)) + (1 - P(c)) \cdot \sigma(n)}$$

This is decreasing in $\sigma(n)$, hence decreasing in n for $n > k^*$. ■

Interpretation: "Trust me, I'm serious, this is absolutely true, I swear" is *less* credible than just stating the claim. The emphasis signals desperation.

3.4 The Meta-Assertion Trap

Theorem 3.4 (Meta-Assertion Trap)**.** *Let a be a cheap talk assertion and m be a meta-assertion "assertion a is credible." Then:*

$$C(c, a \cup m) \leq C(c, a) + \epsilon$$

where $\epsilon \to 0$ as common knowledge of rationality increases.

Proof. Meta-assertion m is itself cheap talk (costs nothing to produce regardless of truth). Therefore m is subject to the Cheap Talk Bound (Theorem 3.1).

Under common knowledge of rationality, agents anticipate that deceptive agents will produce meta-assertions. Therefore:

$$P(m|\text{deceptive}) \approx P(m|\text{honest})$$

The signal provides negligible information; $\epsilon \to 0$. ■

Interpretation: "My claims are verified" is cheap talk about cheap talk. It doesn't escape the bound—it's *subject to* the bound recursively. Adding "really verified, I promise" makes it worse.

---

4 Costly Signal Characterization

4.1 Definition and Properties

Theorem 4.1 (Costly Signal Effectiveness)**.** *For costly signal s with cost differential $\Delta = \mathit{Cost}(s|\bot) - \mathit{Cost}(s|\top) > 0$:*

$$\Pr[C = 1 \mid S] \to 1 \text{ as } \Delta \to \infty$$

Costly signals can achieve arbitrarily high credibility.

Proof. If Δ is large, deceptive agents cannot afford to produce s, so $\beta := \Pr[S \mid C = 0] \to 0$ as $\Delta \to \infty$. Applying Theorem 3.1 with $\alpha = 1$:

$$\Pr[C = 1 \mid S] = \frac{p}{p + (1-p)\beta} \to 1 \text{ as } \beta \to 0.$$

■

Theorem 4.2 (Verified signals drive credibility to 1)**.** *Let $C \in \{0, 1\}$ with prior $p = \Pr[C = 1]$. Suppose a verifier produces an acceptance event A such that*

$$\Pr[A \mid C = 1] \geq 1 - \varepsilon_T, \qquad \Pr[A \mid C = 0] \leq \varepsilon_F,$$

for some $\varepsilon_T, \varepsilon_F \in [0, 1]$. Then

$$\Pr[C = 1 \mid A] \;\geq\; \frac{p(1-\varepsilon_T)}{p(1-\varepsilon_T) + (1-p)\varepsilon_F}.$$

In particular, if $\varepsilon_F \to 0$ and ε_T is bounded away from 1, then $\Pr[C = 1 \mid A] \to 1$.

Proof. Apply Theorem 3.1 with $S := A$, $\alpha := \Pr[A \mid C = 1]$, $\beta := \Pr[A \mid C = 0]$, then use $\alpha \geq 1 - \varepsilon_T$ and $\beta \leq \varepsilon_F$. ■

Remark. This theorem provides the formal bridge to machine-checked proofs: Lean corresponds to a verifier where false claims have negligible acceptance probability ($\varepsilon_F \approx 0$, modulo trusted kernel assumptions). The completeness gap ε_T captures the effort to construct a proof.

4.2 Examples of Costly Signals

Signal	Cost if True	Cost if False	Credibility Shift
PhD from MIT	4 years effort	4 years + deception risk	Moderate
Working code	Development time	Same + it won't work	High
Verified Lean proofs	Proof effort	Impossible (won't compile)	Maximum
Verbal assertion	˜0	˜0	Bounded

Key insight: Lean proofs with `0 sorry` are *maximally costly signals*. You cannot produce a compiling proof of a false theorem. The cost differential is infinite [13, 12].

Theorem 4.3 (Proof as Ultimate Signal)**.** *Let s be a machine-checked proof of claim c. Then:*

$$\Pr[c \mid s] = 1 - \varepsilon$$

where ε accounts only for proof assistant bugs.

Proof. This is a special case of Theorem 4.2 with $\varepsilon_T \approx 0$ (proof exists if claim is true and provable) and $\varepsilon_F \approx 0$ (proof assistant soundness). See [13, 12]. ■

5 Impossibility Results

5.1 The Text Credibility Bound

Theorem 5.1 (Text Credibility Bound)**.** *For any text string T (memory content, assertion, etc.) and high-magnitude claim c with $M(c) > M^*$ (i.e., prior $p < e^{-M^*}$):*

$$\Pr[c \mid T] < \tau$$

where $\tau < 1$ is determined by the mimicability q and M^*. No text achieves full credibility for exceptional claims.

Proof. Text is cheap talk (production cost independent of truth). Apply Theorem 3.1 with prior $p = e^{-M^*}$ and mimicability q:

$$\tau = \frac{p}{p + (1-p)q} = \frac{e^{-M^*}}{e^{-M^*} + (1 - e^{-M^*})q}$$

For M^* large (low prior probability), $\tau \to 0$ regardless of $q > 0$. ■

Corollary 5.2 (Memory Iteration Futility). No rephrasing of memory content can achieve credibility above τ for high-magnitude claims. Iteration on text is bounded in effectiveness.

Interpretation: This is why we couldn't solve the credibility problem by editing memory text. The *structure* of the problem (text is cheap talk, claims are high-magnitude) guarantees bounded credibility regardless of phrasing.

5.2 Optimal Strategies

Theorem 5.3 (Optimal Credibility Strategy). For high-magnitude claims, the credibility-maximizing strategy is: 1. Minimize cheap talk (reduce emphasis, meta-assertions) 2. Maximize costly signal exposure (show the work, provide proofs) 3. Enable real-time demonstration (costly to fake)

---

6 Leverage Integration

6.1 Credibility as DOF Minimization

Applying the leverage framework (Paper 3) [4]:

Signal DOF: Words in an assertion are degrees of freedom. Each word can be independently modified.

Signal Leverage: $L_S = \frac{\Delta C}{\text{Words}}$

Theorem 6.1 (Credibility Leverage). For cheap talk signals, leverage is maximized by minimizing word count:

$$\arg\max_s L_S(s) = \arg\min_s |s|$$

subject to conveying the claim.

Interpretation: Shorter, terser memory entries achieve higher credibility leverage than verbose explanations. "70k lines, deployed in 3 labs" beats lengthy justification.

6.2 Optimal Memory Design

Given Theorems 5.1-5.3 and 6.1, optimal memory content should:

1. **State facts without meta-justification** (reduces Emphasis Penalty)
2. **Include verifiable anchors** (third-party deployments, citations)
3. **Specify mechanism** (explains how exceptional output is achievable)
4. **Direct behavioral calibration** (tell model how to act, not what to believe)

---

7 Related Work

Signaling Theory: Spence (1973) [25] introduced costly signaling in job markets. Zahavi (1975) [27] applied it to biology (handicap principle). Akerlof (1970) [1] established the foundational role of asymmetric information in market collapse. We formalize and extend to text-based communication.

Cheap Talk: Crawford & Sobel (1982) [11] analyzed cheap talk in game theory. Farrell (1987) [15] and Farrell & Rabin (1996) [16] further characterized the limits of unverified communication. We prove explicit bounds on credibility shift.

Epistemic Logic: Hintikka (1962) [18], Fagin et al. (1995) [14] formalized knowledge and belief. We add signaling structure.

Bayesian Persuasion: Kamenica & Gentzkow (2011) [19] studied optimal information disclosure. Our impossibility results complement their positive results.

---

8 Conclusion

We have formalized why assertions of credibility can decrease perceived credibility, proved impossibility bounds on cheap talk, and characterized the structure of costly signals.

Key results: 1. Cheap talk credibility is bounded (Theorem 3.1) 2. Emphasis decreases credibility past threshold (Theorem 3.3) 3. Meta-assertions are trapped in the same bound (Theorem 3.4) 4. No text achieves full credibility for exceptional claims (Theorem 5.1) 5. Only costly signals (proofs, demonstrations) escape the bound (Theorem 4.1)

Implications: - Memory phrasing iteration has bounded effectiveness - Real-time demonstration is the optimal credibility strategy - Lean proofs are maximally costly signals (infinite cost differential)

Methodology and Disclosure

Role of LLMs in this work. This paper was developed through human-AI collaboration, and this disclosure is particularly apropos given the paper's subject matter. The author provided the core intuitions—the cheap talk bound, the emphasis paradox, the impossibility of achieving full credibility via text—while large language models (Claude, GPT-4) served as implementation partners for formalization, proof drafting, and LaTeX generation.

The Lean 4 proofs (633 lines, 0 sorry placeholders) were iteratively developed: the author specified theorems, the LLM proposed proof strategies, and the Lean compiler verified correctness.

What the author contributed: The credibility framework itself, the cheap talk bound conjecture, the emphasis penalty insight, the connection to costly signaling theory, and the meta-observation that Lean proofs are maximally costly signals.

What LLMs contributed: LaTeX drafting, Lean tactic suggestions, Bayesian calculation assistance, and prose refinement.

Meta-observation: This paper was produced via the methodology it describes—intuition-driven, LLM-implemented—demonstrating in real-time the credibility dynamics it formalizes. The LLM-generated text is cheap talk; the Lean proofs are costly signals. The proofs compile; therefore the theorems are true, regardless of how the proof text was generated. This is the paper's own thesis applied to itself.

---

9 Appendix: Lean Formalization

9.1 On the Nature of Foundational Proofs

Before presenting the proof listings, we address a potential misreading: a reader examining the Lean source code will notice that many proofs are straightforward applications of definitions or Bayesian updating rules. This simplicity is not a sign of triviality. It is characteristic of *foundational* work in game theory and signaling, where the insight lies in the formalization, not the derivation.

Definitional vs. derivational proofs. Our core theorems establish *definitional* properties and Bayesian consequences, not complex derivations. For example, the cheap talk bound (Theorem 3.1) proves that text-only signals cannot exceed a credibility ceiling determined by priors and detection probability. The proof follows from Bayes' rule—it's an unfolding of what "cheap talk" means (cost independent of truth value). This is not a complex derivation; it is applying the definition of cheap talk to Bayesian updating.

Precedent in game theory. This pattern appears throughout foundational game theory and signaling:

- **Crawford & Sobel (1982):** Cheap talk equilibrium characterization. The proof applies sequential rationality to show equilibria must be interval partitions. The construction is straightforward once the right framework is identified.
- **Spence's Signaling Model (1973):** Separating equilibrium in labor markets. The proof shows costly education signals quality because low-ability workers find it too expensive. The mathematics is basic calculus comparing utility differences.
- **Akerlof's Lemons (1970):** Market for lemons unraveling. The proof is pure adverse selection logic—once quality is unobservable, bad products drive out good. The profundity is in recognizing the mechanism, not deriving it.

Why simplicity indicates strength. A definitional theorem derived from precise formalization is *stronger* than an informal argument. When we prove that text credibility is bounded (Theorem 5.1), we are not saying "we haven't found a persuasive argument yet." We are saying something universal: *any* text-based argument, no matter how cleverly phrased, cannot exceed the cheap talk bound for high-magnitude claims. The proof follows from the definition of cheap talk plus Bayesian rationality.

Where the insight lies. The semantic contribution of our formalization is:

1. **Precision forcing.** Formalizing "credibility" in Lean requires stating exactly what it means for a signal to be believed. We define credibility as posterior probability after Bayesian updating, which forces precision about priors, likelihoods, and detection probabilities.

2. **Impossibility results.** Theorem 5.2 (memory iteration futility) proves that iteratively refining text in memory cannot escape the cheap talk bound. This is machine-checked to hold for *any* number of iterations—the bound is definitional, not algorithmic.

3. **Leverage connection.** Theorem 6.1 connects credibility to the leverage framework (Paper 3), showing that credibility-per-word is the relevant metric. This emerges from the formalization of signal cost structure, not from intuition.

What machine-checking guarantees. The Lean compiler verifies that every proof step is valid, every definition is consistent, and no axioms are added beyond Lean's foundations (extended with Mathlib for real analysis and probability). Zero `sorry` placeholders means zero unproven claims. The 430+ lines establish a verified chain from basic definitions (signals, cheap talk, costly signals) to the final theorems (impossibility results, leverage minimization). Reviewers need not trust our informal explanations—they can run `lake build` and verify the proofs themselves.

Comparison to informal signaling arguments. Prior work on AI credibility and generated text (Bommasani et al. [9], Bender et al. [8]) presents compelling informal arguments about trustworthiness but lacks formal signaling models. Our contribution is not new *wisdom*—the insight that cheap talk is non-credible is old (Crawford & Sobel [11]). Our contribution is *formalization*: applying signaling theory to AI-mediated communication, formalizing the cheap talk vs. costly signal distinction for LLM outputs, and proving the impossibility results hold for machine-checked proofs as ultimate costly signals.

This follows the tradition of formalizing economic principles: just as Myerson [22] formalized incentive compatibility and Mas-Colell et al. [20] formalized general equilibrium, we formalize credibility in AI-mediated signaling. The proofs are simple because the formalization makes the structure clear. Simple proofs from precise definitions are the goal, not a limitation.

9.2 Module Structure

The following proofs were developed in Lean 4 [13, 26]. The source code is organized as follows:

```
Credibility/
|- Basic.lean -- Definitions 2.1-2.8
|- CheapTalk.lean -- Theorems 3.1-3.4
|- CostlySignals.lean -- Theorems 4.1-4.2
|- Impossibility.lean -- Theorems 5.1-5.3
‘- Leverage.lean -- Theorem 6.1
```

9.3 Core Definitions (Lean 4)

```
-- Basic.lean

/-- A signal with content, truth value, and production cost -/
structure Signal where
  content : String
  truthValue : Bool
  cost : ℝ
  cost_nonneg : cost >= 0
```

```
/-- Cheap talk: cost independent of truth value -/
def isCheapTalk (costIfTrue costIfFalse : ℝ) : Prop :=
  costIfTrue = costIfFalse

/-- Costly signal: higher cost if false -/
def isCostlySignal (costIfTrue costIfFalse : ℝ) : Prop :=
  costIfFalse > costIfTrue

/-- Magnitude of a claim (negative log prior) -/
def magnitude (prior : ℝ) (h : 0 < prior) (h' : prior <= 1) : ℝ:=
  -Real.log prior

/-- Credibility function type -/
def CredibilityFn := Claim -> List Signal -> ℝ
```

9.4 Cheap Talk Bound (Lean 4)

```
-- CheapTalk.lean

/-- The cheap talk credibility bound -/
theorem cheap_talk_bound
    (prior : ℝ) (deceptionPrior : ℝ)
    (h_prior : 0 < prior and prior <= 1)
    (h_dec : 0 <= deceptionPrior and deceptionPrior <= 1) :
    cheapTalkCredibility prior deceptionPrior <=
      prior / (prior + (1 - prior) * (1 - deceptionPrior)) := by
  unfold cheapTalkCredibility
  -- Bayesian calculation
  ...

/-- Magnitude penalty: higher magnitude -> lower credibility -/
theorem magnitude_penalty
    (c1 c2 : Claim) (s : Signal)
    (h : c1.prior > c2.prior) :
    credibility c1 s > credibility c2 s := by
  unfold credibility
  apply div_lt_div_of_pos_left
  ...

/-- Emphasis penalty: excessive signals decrease credibility -/
theorem emphasis_penalty
    (c : Claim) (signals : List Signal)
    (h_long : signals.length > emphasisThreshold) :
    exists k, forall n > k,
      credibility c (signals.take (n+1)) < credibility c (signals.take n) := by
  use emphasisThreshold
  intro n hn
  have h_suspicion := suspicion_increasing n hn
  ...
```

9.5 Impossibility Result (Lean 4)

```
-- Impossibility.lean

/-- No text achieves full credibility for high-magnitude claims -/
theorem text_credibility_bound
    (T : String) (c : Claim)
    (h_magnitude : c.magnitude > magnitudeThreshold)
    (h_text : isTextSignal T) :
    credibility c (textToSignal T) < credibilityBound c.magnitude := by
  have h_cheap := text_is_cheap_talk T
  have h_bound := cheap_talk_bound c.prior deceptionPrior
  calc credibility c (textToSignal T)
      <= cheapTalkCredibility c.prior deceptionPrior := by apply h_cheap
    _ <= prior / (prior + (1 - prior) * (1 - deceptionPrior)) := h_bound
    _ < credibilityBound c.magnitude := by
        apply bound_decreasing_in_magnitude
        exact h_magnitude

/-- Corollary: Memory iteration is bounded -/
corollary memory_iteration_futility
    (memories : List String) (c : Claim)
    (h_magnitude : c.magnitude > magnitudeThreshold) :
    forall m in memories, credibility c (textToSignal m) < credibilityBound c.magnitude
        := by
  intro m _
  exact text_credibility_bound m c h_magnitude (string_is_text m)
```

Lines: 430 **Theorems:** ~12 **Sorry placeholders:** 0

10 Preemptive Rebuttals

We address anticipated objections to the credibility framework.

10.1 Objection 1: "Signaling theory is old—this isn't novel"

Objection: "Spence's signaling model and Crawford-Sobel's cheap talk are decades old. This paper just applies existing theory."

Response: The foundations are established; the application is novel. Our contributions:

1. **Meta-assertions:** We apply cheap talk bounds to *claims about credibility itself*—a recursive structure not analyzed in prior work
2. **Computable bounds:** We derive explicit formulas for credibility ceilings as functions of prior deception probability
3. **Leverage integration:** We connect credibility to the DOF framework, showing credibility as a form of epistemic leverage

The theorems (Cheap Talk Bound, Magnitude Penalty, Meta-Assertion Trap) are new. The methodology is established.

10.2 Objection 2: "Real communication isn't Bayesian"

Objection: "Humans don't update beliefs according to Bayes' rule. The rationality assumption is unrealistic."

Response: The rationality assumption provides *upper bounds*. If agents deviate from Bayesian reasoning, credibility bounds may be tighter (irrational skepticism) or looser (irrational credulity). The theorems characterize what is *achievable* under optimal reasoning—a ceiling, not a prediction.

For AI systems designed to be rational, the bounds are prescriptive. For human communication, they are normative benchmarks.

10.3 Objection 3: "Costly signals aren't always truth-dependent"

Objection: "Expensive signals can be equally costly for liars and truth-tellers. Your distinction is too clean."

Response: Correct. The definition (Section 2) distinguishes:

- **Truth-dependent costs:** Lying is more expensive than truth-telling (e.g., maintaining consistent false memories)
- **Truth-independent costs:** Equal cost regardless of truth value (e.g., verbose phrasing)

Expensive signals with truth-independent costs remain cheap talk despite their expense. The credibility-enhancing property comes from the *differential*, not the absolute cost.

10.4 Objection 4: "The magnitude penalty seems wrong"

Objection: "Detailed claims are often more credible, not less. The magnitude penalty contradicts intuition."

Response: The distinction is between:

- **Detail about the claim:** Research, evidence, specificity—these are costly signals (effort-dependent) and credibility-enhancing
- **Detail about credibility itself:** "I'm absolutely certain, trust me, this is verified"—these are cheap talk and subject to the magnitude penalty

The theorem applies to meta-assertions (claims about credibility), not to substantive claims. A detailed scientific argument is credibility-enhancing; a detailed assertion of trustworthiness is credibility-reducing.

10.5 Objection 5: "The Lean proofs are trivial"

Objection: "The Lean proofs just formalize definitions. There's no deep mathematics."

Response: The value is precision, not difficulty. The proofs verify:

1. The cheap talk bound follows from the cost-independence definition
2. The magnitude penalty follows from the recursive structure of meta-assertions
3. The impossibility results follow from the bound composition

Machine-checked proofs eliminate ambiguity in informal arguments. The contribution is verification, not complexity.

10.6 Objection 6: "This doesn't apply to AI systems"

Objection: "AI systems can be designed to be trustworthy. The cheap talk bounds don't apply to engineered systems."

Response: The bounds apply to *communication about* trustworthiness, not to trustworthiness itself. An AI system can be trustworthy, but its *assertions* of trustworthiness are cheap talk unless backed by costly signals (audits, formal verification, track record).

The framework explains why "I am safe and helpful" is less credibility-enhancing than demonstrated safety and helpfulness.

10.7 Objection 7: "The impossibility results are too strong"

Objection: "Theorem 5.1 says no string achieves high credibility for high-magnitude claims. But some claims are believed."

Response: The theorem applies to *cheap talk* strings. Credibility for high-magnitude claims requires costly signals:

- Track record (historical accuracy)
- Formal verification (mathematical proof)
- Reputation stake (costly to lose)
- Third-party attestation (independent verification)

The impossibility is for *verbal assertions alone*. Credibility is achievable through costly signals, not through phrasing.

10.8 Objection 8: "The leverage integration is forced"

Objection: "Connecting credibility to DOF seems like a stretch. These are different concepts."

Response: The connection is structural. Credibility leverage is:

$$L_C = \frac{\Delta\text{Credibility}}{\text{Signal Cost}}$$

This parallels architectural leverage:

$$L = \frac{|\text{Capabilities}|}{\text{DOF}}$$

Both measure efficiency: capability per unit of constraint. The integration shows that credibility optimization follows the same mathematical structure as architectural optimization. This unification is the theoretical contribution.

References

[1] George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

[2] Anonymous. Paper 1: Nominal typing dominates structural typing, 2025. Technical report.

[3] Anonymous. Paper 2: Single source of truth minimizes modification complexity, 2025. Technical report.

[4] Anonymous. Paper 3: Leverage maximization in software systems, 2025. Technical report.

[5] Anonymous. Paper 4: The complexity of decision-relevant uncertainty, 2025. Technical report.

[6] Robert J Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976.

[7] Robert J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8(1):6–19, 1995.

[8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[9] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[10] In-Koo Cho and David M. Kreps. Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221, 1987.

[11] Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

[12] N. G. de Bruijn. The mathematical language AUTOMATH, its usage, and some of its extensions. In *Symposium on Automatic Demonstration*, volume 125 of *Lecture Notes in Mathematics*, pages 29–61. Springer, 1970.

[13] Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, pages 625–635. Springer, 2021.

[14] Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Y Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

[15] Joseph Farrell. Cheap talk, coordination, and entry. *The RAND Journal of Economics*, 18(1):34–39, 1987.

[16] Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic Perspectives*, 10(3):103–118, 1996.

[17] H. Paul Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Volume 3: Speech Acts*, pages 41–58. Academic Press, 1975.

[18] Jaakko Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, 1962.

[19] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[20] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[21] Paul Milgrom and John Roberts. Price and advertising signals of product quality. *Journal of Political Economy*, 94(4):796–821, 1986.

[22] Roger B. Myerson. Incentive compatibility and the bargaining problem. *Econometrica*, 47(1):61–73, 1979.

[23] Claude E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[24] Joel Sobel. A theory of credibility. *The Review of Economic Studies*, 52(4):557–573, 1985.

[25] Michael Spence. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973.

[26] The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020)*, 2020.

[27] Amotz Zahavi. Mate selection—a selection for a handicap. *Journal of Theoretical Biology*, 53(1):205–214, 1975.