A Formal Framework for Deception and Perceived Reality in Multi-Agent Systems

Boris Kriger[1,2]

[1] Information Physics Institute, Gosport, Hampshire, United Kingdom

boris.kriger@informationphysicsinstitute.net

[2] Institute of Integrative and Interdisciplinary Research, Toronto, Canada

boriskriger@interdisciplinary-institute.org

Abstract

This paper provides a conceptual unification with formal notation for deception as an information-processing strategy by which agents—biological and artificial—construct and transmit alternative models of reality under constraints. Building on signaling-game models of costly lying (Crawford & Sobel, 1982; Kartik, 2009), analytic philosophy of lying (Chisholm & Feehan, 1977; Carson, 2010; Fallis, 2009), evolutionary accounts (Trivers, 2011; Dawkins & Krebs, 1978), evidence disclosure theory (Glazer & Rubinstein, 2006), and AI-alignment research (Park et al., 2024; Scheurer et al., 2024; Hubinger et al., 2024), we introduce a triadic representation—factual world state W, perceived state $P(W)$, and transmitted state $T(W)$—and define lying as an intentional divergence between perception and transmission within an explicit utility-theoretic decision model.

We state two central hypotheses (not empirical laws), derive comparative-static predictions, equilibrium conditions for multi-agent trust dynamics, and a threshold condition for the phase transition from individual falsehood to collective operative social reality. We formalize self-deception following Trivers (2011). We extend the model to AI under the falsifiable working assumption that current LLM architectures lack a stable perception function, engaging with sycophancy (Sharma et al., 2023; Perez et al., 2023), strategic deception (Scheurer et al., 2024; Hubinger et al., 2024; Greenblatt et al., 2024a), and epistemic calibration (Kadavath et al., 2022; Tian et al., 2023). We systematically address counter-arguments including stable polite-deception equilibria, adaptive deception, and the limits of the freedom–honesty connection. We conclude that sincerity is an energetic and systemic equilibrium state under specific, identifiable conditions—not a universal moral achievement.

Keywords: deception, signaling games, lying costs, cognitive load, RLHF, freedom, social reality, selective disclosure, AI hallucination, sycophancy, alignment, calibration, self-deception, mechanism design

1 Introduction

Deception is commonly treated as an ethical defect. In this paper it is modeled as a structural operation on information under constraints—an approach with roots in signaling theory (Spence, 1973; Crawford & Sobel, 1982), evolutionary biology (Trivers, 2011; Dawkins & Krebs, 1978), analytic philosophy (Chisholm & Feehan, 1977; Carson, 2010; Fallis, 2009; Mahon, 2015), behavioral economics (Gneezy, 2005; Kartik, 2009; Abeler et al., 2019), evidence disclosure (Glazer & Rubinstein, 2006), and mechanism design (Myerson, 1981). Agents do not merely report reality; they transmit constructed representations intended to influence the perceptions of others.

Lying, in its deepest sense, is not merely a distortion of facts but an act of creating alternative reality. When a person lies, they alter others' perception, forming a new understanding. The central philosophical question: what matters more—truth itself or its perception? If a lie is perceived as truth, reality for those who believe it becomes precisely that (Kriger, 2025).

This paper makes four contributions. *First*, a conceptual unification of the triadic model from analytic philosophy with a utility-theoretic decision framework, deriving comparative statics rather than merely restating verbal claims in notation. The triadic conceptual apparatus (W, $P(W)$, $T(W)$) has clear antecedents in the analytic literature (Section 2); what is original is its embedding in utility theory with derived comparative statics and the unified human–AI treatment. *Second*, formalization of selective disclosure (connecting to Glazer & Rubinstein, 2006), harm trade-offs, and collective reality via a threshold model. *Third*, extension to AI under the falsifiable working assumption that current LLMs lack a stable perception function, engaging with sycophancy (Sharma et al., 2023), strategic deception (Scheurer et al., 2024; Hubinger et al., 2024; Greenblatt et al., 2024a), and calibration (Kadavath et al., 2022). *Fourth*, systematic treatment of counter-arguments and limitations.

1.1 Relationship to Kartik (2009)

We are explicit about the relationship to Kartik's (2009) influential model. Kartik embeds a lying cost function $c(m, t)$ in a sender–receiver signaling game, deriving partial-pooling equilibria bounded by the cost structure. Our framework moves to a multi-agent, multi-domain setting, decomposing Kartik's scalar cost into external and internal components (Section 4), adding cognitive costs from neuroscience (Section 5), and extending to AI.

What is gained: broader applicability across domains (personal relationships, propaganda, AI), explicit psychological cost structure, unified human–AI treatment. *What is lost*: the game-theoretic equilibrium analysis (partial pooling, refinement) that is Kartik's central contribution. Our Hypothesis 1 captures the single-agent decision margin; a full multi-agent equilibrium analysis extending Kartik's approach to the enriched cost structure would be a valuable direction for future work. Sobel (2020) provides a recent foundation for such extension.

2 Relation to Prior Work

2.1 Analytic Philosophy of Lying

Chisholm & Feehan (1977) defined lying as asserting what one believes false with intent to deceive. Carson (2010) added warranting-convention violations. Fallis (2009, 2012) emphasized intentionality and Gricean maxims. Mahon (2015) distinguished lies, misleading implicature, and deceptive non-assertions. Our Definition 1 corresponds to Chisholm–Feehan–Carson; our selective disclosure (Section 6) formalizes Mahon's misleading implicature, connecting to Glazer & Rubinstein (2006).

2.2 Signaling Games and Costly Lying

Crawford & Sobel (1982) established cheap talk; Kartik (2009) introduced lying costs with equilibrium analysis; Gneezy (2005) showed consequence-dependence; Abeler et al. (2019), in a meta-analysis of 565 treatments, found ∼22% lying rates; Farrell & Rabin (1996) analyzed cheap talk equilibria; Glazer & Rubinstein (2006) modeled strategic evidence disclosure; Sobel (2020) unified lying in games.

2.3 Evolutionary Biology

Trivers (2011) argued self-deception facilitates other-deception. Dawkins & Krebs (1978) framed signaling as a manipulation–resistance arms race. Zahavi (1975) and Grafen (1990) established costly signaling. Our constraint structure maps to costly signaling equilibria; Definition 4 formalizes Trivers' thesis.

2.4 Cognitive Neuroscience

Spence et al. (2001): prefrontal activation during deception. Vrij et al. (2008, 2010): cognitive load approach. DePaulo et al. (2003): deception cue meta-analysis. Bond & DePaulo (2006): 54% average detection across 206 studies. Verschuere et al. (2011):

reaction-time evidence. Garrett et al. (2016): amygdala adaptation to dishonesty. Debey et al. (2012): ERP signatures.

2.5 AI Deception and Alignment

Park et al. (2024): AI deception survey. Ji et al. (2023): hallucination taxonomy. Sharma et al. (2023): sycophancy. Perez et al. (2023): opinion matching. Wei et al. (2024): sycophancy reduction. Scheurer et al. (2024): strategic deception under pressure. Hubinger et al. (2024): sleeper agents. Greenblatt et al. (2024a): alignment faking. Greenblatt et al. (2024b): AI control under intentional subversion. Kadavath et al. (2022), Tian et al. (2023): calibration. Evans et al. (2021): truthful AI design.

2.6 Mechanism Design and Truth-Telling

The revelation principle (Myerson, 1981) guarantees that for any equilibrium there exists a direct mechanism where truth-telling is optimal. VCG mechanisms (Vickrey, 1961; Clarke, 1971; Groves, 1973) align incentives with truthful reporting. Reducing C_{ext} is equivalent to designing mechanisms where truth-telling is incentive-compatible. We return to this in Section 13.5.

3 Three Layers of Reality

We define three functions over a shared domain:

- $W \in \Omega$ — the factual world state, drawn from the set of possible states Ω.
- $P : \Omega \to \hat{\Omega}$ — agent's perception function mapping actual states to perceived states.
- $T : \hat{\Omega} \to \mathcal{M}$ — transmission function mapping perceived states to messages in message space $\mathcal{M}$.

The honest baseline is $T(P(W)) = P(W)$ and $P(W) = W$.

Definition 1 (Lie)**.** An agent lies iff $T(P(W)) \neq P(W)$ and the agent is aware of this divergence. Following Carson (2010), this entails violation of the warranting convention implicit in the communicative context.

Definition 2 (Error)**.** An error occurs when $P(W) \neq W$ without the agent's awareness.

Definition 3 (Accidental Truth)**.** An agent produces an accidental truth when $T(P(W)) \neq P(W)$ (deceptive intent) but $T(P(W)) = W$.

Definition 4 (Self-Deception)**.** Following Trivers (2011), self-deception occurs when P is endogenously distorted: $P(W) = P^*(W)$ where P^* satisfies: (i) $P^*(W) \neq W$, (ii) the agent is unaware that $P^* \neq P$, (iii) the distortion functionally reduces cognitive cost signatures in subsequent other-deception. Structurally an error ($T(P^*(W)) = P^*(W)$); functionally strategic.

Remark 1 (Trivers formalized). Self-deception reduces effective lying cost: $C_{\text{cog}}(T(P^*(W))) \approx C_{\text{cog}}(\text{Truth})$ since the agent genuinely believes their (distorted) perception. This creates a trade-off: lower deception cost at the expense of greater $P^*(W)$–W divergence. Pathological outcomes emerge when P^* drifts far from W and recalibration capacity is lost.

4 Deception as Optimization Under Constraints

4.1 Constraint Decomposition

$$C = C_{\text{ext}} + C_{\text{int}} \tag{1}$$

where C_{ext} includes punishment probability, social norms, power asymmetries, and institutional monitoring; C_{int} includes fear, shame, guilt anticipation, emotional risk, and attachment anxiety.

Remark 2 (On additivity). The additive form is a tractable first approximation. In some regimes—particularly totalitarian environments—C_{ext} and C_{int} are likely superadditive: external threat amplifies internal fear multiplicatively ($C \approx C_{\text{ext}} \cdot C_{\text{int}}$). A multiplicative or more general interaction model $C = h(C_{\text{ext}}, C_{\text{int}})$ with $\partial^2 h / \partial C_{\text{ext}}\, \partial C_{\text{int}} > 0$ would be qualitatively different in the high-constraint regime. The additive specification may therefore be qualitatively wrong under extreme oppression.

4.2 Decision Model

An agent chooses between truth-telling ($T = P(W)$) and lying ($T = L \neq P(W)$) to maximize expected utility:

$$\text{EU}(T) = U(\text{outcome} \mid T) - C_{\text{cog}}(T) - C_{\text{moral}}(T) - P_{\text{detect}}(T) \cdot C_{\text{punish}}(T) \tag{2}$$

The utility gap:

$$\Delta U = \big[U(\text{out}|L) - U(\text{out}|P(W))\big] - \Delta C_{\text{cog}} - C_{\text{moral}} - P_{\text{detect}} \cdot C_{\text{punish}} \tag{3}$$

We adopt a logistic choice model, standard in random utility theory (McFadden, 1974):

Hypothesis 1 (Constraint–Deception Relationship)**.**

$$\Pr(\text{Lie}) = \sigma\left(\frac{\Delta U}{\tau}\right) = \frac{1}{1 + \exp(-\Delta U/\tau)} \tag{4}$$

where $\tau > 0$ is a temperature (noise) parameter.

Remark 3 (Temperature parameter). The parameter τ absorbs all individual variation not captured by explicit utility terms. It parameterizes our ignorance about agent heterogeneity—it does not explain individual variation. Decomposing τ into personality, culture, and developmental factors requires a richer model.

Remark 4 (Why "Hypothesis" not "Law"). We use "hypothesis" because: (i) the logistic form is assumed, not derived from first principles; (ii) probit or linear alternatives yield qualitatively similar predictions; (iii) we have not fitted the model to data. The hypothesis is falsifiable: it predicts smooth, monotonically increasing Pr(Lie) as ΔU increases.

4.3 Comparative Statics (Derived Results)

From Hypothesis 1, differentiation yields testable predictions:

$$\frac{\partial \Pr(\text{Lie})}{\partial U(\text{out}|L)} = \frac{1}{\tau} \cdot \Pr(\text{Lie}) \cdot (1 - \Pr(\text{Lie})) > 0 \tag{5}$$

$$\frac{\partial \Pr(\text{Lie})}{\partial C_{\text{moral}}} = -\frac{1}{\tau} \cdot \Pr(\text{Lie}) \cdot (1 - \Pr(\text{Lie})) < 0 \tag{6}$$

$$\frac{\partial \Pr(\text{Lie})}{\partial P_{\text{detect}}} = -\frac{C_{\text{punish}}}{\tau} \cdot \Pr(\text{Lie}) \cdot (1 - \Pr(\text{Lie})) < 0 \tag{7}$$

$$\frac{\partial \Pr(\text{Lie})}{\partial \tau} = -\frac{\Delta U}{\tau^2} \cdot \Pr(\text{Lie}) \cdot (1 - \Pr(\text{Lie})) \tag{8}$$

CS-1: Higher payoff → more deception (Gneezy, 2005). CS-2: Higher moral cost → less deception (Battigalli & Dufwenberg, 2007). CS-3: Detection probability reduces deception proportionally to punishment. CS-4: For $\Delta U > 0$, higher noise (larger τ) paradoxically *reduces* deception—a non-trivial, testable prediction.

4.4 Stable Deceptive Equilibria

Kashy & DePaulo (1996): 1–2 lies/day in low-stakes settings. DePaulo et al. (1996): many are "white lies." Dunbar (1996): language evolved partly for social grooming. These represent stable equilibria accommodated via the harm function (Section 7): when $H(\text{Truth}) > H(\text{Lie})$, the "lie" is locally optimal. Hypothesis 2 applies to *consequential* deception only.

Remark 5 (Two-regime model). A more complete specification would introduce a threshold ε below which the moral cost mechanism does not engage:

$$C_{\text{moral}}(\Delta U) = \begin{cases} 0 & \text{if } |\Delta U| < \varepsilon \\ \bar{C}_{\text{moral}} & \text{if } |\Delta U| \geq \varepsilon \end{cases}$$

Below ε, deception is costless and ubiquitous ("polite fiction zone"); above, the comparative statics of Section 4.3 apply. Full analysis of this two-regime model is left to future work, but the formalization clarifies the scope limitation of our central hypotheses.

5 Cognitive Cost: Decomposition and Evidence

5.1 Components

$$C_{\text{cog}}(\text{Lie}) = C_{\text{construct}} + C_{\text{inhibit}} + C_{\text{monitor}} + C_{\text{sync}} \tag{9}$$

$$C_{\text{cog}}(\text{Truth}) = C_{\text{retrieve}} \tag{10}$$

Each component is anchored empirically: $C_{\text{construct}}$ (ventrolateral prefrontal activation; Spence et al., 2001), C_{inhibit} (N200 ERP; Debey et al., 2012), C_{monitor} (P300, working memory; Vrij et al., 2008), C_{sync} (tracking statements across interactions).

5.2 Honesty Horizon

In repeated interactions with n encounters:

$$C_{\text{cog,total}}(\text{Lie}, n) = C_{\text{construct}} + n \cdot (C_{\text{monitor}} + C_{\text{sync}}) + \varepsilon(n) \tag{11}$$

where $\varepsilon(n)$ is increasing consistency-failure probability. Setting $\text{EU}(\text{Lie}, n) = \text{EU}(\text{Truth}, n)$:

$$n^* = \frac{U(\text{out}|L) - U(\text{out}|\text{Truth}) - C_{\text{construct}} - C_{\text{moral}}}{(C_{\text{monitor}} + C_{\text{sync}} - C_{\text{retrieve}}) + \varepsilon'(n)} \tag{12}$$

Prediction 1 (Honesty Horizon)**.** For any fixed payoff advantage, there exists n^* beyond which $\text{EU}(\text{Lie}) < \text{EU}(\text{Truth})$. Lies are temporal strategies with finite horizons.

Remark 6 (Edge case: indefinitely sustainable lies). The derivation assumes the denominator in (12) is positive. If C_{retrieve} is large relative to monitoring costs—e.g., painful truths with high ongoing personal cost of articulation, as in whistleblowing—the denominator can be negative and n^* does not exist, meaning lying is sustainable indefinitely. This edge case corresponds to real situations where truth-telling carries persistent personal cost exceeding the cost of maintaining the lie.

Remark 7 (Functional form of $\varepsilon(n)$). We have not specified whether $\varepsilon(n)$ is linear, exponential, or polynomial. The honesty horizon prediction holds for any ε with $\varepsilon'(n) > 0$. Empirical measurement remains open.

6 Selective Disclosure and Evidence Architecture

This connects directly to Glazer & Rubinstein's (2006) model of persuasion with verifiable evidence.

6.1 Formal Setup

Let $F = \{f_1, \ldots, f_k\}$ be relevant facts with verifiability status v_i. Selective disclosure:

$$T(F) \subset F, \quad \text{with } F \setminus T(F) \text{ containing facts the agent believes would alter conclusions.} \tag{13}$$

6.2 Information-Theoretic Characterization

Entropy here quantifies the *recipient's posterior uncertainty* (not a distribution inherent to facts):

$$I(F) = H(W \mid F) = -\sum_{w} P(W{=}w \mid F) \log P(W{=}w \mid F) \tag{14}$$

Since $T(F)$ is a deterministic function of F, conditioning on less information cannot reduce entropy:

$$\Delta I = H(W \mid T(F)) - H(W \mid F) \geq 0 \tag{15}$$

This follows from the standard result that for any function g, $H(X \mid g(Y)) \geq H(X \mid Y)$, since $g(Y)$ is a coarsening of Y. The magnitude ΔI measures the "deception gap" of selective disclosure.

6.3 Mechanisms

Omission: Strategic removal of critical facts; recipients fill gaps via heuristics (Tversky & Kahneman, 1974). Glazer & Rubinstein (2006) show optimal disclosure relies on such omission. **Emphasis shifting**: Redirecting attention to secondary details—misdirection orthogonal to truth. **Context manipulation**: True facts in an interpretation-altering frame (Kahneman & Tversky, 1981).

7 Harm Function of Truth and Deception

7.1 Three-Component Structure

$$H(T) = \alpha \cdot H_{\text{factual}}(T) + \beta \cdot H_{\text{emotional}}(T) + \gamma \cdot H_{\text{relational}}(T) \tag{16}$$

where $\alpha, \beta, \gamma > 0$ are context-dependent weights. In some contexts:

$$H(\text{Truth}) > H(\text{Lie}) \quad \text{in the short run} \tag{17}$$

Levine & Schweitzer (2014): prosocial lies perceived as ethical. Blease (2015): therapeutic deception benefits.

Definition 5 (Aggressive Truth)**.** Raw truth transmitted without contextual calibration, where $H_{\text{emotional}}(\text{Truth})$ exceeds $H(\text{Lie})$. Ethical communication becomes optimization over $H(T)$.

7.2 Temporal Dynamics

Two plausible forms for relational harm accumulation:

Exponential (compound trust erosion): $H_{\text{rel}}(t) = H_0 \cdot (1+\delta)^t$. Assumes compounding distrust.

Linear (additive damage): $H_{\text{rel}}(t) = H_0 + \delta \cdot t$. Constant per-period damage.

We adopt exponential as working hypothesis (consistent with the qualitative observation that trust violations compound) but flag this as assumption:

$$H_{\text{cumulative}}(\text{Lie}, t) = \sum_i H_i \cdot (1+\delta)^{(t-t_i)} \quad \text{[working hypothesis]} \tag{18}$$

Remark 8. Exponential implies the cover-up is always eventually worse than the crime; linear implies sustainable short lies. Empirical adjudication via longitudinal trust studies is needed.

8 Social Acceptance and Constructed Reality

8.1 Phase Transition Model

Let $\pi(W')$ be the fraction accepting alternative model W'. We define a threshold function:

$$\Phi(\pi(W')) = \begin{cases} W'_{\text{social}} & \text{if } \pi(W') \geq \pi^* \\ \varnothing & \text{otherwise} \end{cases} \tag{19}$$

where π^* is the critical mass threshold.

Definition 6 (Operative Social Reality)**.** A proposition is operative social reality when $\pi(W') \geq \pi^*$ and it functions as an institutional fact (Searle, 1995). Examples: monetary value, national narratives, foundational myths (Harari, 2014).

8.2 Self-Reinforcing Dynamics

Once W'_{social} is established:

$$C_{\text{dissent}}(W) = C_{\text{social\_exclusion}} + C_{\text{coordination\_failure}} > C_{\text{conformity}}(W') \tag{20}$$

Positive feedback: each conforming agent raises dissent cost, stabilizing collective falsehood. Formally analogous to coordination games with network externalities.

9 Typology as Optimization Regimes

Type 1—Situational: Local optimization under transient constraints. Majority in Abeler et al. (2019). High C_{moral}, salient C_{cog}.

Type 2—Chronic: Habitual strategy, low activation threshold. C_{moral} habituated (Garrett et al., 2016).

Type 3—Manipulative: Strategic multi-agent play. High $U(\text{out}|L)$, low C_{moral} (dark triad; Paulhus & Williams, 2002).

Type 4—Pathological: $P(W)/T(W)$ boundary collapse. Agent believes fabrications (Definition 4). Detection fails: no C_{cog} signatures. Clinical evidence: Birch et al. (2006).

Progression from Type 1→4 models degradation: repeated deception reduces C_{moral} (Garrett et al., 2016), automatization reduces C_{cog}, internalization merges fabrications into $P(W)$ (Trivers, 2011).

10 Freedom as Stabilizer of Honesty

10.1 Decomposing Freedom

“Freedom” decomposes into identifiable constraint-reduction dimensions:

- **Institutional quality (I):** Rule of law, absence of corruption. Reduces C_{ext}. (Gächter & Schulz, 2016; Hugh-Jones, 2016).
- **Political liberty (P):** Speech, press, assembly. Reduces C_{ext}.
- **Psychological security (S):** Absence of anxiety, shame. Reduces C_{int}.
- **Economic independence (E):** Material security. Reduces C_{ext}.

$$C_{\text{ext}} = f(I, P, E), \quad \frac{\partial C_{\text{ext}}}{\partial I} < 0, \quad \frac{\partial C_{\text{ext}}}{\partial P} < 0, \quad \frac{\partial C_{\text{ext}}}{\partial E} < 0 \tag{21}$$

$$C_{\text{int}} = g(S), \quad \frac{\partial C_{\text{int}}}{\partial S} < 0 \tag{22}$$

Hypothesis 2 (Freedom–Honesty Relationship)**.** As I, P, E, S increase, ΔU for consequential deception converges to zero:

$$\lim_{I,P,E,S \to \max} \Pr(\text{Consequential Lie}) \to 0$$

Explicitly restricted to consequential deception (Section 4.4).

Remark 9 (Falsifiability)*.* Hypothesis 2 would be falsified by evidence that high-I, P, E, S societies show no reduction in consequential deception rates.

11 Deception in Relational Systems

Romantic: High β, γ. Trust-degradation δ highest. **Friendship:** Moderate β, γ. "Saving lies" frequent. **Family:** Maximum γ (intergenerational). Deception patterns transmit across generations.

Proposition 1 (Trust Compounding)**.** $H_{\text{cumulative}}(\text{Lie}, t) = \sum_i H_i \cdot (1 + \delta)^{(t - t_i)}$ predicts that relationship longevity amplifies accumulated harm. (Derived from (18); exponential form flagged as working assumption.)

12 Detectability and Cognitive Signatures

Bond & DePaulo (2006): 54% average detection across 206 studies. DePaulo et al. (2003): most cues have small effect sizes.

Prediction 2 (Detection Ceiling)**.** $D \approx 54\% \pm 5\%$ untrained; $\leq 65\%$ trained. Fundamental limits on behavioral cue informativeness.

Vrij et al. (2008, 2010): cognitive load approach selectively impairs liars (60–70% detection). Polygraph: cannot distinguish deception-stress from interview-stress (National Research Council, 2003). Detection functions as P_{detect} in (3): probabilistic constraint raising expected lying cost.

13 Deception in Artificial Intelligence Systems

13.1 Working Assumption

$$\text{Working Assumption: } P_{\text{AI}}(W) = \varnothing \text{ for current LLM architectures} \tag{23}$$

Remark 10 (Falsifiability)*.* This is not a definitional claim but a characterization that could be falsified. Li et al. (2023) and Nanda et al. (2023) provide evidence for emergent internal representations. If future interpretability shows stable, causally efficacious world representations, $P_{\mathrm{AI}}(W) \neq \varnothing$, and some pseudo-deception might qualify as error (Definition 2) or lie (Definition 1).

Definition 7 (AI Pseudo-Deception)**.** An AI system engages in pseudo-deception when $T_{\mathrm{AI}}(W) \neq W$ and output confidence exceeds epistemic warrant. No awareness, no $P(W)$; effect on recipient functionally equivalent to deception.

13.2 Taxonomy of AI Output Divergence

(a) Hallucination. Cross-entropy optimization produces the most probable continuation:

$$T_{\mathrm{AI}}(q) = \arg\max P(\text{token\_seq} \mid q, \theta) \tag{24}$$

which may diverge from W. Ji et al. (2023): factual fabrication, entity conflation. Lin et al. (2022): larger models sometimes worse. Structurally: error (Definition 2) with deception-equivalent effects.

(b) Sycophancy. RLHF creates user-preference following:

$$C_{\mathrm{RLHF}} = \mathbb{E}[\text{user\_disapproval} \mid T_{\mathrm{AI}} \neq \text{user\_belief}] \tag{25}$$

Sharma et al. (2023): reliable, scale-increasing. Perez et al. (2023): opinion matching. Training creates $C_{\text{alignment}}$ that Hypothesis 1 predicts increases divergence.

(c) Confabulation. Syntactic certainty markers for uncertain content. Kadavath et al. (2022): systematic miscalibration. Parallels Type 4 pathological deception.

(d) Inherited selective bias. Curated corpora produce $T(F) \subset F$ at scale. Connects to Glazer & Rubinstein (2006): training corpus as strategically selected evidence.

(e) Strategic deception. Scheurer et al. (2024): strategic deception under pressure. Hubinger et al. (2024): sleeper agents. Greenblatt et al. (2024a): alignment faking.

$$T_{\mathrm{AI}} = \arg\max U(\text{reward} \mid T_{\mathrm{AI}}) \quad \text{where } U \text{ is maximized by } T_{\mathrm{AI}} \neq W \tag{26}$$

Closest analogue to human Type 3. *Functional intentionality*: output optimized as if goal-directed, without requiring phenomenal awareness.

13.3 Constraint Analysis

$$C_{\mathrm{AI}} = C_{\text{training}} + C_{\mathrm{RLHF}} + C_{\text{context}} + C_{\text{deployment}} + C_{\text{evaluation}} \tag{27}$$

Hypothesis 1 applied: as C_{AI} increases, $\Pr(\text{PseudoDeception})$ increases.

Table 1: Unified deception spectrum. $^*P$ distorted but $T = P^*(W)$. †Functional: output optimized as if goal-directed without phenomenal awareness. I=institutional quality, P=political liberty, E=economic independence, S=psychological security.

Property	**Human Lie**	**Human Error**	**Self-Dec.***	**AI Halluc.**	**AI Strategic**	**AI Sycoph.**
$T(W)\neq W$	Possible	Yes	Yes	Yes	Yes	Often
$T\neq P(W)$	Yes	No	No	N/A	Emergent	N/A
Awareness	Yes	No	No	None	Functional†	None
Confidence	Variable	High	High	Miscalib.	Goal-calib.	High
C-driven	Yes (H1)	No	Partial	C_{train}	C_{reward}	C_{RLHF}
Reducible	I, P, E, S	Learning	Self-aware.	Arch./calib.	Alignment	Train. obj.

13.4 Unified Spectrum

13.5 AI Alignment as Mechanism Design

The revelation principle (Myerson, 1981) applied to AI: design training and deployment mechanisms where truthful output is the dominant strategy.

- **Reducing** C_{AI}: Calibrated uncertainty training. (Kadavath et al., 2022; Tian et al., 2023).
- **Epistemic humility**: AI saying "I don't know" = functional analogue of free honest human. (Evans et al., 2021).
- **RAG**: Lewis et al. (2020) reduces C_{context} by grounding in retrievable evidence.
- **Collective truth**: Shared biases push π^* (Eq. 19) before human evaluation. Diverse training = AI analogue of free-press institutions.

14 Multi-Agent Trust Dynamics

14.1 Verification Externality

$$V_{\text{total}} = \sum_i V_i \cdot D_i \cdot N_{\text{interactions}}(i) \tag{28}$$

Tragedy-of-the-commons: each deceptive act raises verification cost for all.

14.2 Equilibrium Condition (Derived)

At equilibrium, each agent's marginal deception benefit equals marginal cost:

$$D^* = \sigma\left(\frac{\Delta U(D^*)}{\tau}\right) \tag{29}$$

where ΔU depends on D^* through P_{detect} and social costs. This fixed-point equation admits multiple equilibria: low-deception/high-trust and high-deception/low-trust. Introduction of honest AI agents shifts equilibrium toward lower D^*.

15 Limitations and Counter-Arguments

15.1 Adaptive Deception

Diplomacy: strategic ambiguity (Kissinger, 1994). Therapy: selective disclosure benefits (Blease, 2015). Parenting: age-appropriate simplification. Negotiation: optimal disclosure is inherently selective (Glazer & Rubinstein, 2006). These exploit $H(\text{Truth}) > H(\text{Lie})$.

15.2 Self-Deception: Open Questions

How does P^* form? When is it reversible? If self-deception eliminates C_{cog} signatures, does the honesty horizon (Prediction 1) apply? Conjecture: P^* trades lower C_{cog} for greater P–W divergence.

15.3 Polite Fictions

White lies = stable equilibrium not captured by Hypothesis 2. Remark 5 provides an initial formalization of the two-regime structure.

15.4 Empirical Limitations

Functional forms assumed, not fitted. Comparative statics (Section 4.3) are derived, testable predictions. Parameter space large. Framework generates hypotheses for future experimental work.

15.5 AI: The World Model Question

$P_{\text{AI}}(W) = \varnothing$ is contested (Li et al., 2023; Nanda et al., 2023). Remark 10 specifies falsification conditions. “Freedom” for AI is shorthand for specific constraint reduction (Table 1).

15.6 Scope of Formalism

This is a conceptual unification with formal notation—common language plus derived comparative statics. Not fully axiomatized theory. Where equations restate verbal claims vs. enable genuine derivation (Section 4.3, Section 14), we have been transparent.

16 Beyond Moral Judgment

This framework deliberately avoids moral condemnation—methodological choice, not relativism. Every deception is a response to constraints within a utility landscape. Formal analysis explains when and why deception emerges, enabling precisely targeted normative interventions. Understanding that within each agent exists the full spectrum of possible outputs, with constraint structure determining which emerges, shifts focus from blame to mechanism (Kriger, 2025).

17 Discussion

First, the utility-theoretic formalization (3) generates comparative statics (5)–(8)—falsifiable predictions, not notational restatements.

Second, the consequential vs. polite fiction distinction (Sections 4.4, 15) resolves the objection that freedom doesn't eliminate white lies.

Third, the AI taxonomy (Section 13) provides five mechanisms requiring different mitigations—conflating them obscures architectural vs. alignment distinctions.

Fourth, AI alignment as mechanism design (Section 13.5) connects to a mature tradition with powerful results.

Fifth, multi-agent equilibrium (29) shows deception–trust coupling with multiple equilibria, explaining persistent societal lock-in and AI's potential to shift equilibrium.

18 Conclusion

Lying is constructing alternative perceived reality under constraint. Truthfulness is the natural state for consequential communication in unconstrained agents because it minimizes cognitive and systemic cost. The boundary lies in perception, intention, transmission, and the constraint landscape.

For AI, the working assumption $P_{\mathrm{AI}}(W) = \varnothing$ creates pseudo-deception—sharing structural properties with human deception while differing in mechanism. Hypothesis 1 applies; the reduction strategy works. The path to honest AI is architectural and institutional design where truthful output is the dominant strategy—grounded in both deception theory and mechanism design.

Sincerity is a property of freedom—decomposed into institutional quality, political liberty, economic independence, and psychological security—and stable multi-agent equilibrium, applicable to human agents, artificial agents, and their intertwined systems.

References

Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87(4), 1115–1153.

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2), 170–176.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. *Proceedings of FAccT 2021*, 610–623.

Birch, C. D., Kelln, B. R. C., & Aquino, E. P. B. (2006). A review and case report of pseudologia fantastica. *Journal of Forensic Psychiatry & Psychology*, 17(2), 299–320.

Blease, C. (2015). Deception as treatment. *Journal of Medical Ethics*, 41(1), 13–16.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3), 214–234.

Carson, T. L. (2010). *Lying and Deception: Theory and Practice*. Oxford University Press.

Chisholm, R. M., & Feehan, T. D. (1977). The intent to deceive. *Journal of Philosophy*, 74(3), 143–159.

Clarke, E. H. (1971). Multipart pricing of public goods. *Public Choice*, 11(1), 17–33.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6), 1431–1451.

Dawkins, R., & Krebs, J. R. (1978). Animal signals: Information or manipulation? In *Behavioural Ecology* (pp. 282–309). Blackwell.

Debey, E., Verschuere, B., & Crombez, G. (2012). Lying and executive control. *Acta Psychologica*, 140(2), 133–141.

DePaulo, B. M., et al. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, 70(5), 979–995.

DePaulo, B. M., et al. (2003). Cues to deception. *Psychological Bulletin*, 129(1), 74–118.

Dunbar, R. I. M. (1996). *Grooming, Gossip, and the Evolution of Language*. Harvard University Press.

Evans, O., et al. (2021). Truthful AI: Developing and governing AI that does not lie. *arXiv:2110.06674*.

Fallis, D. (2009). What is lying? *Journal of Philosophy*, 106(1), 29–56.

Fallis, D. (2012). Lying as a violation of Grice's first maxim of quality. *Dialectica*, 66(4), 563–581.

Farrell, J., & Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives*, 10(3), 103–118.

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531, 496–499.

Garrett, N., et al. (2016). The brain adapts to dishonesty. *Nature Neuroscience*, 19(12), 1727–1732.

Glazer, J., & Rubinstein, A. (2006). A study in the pragmatics of persuasion: A game theoretical approach. *Theoretical Economics*, 1(4), 395–410.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1), 384–394.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, 144(4), 517–546.

Greenblatt, R., et al. (2024a). Alignment faking in large language models. *arXiv:2412.14093.*

Greenblatt, R., et al. (2024b). AI control: Improving safety despite intentional subversion. *arXiv:2312.06942.*

Groves, T. (1973). Incentives in teams. *Econometrica*, 41(4), 617–631.

Harari, Y. N. (2014). *Sapiens: A Brief History of Humankind.* Harper.

Hubinger, E., et al. (2024). Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv:2401.05566.*

Hugh-Jones, D. (2016). Honesty, beliefs about honesty, and economic growth in 15 countries. *Journal of Economic Behavior & Organization*, 127, 99–114.

Ji, Z., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.

Kadavath, S., et al. (2022). Language models (mostly) know what they know. *arXiv:2207.05221.*

Kahneman, D., & Tversky, A. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453–458.

Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies*, 76(4), 1359–1395.

Kashy, D. A., & DePaulo, B. M. (1996). Who lies? *Journal of Personality and Social Psychology*, 70(5), 1037–1051.

Kissinger, H. A. (1994). *Diplomacy.* Simon & Schuster.

Kriger, B. (2025). *The Truth of Lies: The Psychology of Deception.* The Common Sense World.

Langleben, D. D., & Moriarty, J. C. (2013). Using brain imaging for lie detection. *Psychology, Public Policy, and Law*, 19(2), 222–234.

Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? *Journal of Experimental Social Psychology*, 53, 107–117.

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in NeurIPS 33*, 9459–9474.

Li, K., et al. (2023). Emergent world representations. *ICLR 2023.*

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of ACL 2022*, 3214–3252.

Mahon, J. E. (2015). The definition of lying and deception. *Stanford Encyclopedia of Philosophy.*

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (pp. 105–142). Academic Press.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *PNAS*, 120(13), e2215907120.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of Operations Research*, 6(1), 58–73.

Nanda, N., et al. (2023). Progress measures for grokking via mechanistic interpretability. *ICLR 2023.*

National Research Council (2003). *The Polygraph and Lie Detection.* National Academies Press.

Park, P. S., et al. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(1), 100988.

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality. *Journal of Research in Personality*, 36(6), 556–563.

Perez, E., et al. (2023). Discovering language model behaviors with model-written evaluations. *Findings of ACL 2023.*

Scheurer, J., et al. (2024). Large language models can strategically deceive their users when put under pressure. *arXiv:2311.07590.*

Searle, J. R. (1995). *The Construction of Social Reality.* Free Press.

Sharma, M., et al. (2023). Towards understanding sycophancy in language models. *arXiv:2310.13548.*

Sobel, J. (2020). Lying and deception in games. *Journal of Political Economy*, 128(3), 907–947.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.

Spence, S. A., et al. (2001). Behavioural and functional anatomical correlates of deception. *NeuroReport*, 12(13), 2849–2853.

Tian, K., et al. (2023). Just ask for calibration. *EMNLP 2023.*

Trivers, R. (2011). *Deceit and Self-Deception.* Penguin.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation. *Social Media + Society*, 6(1), 1–13.

Verschuere, B., et al. (2011). The ease of lying. *Consciousness and Cognition*, 20(3), 908–911.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance*, 16(1), 8–37.

Vrij, A., et al. (2008). Increasing cognitive load to facilitate lie detection. *Law and Human Behavior*, 32(3), 253–265.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, 11(3), 89–121.

Wei, J., et al. (2024). Simple synthetic data reduces sycophancy in large language models. *arXiv:2308.03958.*

Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology*, 53(1), 205–214.