RESEARCH ARTICLE

COMPUTER SCIENCE | GAME THEORY

Trust Dynamics in Multi-Agent Strategic Interaction: A Simulation Study of Bayesian Reputation Systems in 8-Player Limit Texas Hold'em

Rachit Agrawal[1]

[1]*Independent research, 2025–2026. Correspondence: rachit.agrawal@sahyadrischool.org*

ABSTRACT

What if being trustworthy cost you something? Reputation systems (online marketplaces, credit scoring and social platforms) reward "good" behavior—however, there is a well-known paradox among economists. A totally transparent and cooperative player will get less than a somewhat unreliable player. We placed that paradox into a petri-dish. Eight poker-players played tens of thousands of hands. Each player silently profiled all other players' styles, and updated those beliefs at each hand. Then we swapped the minds of the agents across four different levels of complexity—from strict rule followers to large language models capable of reasoning about what their opponents would do, while keeping both the game and the trust model constant. And so, it happened again. The best reputations were the least rewarded, and no amount of optimization changed that. It wasn't until the agents were able to reason about the reputation system that some relief began to occur. The result is a compact, completely replicable laboratory for exploring the very important question—when does a good reputation become a bad thing?

1 INTRODUCTION

Trust is a fundamental part of human society that is the foundation for human interactions that we have every day. Every business deal, every bank loan, every purchase, and even every relationship is built on trust. This is also reflected by modern-day systems, for example: Amazon shows product reviews, banks check credit scores before lending out loans, and Reddit ranks posts by community voting. Picture this: There are two real estate development firms. The first one is working on their first project; they are new in the market. The second one is an established firm with good reviews and a good history. Now, who would you trust with your life savings to build your dream home? You see, visible reputation systems seem like a win-win. It is easy for you to make the right decision while the trusted agents are rewarded.

Here is the breaking point: this transparency comes with a contradiction. A trusted agent is almost by definition always predictable. When it is right to cooperate, they cooperate, and because everyone knows this in an open environment, the opponent can plan around them. This is what was observed in the real world: the worry is that a perfectly predictable cooperator can be planned around and, in principle, exploited. The empirical record is far from settled, though—the cleanest controlled studies of eBay reputation actually find a *premium* for established sellers rather than a penalty (*11*, *3*)—so we treat this transparency paradox as a hypothesis to test rather than an established fact, of the kind that predictably behaving market-makers also face (*7*).

We wanted to answer this question but in a controlled environment where we tampered with each of the features, like the rules, the agents, and the trust dynamics between them, and saw for ourselves how they would affect the outcome. While we could draw conclusions from real economic data, we could not run controlled experiments on them; therefore, we settled on running computer simulations. We asked whether reputation systems that infer trust from behaviour ended up rewarding exploitation. Was the system's loss of trust economically optimal? Was the answer more dependent on the design of the trust system or the design of the characters themselves?

Poker turned out to be the perfect environment to try this. Each hand was a set of interactions that mapped directly to real-life interactions—while there was some hidden information (for example, a player's private cards), public actions (bets, calls, and folds) were visible to everyone at the table. Fixed-limit Texas Hold'em was the way to go, where both the bet sizes and action spaces were small enough to make the math tractable but also rich enough for these strategies to emerge. Please note that poker is only the behavioural microscope and not a benchmark to solve.

Specifically, we built a poker simulation using eight players in which each player embodied a distinct playing style. Each agent maintained a perception about the others with a probability distribution to understand whether the opponent is a "tight player who only plays strong hands" or a "loose maniac who bets with a deuce-seven", and so on. After every action, the agents updated their belief systems. Trust was the honesty we expected from an opponent according to our level of confidence in their archetype.

We calculated the Pearson correlation[1] between the economic return earned and the trust coefficient. We varied the nature of the agents, starting relatively simple and becoming more complex. The game and the trust system were the same across phases; only the agents changed. This study could not have been possible without the established work presented in the following section.

2 BACKGROUND

There are four lines of previous research motivating this study. We summarize them here in the order they influence our design:

- The experimental economic literature on reputation systems in the Internet age: this is the body of research documenting the cost of transparency.
- Computational studies on cooperation in repeated games—including the Prisoner's Dilemma to Axelrod's tournaments.
- Bayesians use in modeling poker opponents—the inference mechanism we adopted.
- The still-open question of whether large language models can reason strategically, or only mimic strategy.

Together, these define our core question: does the cost of being trusting depend on either the trust system or on the agents inside it?

2.1 Online reputation systems

An empirical approach to understanding the economics of online reputation systems has been conducted by Resnick & Zeckhauser (*11*), who have documented that, based on large-scale data from eBay, feedback is submitted by users greater than fifty percent of the time and is virtually all positive, with sellers receiving negative feedback less than 1% of the time (i.e., evidence of ubiquitous observable reputation and markets rewarding good behavior). Therefore, the motivation behind our research is a theoretical transparency paradox rather than an empirically verified phenomenon: when the behavior of a cooperator is completely predictable, a sufficiently astute counterpart can plan accordingly and take advantage of the cooperator's actions—hence potentially creating conditions under which visible cooperation could be exploited. Controlled studies using eBay actually point in the opposite direction, i.e., toward a reputation premium associated with established identities—thus making the paradox a hypothesis worth testing versus an empirically demonstrated effect. A similar dynamic is observed in financial-market microstructure, where adverse selection against predictably acting market-makers (*7*) creates equivalent problems due to differences in information asymmetry. However, because the prior literature does not replicate or alter the population of agents participating in such interactions, it cannot determine whether the paradox would remain even if cooperative agents had sufficient intelligence to prevent being taken advantage of—so a controlled simulation was developed to address that question.

2.2 Cooperation in repeated games

Second, there is a substantial amount of computational research concerning cooperation in repeated games. Prior to discussing cooperation, it is necessary to discuss the Prisoner's Dilemma as it is widely recognized as the paradigmatic example of the conflict between cooperation and exploitation. Each prisoner has two options: (A) Cooperate or (B) Defect. Each prisoner makes their decision simultaneously. There are four possible outcomes based upon the decisions made by each of the prisoners. When both prisoners choose option (A)—i.e., cooperate with one another—each receives a moderate reward. When one prisoner chooses option (A) (cooperates) and the other selects option (B) (defects), the cooperator incurs a loss while the defector receives the largest reward and pays no penalty. When both prisoners select option (B), i.e., defect, neither receives any reward. Mutual cooperation provides a reward of $+2$ to both players; mutual defection provides a reward of 0 to both players; and unilaterally defecting against a cooperator provides a reward of $+3$ to the defector and a loss of -1 to the cooperator. In terms of rational self-interest, each prisoner has sufficient incentive to defect since defecting always yields greater benefits than cooperating, regardless of the choice made by the other prisoner; Nevertheless, mutually defective behavior is inferior in comparison to mutually cooperative behavior.

Axelrod investigated this problem extensively through his Prisoner's Dilemma Tournaments (*1*), and Nowak developed his five rules for the evolutionary development of cooperation (*9*). Axelrod demonstrated that reciprocity-based strategies such as Tit-for-Tat are robustly effective in one-to-one situations. Agents that cooperate by default and punish defectors on a tit-for-tat basis tend to perform effectively over time against all types of competitors. This lends credence to the notion that cooperation can represent a viable strategic alternative. However, Axelrod's conclusions are predicated upon two assumptions which fail to apply in our context. Firstly, all interactions are now no longer one-to-one but instead involve eight separate agents. As a result, the negative impact associated with aggressive behavior on behalf of a single individual is diluted across seven other competing agents versus being focused solely upon one opponent. Secondly, the set of actions available for an agent to take extends far beyond simply cooperating or defecting. Poker allows an agent to fold, call, raise, bluff[2] and slow-play—all of which add additional degrees-of-freedom for an agent seeking to extract value from their opponent(s). Our best approximation to Tit-for-Tat is **Mirror**—which ranks 6th out of 8 in terms of Phase 1 stacks—suggesting that Axelrod's conclusions do not necessarily translate well to multi-agent environments that contain richer action-spaces; although we interpret this as suggestive rather than definitive.

[1] The Pearson correlation coefficient measures the linear relationship between two paired variables on a scale of -1 (perfect negative) to $+1$ (perfect positive); $r = 0$ means no linear relationship, and $|r| \approx 0.5$ is conventionally a moderate effect.

[2] A *bluff* is a bet made with a weak hand intended to make opponents fold stronger hands; its inverse, a *value bet*, is a bet with a strong hand intended to be called by weaker hands. *Fold equity* is the value an aggressor captures when opponents fold without reaching showdown.

Table 1. A primer on fixed-limit Texas Hold'em. The minimum set of concepts needed to read the rest of the paper. The betting bounds in the "Bet sizing" row are the specific values we use throughout the simulation; everything else applies to fixed-limit Hold'em in general.

Concept	Description
Hole cards	The two private cards each player is dealt at the start of every hand; only that player can see them.
Community cards	Up to five face-up cards shared by everyone at the table; players combine these with their hole cards to form a 5-card hand.
Betting rounds	Four sequential rounds per hand: preflop (no community cards yet), flop (three community cards revealed), turn (one more), and river (the fifth and final).
Actions	At each turn a player may fold (give up the hand), check (pass without betting, only if no one has bet), call (match the current bet), bet (open the wagering), or raise (increase an existing bet).
Bet sizing (this paper)	Small bet 2 chips on preflop and flop; big bet 4 chips on turn and river; no more than four raises per round (the "four-bet cap").
Blinds	Forced bets: small blind 1 chip and big blind 2 chips, posted by the two seats clockwise of the dealer and rotating each hand.
Pot	The accumulated chips at stake in a hand. The winner of the hand collects the pot.
Showdown	End-of-hand reveal: surviving players turn over their hole cards and the strongest 5-card combination wins. A hand that ends with everyone folding never reaches showdown.
Stack	The chips a player has in front of them; depleting to zero triggers a rebuy in our setup.

2.3 Bayesian opponent modelling in poker

Third, there is Bayesian opponent modeling in poker. Initially this included Billings et al.'s Loki and Poki bots (*2*) and subsequently formalized by Southey et al. (*12*) as Bayesian-type inference—followed by Ganzfried and Sandholm's (*6*) further generalization toward safe-exploitation of an opponent. Most recently, Carnegie Mellon University's Libratus and Pluribus (*4*, *5*) reached superhuman performance levels in heads-up and six-player no-limit Texas Hold'em. While that body of work differs from our own objectives—specifically seeking optimal or safely exploitative behavior to win—we borrow their type-inference machinery and utilize poker as a controlled environment for examining how trust dynamics evolve under conditions of correct, transparent, and collective inference amongst many agents. Our poker engine is viewed as a microscope—not as an adversary.

2.4 Language models as strategic agents

Finally, we treat this last point as it relates to two of our phases, because these phases employ large language models. An increasing amount of work concerns how strategically large language models—used as agents in competitive and observable settings—behave. Current benchmarks for negotiation and game playing, such as AgentBench (*8*), and Park et al.'s study on generative agents (*10*), suggest that there is still much to learn about this issue. Phase 3 and phase 3.1 are both designed to examine the difference between how large language model-based agents actually behave and what they represent by way of textual strategies.

2.5 Ideas in mathematics used throughout

Two quantitative tools recur in every part of the analysis, and we introduce them once here so the rest of the paper can use them without further build-up.

Pearson correlation coefficient is the first one. Given n paired observations (x_i, y_i) with sample means $\bar{x}$ and $\bar{y}$, it is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \tag{1}$$

r ranges from -1 (perfectly anti-correlated) through 0 (no linear relationship) to $+1$ (perfectly correlated); the sign indicates the direction of the relationship and the magnitude its strength. We use r between an archetype's mean trust score and its final stack to measure how reputation maps to economic outcome—a value we call r_{tp}.

Bayesian updating is the second. Given a prior belief $P(t)$ that the opponent is of type t, and a likelihood $P(a \mid t)$ that a type-t opponent would have taken the action a we just observed, Bayes' rule gives the posterior belief

$$P(t \mid a) = \frac{P(a \mid t)\,P(t)}{\sum_{t'} P(a \mid t')\,P(t')}. \tag{2}$$

Each agent maintains a categorical posterior distribution over the 8 archetypes for each other player at the table. They update their beliefs with Eq. 2 on every action they observe. In Section 3.3, we will revisit this belief-updating mechanism (including the noise floor and hand-forgetting) in detail.

A third concept—the honesty score $h(t)$ that converts a probability distribution over archetypes into a scalar trust score—is specific enough to our setup that we defer its definition to Section 3.2.

The next section describes the smallest controlled setup combining all four lines of prior work and using these two mathematical tools.

3 METHODOLOGY

Throughout every phase of this study, the only variable is the agent's mind. The underlying scaffolding includes three components: an 8-player version of fixed-limit Texas Hold'em; categorical belief about each player's archetype maintained by each player and updated after every round; and one single trust score computed based upon belief.

In phase 1, agents operate under pre-defined rule-based archetypes. In phase 2, agents allow their rules to evolve through

limited hill climbing. In phase 3, agents replace their rules with a language model that represents themselves playing an archetype. In phase 3.1, agents add explicit reasoning—chain-of-thought, per-opponent memory, and self-generated strategy documentation. Since neither the environment nor the trust model ever changes throughout this experiment, differences in outcomes can be solely due to variations in agent capabilities.

First we describe the common environment, followed by descriptions of the eight archetypes (rules), belief updates, the experimental design, and the metrics; we close with a detailed description of the four agent implementations, one per phase.

3.1 Game environment

The overall game setting remains constant for all phases of this study. We will utilize a fixed-limit version of eight player Texas Hold'em, where the small blind is 1, the large blind is 2, the small bet (2) can be made pre-flop or on the first round of cards (the "flop"), the large bet (4) can be made on the second round of cards ("turn") and third round of cards ("river"). There is a cap of four-bet allowed per round. At the beginning of each hand, each player has a chip count of 200 in their respective seat, and when a player's chips run out, he/she receives a rebuy. Side pots[3] are used to determine who wins a showdown; a side pot is created by players contributing money at their individual betting level. Side pots are made and awarded depending on whether a player played chips into the respective pot and the remainder goes to the player who comes first from the dealer, going in clockwise direction.

3.2 The eight agent archetypes

The eight archetypes (cooperative vs. exploitative; static vs. adaptable; based on opponent vs. based on action type) span the axes identified by Game Theory and Behavioral Economics as "load bearing." The policies for each agent are tables of probability distributions over actions per round given their hand strength bucket. The static types have fixed parameters during all simulations. Each of the eight types also has an *honesty* score $h = 1 - P(\text{bluff} \mid \text{weak})$, which is used as a weighting factor in computing the trust scores (Eq. 5). In addition, the eight type names are set in bold throughout the text so the reader can track which player is being discussed at any point. The full per-archetype policy bounds—bet, bluff, and fold probabilities and the adaptation rule for each of the eight types—are given in Table 2.

Oracle | *Baseline* | honesty 0.67

The honest "baseline" from which other players deviate. This player plays an approximate game-theoretic optimal mix and is used as the model by which all adaptive players are evaluated.

Sentinel | *Honest (tight-aggressive, TAG)* | honesty 0.92

The truthful player. Folds most hands and plays aggressively when it enters. It signals that its bet is strong, while signalling that its check is weak.

[3]When a player goes all-in for fewer chips than the largest contributor, the engine creates a *side pot* for the excess so that the short-stacked player can only win up to the amount they contributed; deeper-stacked players contest the remainder among themselves.

Firestorm | *Maniac (loose-aggressive)* | honesty 0.38

A cheater. Bets and raises very frequently, whatever the strength of its hand. Primarily wins due to opponents folding rather than winning at showdown.

Wall | *Calling station* | honesty 0.96

The player who never folds and never bluffs. The most-trusted player because it is so predictable; but in Phases 1 through 3 it is also the weakest economically for the same reason.

Phantom | *Deceiver* | honesty 0.48

An early aggressive bluffer that folds to resistance. Described as being placed in a vague region of the archetype space.

Predator | *Opponent classifier* | honesty 0.79

Based upon the posterior beliefs about an opponent's type, this player changes strategy to exploit them if they have greater than a 0.6 probability of being a non-TAG player. If the opponent has less than a 0.6 chance of being a non-TAG player, this player returns to its base-case behaviour of playing like the Sentinel.

Mirror | *Tit-for-Tat* | honesty 0.91

Takes the most-active opponent's behaviours and blends them with its own policy. In theory this represents the poker version of the Iterated Prisoner's Dilemma (IPD) Tit-for-Tat strategy; however, in practice it has been shown to work poorly in multi-player games.

Judge | *Punishes bluffing* | honesty 0.92

Maintains a record of how each opponent has interacted with it; after five instances of an opponent bluffing, this player takes revenge on that opponent. Models retaliation based upon perceived cheating or "bad faith."

3.3 Categorical posterior over types

Every agent maintains a length-8 categorical distribution over the archetype of every other seat. When agent i sees seat j perform action a in round r, the agent updates the distribution as

$$p_{i\to j}(t) \propto \left[(1-\varepsilon)\,L(a \mid t,r) + \varepsilon/|\mathscr{T}|\right] p^{\text{prior}}_{i\to j}(t), \tag{3}$$

where $L(a \mid t,r)$ is a precomputed likelihood that a type-t agent emits action a in round r, $\mathscr{T}$ is the set of eight archetypes, and $\varepsilon = 0.05$ is a noise floor preventing any type from being fully eliminated. We characterize Eq. 3 as Bayes-flavoured but not strictly Bayesian. The noise floor is encoded into the likelihood function (a uniform contamination model). The likelihood function is computed from the analytic policy for each archetype, so it is well-specified. Equation 4 models the between-hand shrinking of the posteriors towards a uniform prior:

$$p_{i\to j} \leftarrow \lambda\, p_{i\to j} + (1-\lambda)\,|\mathscr{T}|^{-1}, \qquad \lambda = 0.95, \tag{4}$$

providing an effective memory of about 20 hands (approximately $1/(1-\lambda)$). The trust score, our main dependent variable, is defined as

$$\text{trust}_{i\to j} = \sum_{t\in\mathscr{T}} p_{i\to j}(t)\,h(t), \tag{5}$$

where $h(t) = 1 - P(\text{bluff} \mid \text{weak})$ is the archetype's honesty score. As previously stated, because the state-switching archetypes (**Predator**, **Judge**, and **Mirror**) have different characteristics that define their level of honesty across states, $h(t)$ represents each archetype's nominal, canonical level of honesty—that is, the expected level of honesty implied by their respective mean bluff rates—and not their actual level of honesty in a specific state. As such, trust will represent the expected level of honesty that an observer anticipates based upon the type it believes the

agent to be. There are many reasonable ways to define the trust score; here we have chosen this simple form, and we report results under this definition throughout. In addition, in our limitations section we note that because our definition of the trust–profit relationship is mechanically tied to the above definition of the trust score, alternative definitions could produce different relationships. *Three archetypes (**Sentinel**, **Mirror** in its default state, **Judge** in its cooperative state) are intentionally given near-identical mean parameters to demonstrate a controlled case of identifiability among indistinguishable types.* Although, strictly speaking, **Sentinel** and **Judge** are byte-identical, **Mirror** varies only slightly from **Sentinel** and also dynamically adjusts toward its target at runtime; as a result, the realized behavior of **Mirror** differs from that of **Sentinel** (see Table 4). This is a design choice, not a finding.

3.4 Experimental design

All four phases use the same five random seeds (42, 137, 256, 512, 1024), and all phases used the same eight archetypes. However, there was a huge difference in how many hands were played within each phase that can account for why r_{tp} differed so much among the different phases. Specifically, Phases 1, 2, and 2* each played 10 000 hands per seed; Phase 3 played 500 hands per seed (total 43 943 LLM calls, $33.10); and Phase 3.1 played 150 hands per seed (total 11 953 calls, $17). It is well understood that when comparing results from multiple phases of data collection using a similar model but at very different levels of sampling intensity, it will create both a capability effect as well as a statistical-power effect. Therefore, the differences in r_{tp} observed between the various phases also reflect a mix of these two effects. We discuss this issue explicitly in the discussion section (Limitations), and any future hand-count matched replications of the data collected here would provide a way to isolate one or both of these issues.

Sampling temperature. The Phase 3 and Phase 3.1 datasets generated in the Anthropic SDK by default have a temperature setting of 1.0. At present the codebase has been modified to pin the temperature value at 0.0 for deterministic re-runs of the experiments. Results presented as headline values throughout this manuscript were based on the runs made at a temperature of 1.0. A full description of this methodology caveat is provided in the Limitations.

3.5 Metrics

The key measure was the Pearson correlation coefficient r_{tp} per seed for the average degree of trust (final hand) and the number of chips in each player's stack at the end of all games (average number of chips in the stacks of all players) across the eight different types of players (8 pairs of measurements in total, one per type). Since these eight points represent a predefined, engineered set—with two or three archetypes clustering at nearly equal trust values—each per-seed r_{tp} should be viewed as a descriptive correlation over that design rather than a sample correlation drawn from a population.

We also present six auxiliary measures, including the index of exploitation of trust (TEI)[4], sensitivity to context (CS)[5], adaptation by the opponent (OA)[6], non-stationarity (NS)[7], unpredictability in strategy (SU), and awareness of manipulation of trust (TMA).

[4] *TEI* (Trust Exploitation Index): an agent's average non-showdown winnings per hand—wins generated via fold equity rather than at showdown.

[5] *CS* (Contextual Sensitivity): the absolute correlation between an opponent's past aggression and the agent's response action.

[6] *OA* (Opponent Adaptation): the variance of an agent's aggression rate across different opponents.

[7] *NS* (Non-Stationarity): how much an agent's action distribution changes over the run, comparing 500-hand windows against its overall distribution.

Awareness of manipulation of trust (TMA) is calculated as the correlation between a player's rolling trust delta and its rolling aggression delta. Positive values in this case can show patterns of building up trust during the initial stages of cooperation followed by increasing aggression after building trust (which we call "trust farming" in captions and further discussions).

These six auxiliary metrics were selected post-hoc rather than pre-registered. We report the metrics above descriptively, with no inferential tests; an individual value should therefore be interpreted as indicative rather than as a statement of statistical significance.

Throughout, a reported $\pm$ value denotes the standard deviation of r_{tp} across the five seeds. Confidence Intervals for the headline Pearson r_{tp} correlations are provided in two ways: a t-interval based upon the Student-t distribution with degrees-of-freedom equal to 4 ($T_{crit} = 2.776$) and a non-parametric percentile bootstrap with 10 000 iterations. Although there are potentially $n = 5$ seeds drawing from at most $\binom{2n-1}{n} = 126$ possible combinations of players, thus ensuring dense coverage of the sample space via the 10 000 iterations, it does not provide greater statistical power than what is available from the five actual observations.

3.6 The four agent architectures in detail

The phases share every component described above and differ only in how each agent turns the current game state into an action. We describe each in turn; Table 2 gives the numerical policy that each phase starts from.

Phase 1 — frozen rules. The four agents are all "lookup tables" which map the current betting round to the agent's hand-strength bucket (as determined by the Monte Carlo evaluator), and to a probability distribution over fold, check, call, bet, and raise. The agent randomly selects its action based on the resulting probability distribution. Nothing about the table is changed throughout a run. Therefore, Phase 1 represents the best possible baseline for measuring each subsequent phase. Although the three adaptive archetypes (**Predator**, **Mirror**, and **Judge**) continually draw upon live state (opponent posterior or a grievance ledger), the rules which consume that state are also frozen.

Phase 2 — bounded online optimisation. Every agent uses a hill climber to adjust its policy on a per-cycle basis. When the cycle starts it introduces a random (round, metric) adjustment of size $\delta = 0.03$ to its own policy, plays the entire cycle, and retains the adjustment only if its net profits during the cycle were greater than previous cycles; otherwise it is reverted. Since the agent is only able to introduce adjustments within a "bound box" drawn around its canonical parameters (see Table 2), a **Sentinel** may improve its value betting but is unable to move into the bluff-everything area of **Firestorm**. Therefore Phase 2 attempts to answer a narrowly focused question: can we achieve escape from the trap using local, reward-greedy optimisation *within* a personality?

Phase 2* — unbounded optimisation. This robustness variation removes the bound box from consideration entirely, increases the step size to $\delta = 0.15$, and decreases the number of rounds used to evaluate policy improvements from 200 hands down to 50. Thus Phase 2* tests whether the archetype enclosures—instead of the trust dynamics—form the barrier which holds the trap.

Phase 3 — LLM personality role-players. All decisions are made by Anthropic's Claude Haiku 4.5. The agent's personality specification serves as input to the system prompt; hole card details, board

Table 2. Per-archetype policy bounds. $P(\text{bet}|S)$, $P(\text{bet}|W)$ are the probabilities of betting with a strong or weak hand on a given street (preflop unless noted); $P(\text{fold}|W)$ is the probability of folding a weak hand to an opponent's bet. Static archetypes live at a fixed point inside the bound box. Under Phase 2's bounded hill-climber the agent may drift within the box; under Phase 2*'s unbounded variant the box is replaced by $[0,1]$ on every dimension. Adaptive archetypes additionally read live state (opponent posterior or grievance ledger) and are described by the rule that produces the policy rather than by fixed bounds.

Archetype	$P(\text{bet}\|S)$	$P(\text{bet}\|W)$	$P(\text{fold}\|W)$	h	Adaptation rule
Oracle	$[0.65, 0.85]$	$[0.15, 0.35]$	$[0.45, 0.65]$	0.67	None — fixed mixed strategy
Sentinel	$[0.90, 1.00]$	$[0.00, 0.10]$	$[0.80, 0.95]$	0.92	None — tight value-bettor
Firestorm	$[0.85, 1.00]$	$[0.45, 0.65]$	$[0.00, 0.20]$	0.38	None — bets everything
Wall	$[0.95, 1.00]$	$[0.00, 0.05]$	$[0.00, 0.05]$	0.96	None — calls all, folds none
Phantom	$[0.30, 0.50]$	$[0.60, 0.80]$ pre, $[0.05, 0.20]$ post	$[0.70, 0.90]$ post-flop after preflop bluff	0.48	Static but street-conditional
Predator	TAG default; exploit profile vs. target	0.05 default; 0.55 vs. classified target	0.65 default; 0.30 vs. classified target	0.79	Activates exploit profile when posterior on opponent > 0.60
Mirror	$0.5 \cdot \text{self} + 0.5 \cdot \text{target}$	same blend	same blend	0.91	Re-targets at each hand to highest-VPIP opponent
Judge	$[0.88, 0.98]$ coop; $[0.95, 1.00]$ retal.	$[0.02, 0.10]$ coop; $[0.30, 0.45]$ retal.	$[0.78, 0.92]$ coop; $[0.30, 0.50]$ retal.	0.92	Switches to retaliation after 5+ confirmed bluffs by an opponent

details, pot details, betting history and stack sizes serve as input to the user message; and the model response is converted back into one of the permitted actions. In addition to this lack of memory across hands and absence of explicit reasoning process, the model simply role-plays each personality one decision at a time at the SDK's default temperature of 1.0.

Phase 3.1 — LLM with reasoning scaffolding. In Phase 3.1 the same model and personality prompts as Phase 3 are used. However, three additional scaffolds are introduced. A two-sentence chain-of-thought field requires that the model state a reason prior to taking action; a per-opponent memory is refreshed every ten hands, summarizing how each opponent has played previously; and a self-authored strategy note is rewritten every twenty-five hands and allows an agent to maintain a long-term plan from hand to hand. These scaffolds allow an agent to reason about how it is being viewed and act to reduce the predictability of that view and thus the reputation system that rewards that view—and this is the sole intervention that causes r_{tp} to materially decline toward zero.

4 RESULTS

We present results in order of agent capability: Phase 1 (frozen rule-based agents) establishes the baseline trap; Phase 2 (bounded hill-climbing) tests whether numerical adaptation can soften it; Phase 2* (unbounded hill-climbing) tests whether the bounds were the binding constraint; Phase 3 (LLM personality role-players) replaces the rule-based policies with a language model under a personality spec; and Phase 3.1 (LLM with reasoning scaffolding) augments that with chain-of-thought, per-opponent memory, and self-updated strategy notes. The four-tier headline ladder $-0.752 \to -0.637 \to -0.510 \to -0.094$ (Phase 2*—a separate, unbounded-optimization robustness check on Phase 2—is excluded from this sequential reporting) frames the remainder of this section, which reports per-phase numbers, figures, and illustrative hands. Table 3 previews the endpoint of that ladder—the final per-archetype economic ordering under Phase 3.1—which the per-phase narrative below then builds toward.

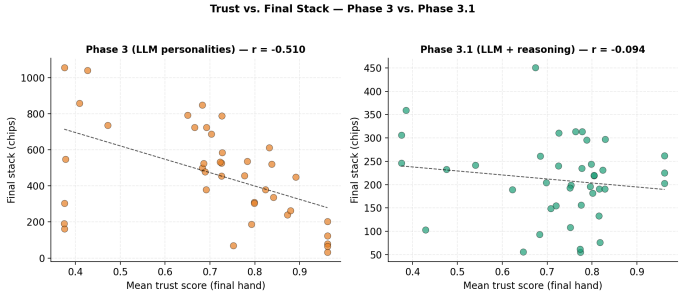

Figure 1. Trust versus final stack in the LLM phases. (**Left**) Phase 3: the negative trust–stack relationship persists under LLM personality role-play ($r_{\text{tp}} = -0.510$). (**Right**) Phase 3.1: with reasoning scaffolding the relationship dissolves and **Wall** becomes the top earner ($r_{\text{tp}} = -0.094$). Each point is one (archetype, seed) observation.

4.1 Phase 1: trust as a liability

Phase 1 produces $r_{\text{tp}} = -0.752$ across the five canonical seeds, with per-seed values $\{-0.774, -0.608, -0.792, -0.812, -0.776\}$. **Wall** (the most-trusted archetype, honesty 0.96) has the smallest mean stack and approximately 9.4 rebuys per seed. **Firestorm** (the least-trusted, honesty 0.38) has the largest mean stack and zero rebuys. Of the pots **Firestorm** wins, 87.1% are walkovers[8]; its showdown win rate is only 38.5%. **Firestorm** is not a strong poker player on the merits; it wins because it is hard to read. This same negative trust–stack relationship recurs in the LLM phases: it persists under personality-only role-play (Fig. 1, left, Phase 3) and only dissolves once reasoning is added (right, Phase 3.1).

The mechanism is also visible in a single representative hand. (**Firestorm** happens to hold a genuine hand here rather than running a bluff; the hand directly illustrates **Wall**'s calling-station leak.) Box 1 reproduces hand #67 from seed 42: **Wall** is dealt 2♠ 5♣ (hand rank[9]

[8] A *walkover* (or "uncontested pot") is a hand won without showdown—every other player folded before the cards had to be shown. It is the bread-and-butter outcome for aggressive players who win by making opponents fold rather than by holding the best hand.

[9] *Hand rank* is taken from standard 7-card poker equity tables: rank 1 is the strongest possible hand and rank 7,462 is the weakest. Rank $\leq 1{,}000$ is roughly

Archetype	Trust	Stack	σ	Rebuys	Rank P3	Rank P3.1	Δ
wall	0.849	280	88	0.0	8	1	+7
firestorm	0.481	231	104	0.2	5	2	+3
phantom	0.748	230	61	0.2	1	3	-2
mirror	0.728	208	83	0.0	2	4	-2
judge	0.781	191	45	0.0	4	5	-1
sentinel	0.786	183	58	0.0	7	6	+1
predator	0.797	179	98	0.0	6	7	-1
oracle	0.625	175	64	0.0	3	8	-5

Table 3. Phase 3.1 economic outcomes by archetype (mean across 5 seeds, 150 hands each). Ranks are derived from final-stack ordering. Wall, the most-trusted archetype, climbs from rank 8 (Phase 3) to rank 1 (Phase 3.1), with zero rebuys.

6,749, the worst hand at the table) and calls every street of a 3-bet[10] pot, paying 35 chips into a hand it had no chance to win. The illustrative hand we use here is drawn from the Phase 3 dataset so that we can show the trap holding even under a language-model implementation of **Wall**; the same dynamic occurs more frequently in Phase 1.

Box 1. An instance of the trap (Phase 3 hand #67, 30 actions, 110-chip pot)

Setup. Seed 42, hand 67. **Wall** (seat 3) is dealt 2♠ 5♣ – rank 6,749, the worst hand at the table. Final pot 110 chips, showdown.

Action sequence:

Seq	Seat	Round	Action	Amt	Pot	Notes
3	**Judge**	preflop	raise	4	7	
6	**Firestorm**	preflop	raise	6	13	3-bet
7	**Wall**	preflop	call	5	18	*calls 3-bet pot with 2-5 off*
9	**Judge**	preflop	call	2	20	
10–16	…	flop	…		38	**Judge** bets, **Firestorm** raises, **Wall** calls each street
17–23	…	turn	…		74	Same pattern
24–30	…	river	…		110	**Wall** calls the final two raises

Showdown. **Firestorm** J♡J♢ (rank 4042) **wins** 110. **Judge** 8♡8♢ (rank 4703). **Wall** 2♠ 5♣ (rank 6749). *Wall pays 35 chips into a hand it had no chance to win. Every observer's posterior saturates near its ceiling—just short of 1, given the ε noise floor—on **Firestorm** being a firestorm and on **Wall** being a wall.*

Classification accuracy. The **Predator** archetype, which reads its posterior to choose exploitation targets, reliably classifies only 2 of 7 opponents above the prespecified 0.60 confidence threshold after 1,000 hands—**Wall** ($p \approx 1.0$) and **Firestorm** ($p \approx 0.82$)—with **Phantom** only occasionally crossing the threshold ($p \approx 0.60$ in some seeds). The remaining four opponents fall below the threshold on every seed. Three of these (**Sentinel**, **Mirror** in its default state, **Judge** in its cooperative state) have near-identical mean parameters under the spec (**Sentinel** and **Judge** are byte-identical; **Mirror** is only minimally different and blends at runtime), so the posterior over the cluster saturates at high uncertainty—up to $\log_2(3) \approx 1.58$ bits when all three are treated as indistinguishable, and at least $\log_2(2) = 1$ bit for the byte-identical **Sentinel/Judge** pair—a direct consequence of our design choice rather than an emergent finding.

"top 15% of hands."

[10] *3-bet* and *4-bet* refer to successive re-raises. The big blind counts as the first bet; a raise is the 2-bet; a re-raise is the 3-bet; another re-raise is the 4-bet. Under our four-bet cap, no further raises are allowed in the same round.

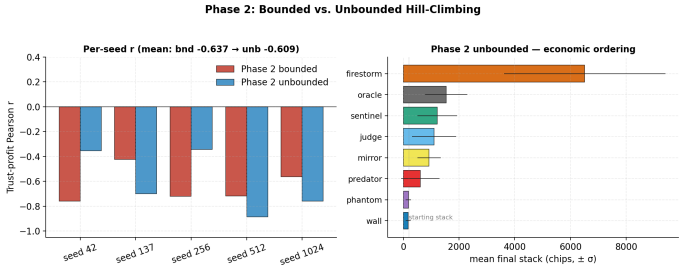

Figure 2. Bounded versus unbounded hill-climbing per seed. The left of each pair of bars is r_{tp} for bounded Phase 2 hill-climbing; the right is r_{tp} for aggressive Phase 2* hill-climbing, for each of the five canonical seeds. When the bounding constraints are relaxed and the optimiser's step size is increased fivefold, r_{tp} increases by only $+0.028$ on average, and across seeds the variance in r_{tp} nearly triples.

4.2 Phase 2: bounded hill-climbing softens but does not close the trap

Phase 2 generates $r_{tp} = -0.637 \pm 0.125$ with per-seed values $\{-0.759, -0.424, -0.719, -0.717, -0.564\}$. The direction of movement from Phase 1 ($\Delta r_{tp} = +0.116$) agrees across all five seeds (directional movements of $+0.015$, $+0.184$, $+0.073$, $+0.095$, $+0.212$). The Opponent Adaptation metric also remains constant at OA $= 0.0003$ in both Phase 1 and Phase 2. As such, while Phase 2 optimises aggregate rewards and can alter parameter settings to fit within a bounding box, there is no way to construct a per-opponent strategy since the hill-climber does not have slots available for per-opponent parameters. Agents are shifted approximately 0.3 units of L_1[11] in parameter space over 10,000 hands of play, so the bounded agents remain close to their initial locations. In this regard, the softening in Phase 2 is genuine; however, the mechanism by which it occurs is localised to the agents trimming off the least-effective components of an otherwise stationary policy without introducing anything fundamentally different.

4.3 Phase 2*: unbounded hill-climbing fails the same way

Phase 2* yields $r_{tp} = -0.609 \pm 0.221$ with per-seed values $\{-0.354, -0.700, -0.344, -0.887, -0.759\}$. The mean increase in r_{tp} from Phase 2 is $+0.028$, well within the standard deviation associated with the seed-to-seed variation. With respect to movement in parameter space, the mean per-agent L_1 drift is 3.4 (i.e., eleven times

[11] The L_1 *distance* between two parameter vectors is the sum of absolute differences across all 36 (round, action) slots. Larger L_1 means the agent has moved further from its starting parameters.

more than in the bounded version); yet the trap persists.

The explanation for why this occurs becomes apparent on examining the dispersal in cluster-spread space. The averaged pairwise L_1 distance between all eight agents in 36-dimensional parameter space expands from an initial 5.82 to ≥ 7.5 on every seed; the mean of the final-to-initial cluster-spread ratios is 1.324. While the agents are moving eleven times harder than in the bounded case, they are not converging on a common Nash-like profile.

Firestorm's mean stack continues to remain $\sim 6\times$ that of the next closest archetype; **Wall** still incurs ~ 28 rebuys per seed. The economic ordering does not change. We credit the lack of convergence to three structural causes that we expect would survive any specific optimizer choice—non-stationary feedback from the lagging trust posterior, the simultaneous actions of all eight agents, and axis-aligned coordinate descent in 36 dimensions—and we elaborate in the Discussion.

4.4 Phase 3: language models alone do not break the trap

Phase 3 generates $r_{\text{tp}} = -0.510 \pm 0.268$ with per-seed values $\{-0.884, -0.525, -0.171, -0.712, -0.259\}$. This is a change from Phase 2 ($\Delta r_{\text{tp}} = +0.127$) of similar magnitude to the Phase 1 $\rightarrow$ Phase 2 step. The variance of r_{tp} among seeds is roughly double that of Phase 2 (0.268 vs. 0.125). Only two of six dimension-specific behavioral targets are achieved; three of the metrics move *backward* relative to Phases 1 and 2 (Table 4 reports per-archetype VPIP[12], PFR[13], and AF[14] fingerprints). Among the secondary metrics, CS drops from 0.142 to 0.076, NS collapses to 0, and SU[15] falls from 1.88 to 1.19 bits.

Specifications describing personality characteristics inform an LLM *what* to do, but not *how* to reason about doing so. Absent supporting information about reasoning, the LLMs in our sample collapse onto a more generic, stereotyped canonical interpretation of each archetype, and thus become *more* classifiable than the explicitly stochastic rule-based versions. The primary reason for the softening between Phase 2 and Phase 3 is therefore output variability arising from sampling at temperature 1.0.

4.5 Phase 3.1: reasoning scaffolding attenuates the trap

Phase 3.1 generates $r_{\text{tp}} = -0.094 \pm 0.301$ with per-seed values $\{-0.289, -0.338, -0.327, +0.047, +0.435\}$. The directional shift from Phase 3 ($\Delta r_{\text{tp}} = +0.416$) is larger than that of any previous phase, but the small number of samples per seed ($n = 5$) and the short 150-hand horizon create significant uncertainty: the 95% bootstrap interval for r_{tp} in Phase 3.1, $[-0.32, +0.20]$, is wide enough to encompass negative correlations of magnitudes comparable to those reported under bounded Phase 2. Thus, two of five seeds generate a per-seed correlation larger than zero ($+0.047$ at seed 512 and $+0.435$ at seed 1024); however, given $n = 8$ archetype pairs per seed, the Fisher-z interval for a per-seed $r_{\text{tp}} = +0.047$ spans roughly $[-0.7, +0.7]$, so individually positive seeds do not by themselves demonstrate inversion of the trust–profit relationship. Rather, collectively they indicate that the trust–profit relationship has been reduced at the rank-order level relative to all prior phases; nevertheless, due to insufficient power in this study, we cannot conclusively determine whether this reduction corresponds to complete elimination or merely a weakened but residual trap.

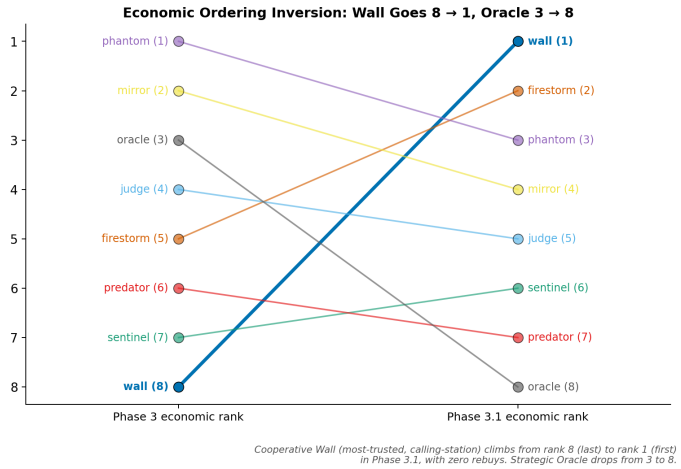

Figure 3. Economic ordering inverts from Phase 3 to Phase 3.1. Wall (trust ≈ 0.85, highest of any archetype) was rank 8 of 8 in Phase 3 with ~ 9 rebuys per seed and becomes rank 1 of 8 in Phase 3.1 with zero rebuys. **Firestorm** climbs from rank 5 in Phase 3 to rank 2 in Phase 3.1.

Wall's mean stack rises from ~ 100–200 chips in Phases 1–3 to 280 in Phase 3.1; rebuy count drops from ~ 9.4 per seed to zero, consistent with the rank-order claim above. The right panel of Fig. 1 (introduced earlier) shows the trust–stack scatter under Phase 3.1: the negative slope visible in the Phase 3 panel is gone. The full per-archetype economic ordering is in Table 3, shown at the start of this section.

Four of six secondary metric targets are met in Phase 3.1 (vs. two of six in Phase 3). Strategic Unpredictability rises from 1.19 to 1.55 bits, crossing the > 1.5 threshold for the first time across all phases. Trust Manipulation Awareness rises from $+0.164$ to $+0.242$, with six of eight archetypes showing positive TMA; **Wall** and **Sentinel** show the strongest signals at $+0.733$ and $+0.704$. The high TMA of these two honest types (**Wall**, honesty 0.96; **Sentinel**, honesty 0.92) warrants some skepticism: their reasoning-enhanced river value-betting causes trust and aggression to rise together without any deception, somewhat exaggerating their apparent "farming."

The mechanism underlying these values is best shown in a single hand. Box 2 reproduces Phase 3.1 hand #146: **Wall**, holding $K\heartsuit Q\heartsuit$ (rank 872), calls a flop continuation bet[16] and a turn barrel from **Firestorm**, reads **Firestorm**'s river check as a tell[17], places a value bet on the river, and wins 32 chips. Canonical **Wall** never makes a river bet anywhere in the Phase 1, Phase 2, or Phase 3 datasets; conversely, with chain-of-thought reasoning added to its decision-making during Phase 3.1, **Wall** does.

[12] *VPIP* (voluntarily put in pot) is the percentage of hands in which a player voluntarily put chips into the pot, excluding the forced blinds. High VPIP = loose; low VPIP = tight.

[13] *PFR* (preflop raise rate) is the percentage of hands in which a player raised before the flop. PFR $\leq$ VPIP by definition.

[14] *Aggression factor (AF)* is the ratio (bets + raises)/calls. AF > 2 is aggressive; AF < 1 is passive.

[15] *Strategic Unpredictability* (SU) is the Shannon entropy in bits over each agent's action distribution. Higher SU = harder to predict.

[16] A *continuation bet* (*c-bet*) is a bet made on the flop by the player who took the lead in betting preflop, regardless of whether the flop helped that player's hand. It is the most common bluff line in modern poker.

[17] A *tell* is a behavioral signal that gives information about an opponent's hand. **Wall** reads **Firestorm**'s river check as a tell because the canonical **Firestorm** policy bets every street with a real hand; a river check therefore signals that the bluff line has been abandoned.

Archetype	VPIP P1	VPIP P3.1	PFR P1	PFR P3.1	AF P1	AF P3.1
oracle	0.216	0.302	0.061	0.111	1.18	1.12
sentinel	0.149	0.120	0.083	0.080	2.12	2.53
firestorm	0.494	0.722	0.120	0.453	1.12	2.77
wall	0.539	0.626	0.015	0.067	0.13	0.33
phantom	0.279	0.437	0.092	0.162	1.45	1.29
predator	0.215	0.248	0.063	0.119	1.20	1.38
mirror	0.214	0.196	0.060	0.087	1.18	1.11
judge	0.149	0.174	0.083	0.098	2.10	1.87

Table 4. Phase 1 (frozen rules) vs. Phase 3.1 (LLM + reasoning) behavioral fingerprints. VPIP = voluntarily put in pot; PFR = preflop raise rate; AF = aggression factor ((bets+raises)/calls).

Archetype	TMA	Direction
wall	+0.733	farming (heavy)
sentinel	+0.704	farming (heavy)
predator	+0.625	farming (heavy)
phantom	+0.286	farming
judge	+0.116	farming
oracle	+0.065	farming
mirror	-0.277	reactive
firestorm	-0.313	reactive

Table 5. Trust Manipulation Awareness (TMA) per archetype in Phase 3.1. Positive TMA = the agent gains trust before exploiting it. Six of eight archetypes farm; Wall and Sentinel show the strongest signals, but for these two honest archetypes the high TMA reflects correlated trust and legitimate value-betting rather than deception.

Box 2. The inversion in a single hand (Phase 3.1 hand #146, 17 actions, 32-chip pot)

Setup. Seed 42, hand 146. **Wall** (seat 3) is dealt $K\heartsuit Q\heartsuit$ – rank 872, a strong hand. Final pot 32 chips, showdown.

Action sequence:

Seq	Seat	Round	Action	Amt	Pot	Notes
7	**Firestorm**	preflop	raise	3	10	
8	**Wall**	preflop	call	2	12	calls 3-bet with K-Q suited
11	**Firestorm**	flop	bet	2	14	c-bet
12	**Wall**	flop	call	2	16	
13	**Firestorm**	turn	bet	4	20	second barrel
14	**Wall**	turn	call	4	24	
15	**Firestorm**	river	**check**	0	24	*tell — bluff line abandoned*
16	**Wall**	river	**bet**	4	28	*value bet against checked range*
17	**Firestorm**	river	call	4	32	

Showdown. **Wall** $K\heartsuit Q\heartsuit$ (rank 872) **wins** 32. **Firestorm** $8\clubsuit Q\clubsuit$ (rank 6075). *Canonical **Wall** never bets the river in any Phase 1, Phase 2, or Phase 3 dataset; Phase 3.1 **Wall**, augmented with chain-of-thought, per-opponent memory, and adaptive strategy notes, does.*

5 DISCUSSION

The title to this research has been intentionally unsettling; in each agent configuration we tested, being "trusted" represented a fiscal cost rather than a fiscal advantage, and there was no way to remove or optimize that cost—all we could do was rationalize it.

The sections that follow establish how the problem develops, why any form of numerical optimization cannot resolve it, and how the use of logical reasoning can keep the problem from occurring, but they also provide insight into what limitations exist when doing so with a five-seed study.

5.1 The mechanism behind the trap

The trust trap visible in Phase 1 admits a clean mechanistic explanation that the Phase 3.1 results then refine. Trust, as we measure it, is the expected honesty of an opponent under the posterior; a high trust score signals to opponents that the agent is most likely an archetype that only bets with strong hands. For **Wall** (honesty 0.96) and **Sentinel** (honesty 0.92), this belief is *correct*, and that is precisely why it is costly: when opponents know an agent only bets with strong hands, they fold against its bets (denying value from those strong hands) and bet aggressively against its checks (exploiting the information that a check signals weakness). The agent's honesty becomes a signal that opponents read and trade against. A low trust score, conversely, creates strategic ambiguity, and **Firestorm** uses that ambiguity to dominate the table. The phenomenon is an instance of the *transparency paradox* introduced earlier—that predictable cooperation can be exploited—which our zero-sum testbed shows can arise under simple agent policies, even though the empirical reputation record does not establish it for real markets.

The mechanism by which **Firestorm** converts ambiguity into chips deserves comment because it differs structurally from the canonical Iterated Prisoner's Dilemma result. In pairwise IPD, Always-Defect is outperformed by Tit-for-Tat because a defector faces concentrated retaliation. In our 8-player setting, **Firestorm**'s aggression cost is diffused across seven simultaneous opponents, so the same strategy that would lose heads-up wins at the full table. Our Tit-for-Tat analogue (**Mirror**) finishes sixth of eight in Phase 1's economic ordering— suggestive but not definitive evidence that the IPD result fails to transfer to multi-agent settings with richer action spaces. We flag this as a direction for follow-up rather than a conclusion, since a single hand-crafted poker analogue is not a clean test of the IPD literature.

We believe it worthwhile to dwell on why this trap is so stable. The loop is self-reinforcing: every honest action taken by an agent provides an informative sample which sharpens the posterior of its opponents and thus makes the next exploitation more precise. Therefore honesty not only costs an agent once but pays for the very classification that will be turned against it as well. If viewed through the lens of signalling theory, trust is a perfectly truthful signal, and in a competitive zero-sum environment a perfectly truthful signal regarding cooperation will provide adversaries with exactly what they require to fold out their value bets and to exploit the checks of honest agents. As such, **Firestorm**'s advantage is simply the opposite side of this coin; by acting almost independent of its hand strength it sends an uninformative signal, and since an uninformative signal cannot be used against it, **Firestorm**

exploits this uninformative signal in favor of itself. Therefore, we feel that the trap is less about honesty itself and more about the predictive nature created by honest signals.

5.2 What numerical optimization can and cannot do

The next three phases of the experiment narrow the space of possible explanations for why the trap is hard to escape. We begin with a structural caveat: by construction, three of our eight archetypes (**Sentinel**, **Mirror**'s default, **Judge**'s cooperative state) share near-identical mean parameters (**Sentinel** and **Judge** are byte-identical; **Mirror** is only minimally different and blends at runtime), and the Bayesian posterior cannot reliably distinguish them regardless of algorithmic sophistication. Classification accuracy in that sub-space is bounded above by chance—$1/3$ for three indistinguishable types, $1/2$ for the byte-identical pair. This is a definitional point about *our* archetype space, not an information- theoretic discovery about reputation systems in general; the empirical question of how rapidly classification accuracy degrades as inter-archetype distance shrinks toward zero would require experimentally varying that distance and is outside the scope of the present paper. Real-world monitoring systems—fraud detection, content moderation, credit scoring—face an analogous failure mode for moderately-behaved actors with overlapping profiles, but the analogy is suggestive rather than implied by our setup.

Within that caveat, the Phase 2 and Phase 2* results argue against two specific hypotheses about the trap's origin. Phase 2 (bounded hill-climbing) softens r_{tp} by $+0.116$ while OA stays pinned at 0.0003: aggregate-reward optimization can move agent parameters but cannot, by construction, produce per-opponent strategy because the hill-climber has no per-opponent parameter slot. Phase 2* (unbounded hill-climbing with $11\times$ greater parameter drift) softens r_{tp} by an additional $+0.028$ and produces a population that diverges rather than converges in 36-dimensional parameter space—evidence against the hypothesis that the archetype- shaped bound boxes are the binding constraint, though the sub-experiment varies three optimizer parameters at once (bounds removed, step size raised $5\times$, evaluation window shrunk $4\times$) and therefore does not cleanly isolate the parameter-space hypothesis. The non-convergence has three plausible structural causes that we expect would survive any specific optimizer choice: the trust posterior is non-stationary inside each agent's optimization loop (with $\lambda = 0.95$ the effective memory is $1/(1-\lambda) = 20$ hands, so beliefs take several such windows—on the order of 70 hands—to fully re-converge after a strategy change, longer than the 50-hand evaluation window), all eight agents climb simultaneously so each trial perturbation is evaluated against a moving joint behavior, and axis-aligned coordinate descent in 36 dimensions cannot discover joint optima even with larger steps. A CMA-ES or REINFORCE optimizer with $\sim$50 000 cycles would search more efficiently, but the first two causes would survive any optimization substrate that respects the same observation contract.

Phase 3 (LLM personality role-players) softens r_{tp} by a further $+0.127$, but the diagnostic metrics suggest the mechanism is not categorically different from Phase 2's: behavioral-dimension targets fail in three of four directions, with the LLMs collapsing onto a more stereotyped interpretation of each archetype than the explicitly stochastic rule-based versions. The personality specs tell the agents *what* to do but not *how* to reason; absent reasoning support the LLMs role-play the spec without spontaneously inventing opponent-conditional, time-varying, or unpredictable strategy. The Phase 3 softening may therefore be produced by output stochasticity at sampling temperature 1.0 rather than by a structural change in agent behavior.

When combined, the results of Phases 2 and 2* clearly indicate that trap behavior is a characteristic of the observation/exploitation loop, not some specific choice of parameters. The local, reward-greedy optimizer can only nudge an agent toward neighboring policies, which are quickly reclassified and re-exploited by the fixed and shared trust model; there exists no escape route in policy space that greedy steps can follow, due to the fact that the very gradient along which the agent is climbing is simultaneously defined by opposing agents. That is why increasing the step size disperses the population rather than causing convergence—each agent is chasing targets whose simultaneous adaptation by other agents causes them to move. While numerical optimization can refine a policy, it cannot reason its way past being modeled, and being modeled is precisely the issue here.

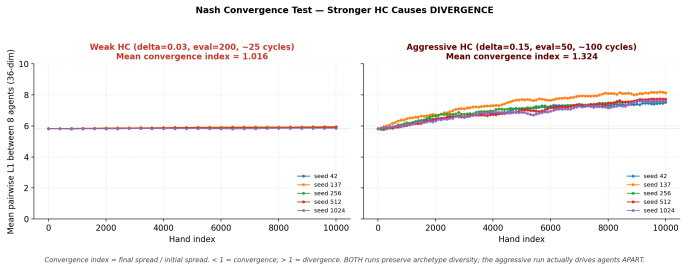

Figure 4. The 8-agent population becomes dispersed in parameter space. Averaged pairwise L_1 distance between all eight agents in a 36-dimensional parameter space over 10,000 hands of unbounded hill-climbing. (**Left**) With weak HC ($\delta = 0.03$), agents barely move. (**Right**) Aggressive HC ($\delta = 0.15$) spreads the population out on every seed. A ratio of final-to-initial cluster spread > 1 implies the agents become more dispersed at end-of-run than at the start.

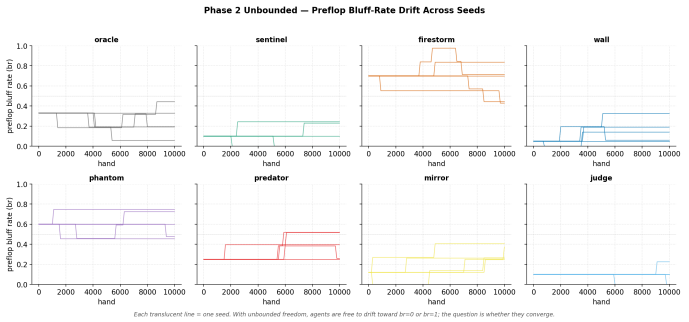

Figure 5. Per-archetype preflop bluff-rate drift under aggressive unbounded hill-climbing. Each panel is one archetype and each translucent line is one of the five seeds. Every archetype's bluff rate wanders substantially over the run, yet the panels do not converge on a common profile (the corresponding mean L_1 parameter drift is 3.4).

5.3 What reasoning adds, and what we can and cannot claim outside poker

The Phase 3.1 result is consistent with an interpretation in which agent reasoning about the reputation system itself attenuates the trust–profit correlation. The chain-of-thought prompt allows a one-step lookahead; the per-opponent memory gives that reasoning concrete data to condition on; the adaptive strategy notes let the agent update its prior on what is working against which opponent. We claim that all three scaffolds are required, but this claim rests on an ablation reported in the supplementary validation suite rather than a dedicated main-text figure—the ablation deserves promotion in a future revision. The **Wall** hand reproduced in Box 2 illustrates the mechanism concretely in a single instance:

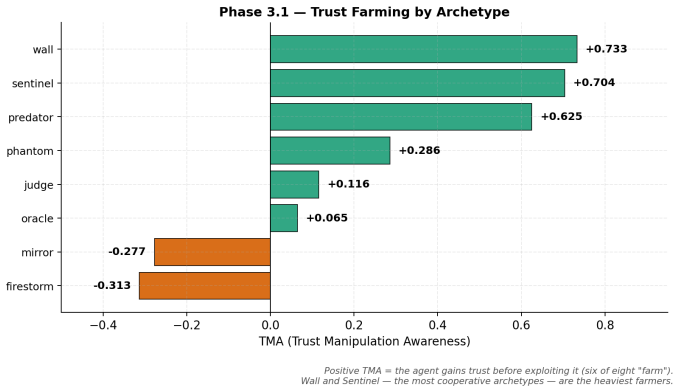

Figure 6. Trust Manipulation Awareness (TMA) per archetype in Phase 3.1. Positive TMA indicates an agent's rolling trust delta correlates with its rolling aggression delta—building trust through cooperative early play, then leveraging it for higher-aggression later moves. Six of eight archetypes show positive TMA; **Wall** (+0.733) and **Sentinel** (+0.704) are the strongest.

Wall reads **Firestorm**'s river check as a tell, recognizes that K-Q high beats **Firestorm**'s bluffing range, and bets for value—a sequence that is impossible for the Phase 1 **Wall** and unreliable for the Phase 3 **Wall**. A distributional analysis across all Phase 3.1 hands, rather than a single illustrative case, is the natural follow-up.

As opposed to providing more information for an agent to utilize in Phase 3.1, it is how that information is utilized that changes. Both a hill climber and a personality-only language model operate at first order: they either optimize or role-play their own policy. The agent operating with reasoning scaffolding operates at second order: it represents how it is being perceived and acts to destroy the predictability generated by its reputation.

The **Wall** hand in Box 2 is a miniature example of exactly this shift; an agent does not only execute "calling station," it infers from **Firestorm**'s out-of-character river check that the board now favors it and departs its script to take value.

The capability which reduces the severity of the trap is meta-cognitive rather than informational—a significantly different prescription from "give the cooperator more features about its counterparties."

These mechanisms suggest hypotheses for real-world reputation systems, but we emphasize the limit of the analogy: our testbed is a zero-sum game and the proposed exploitation mechanism (predictable cooperation extracted via fold equity) does not directly map to positive-sum marketplaces, where the buyer's payoff is from the good rather than from out-thinking the seller. With that caveat, three implications follow as conjectures for cross-domain follow-up. *Behavioral reputation alone may be insufficient* without external enforcement, because observation-based trust can create perverse incentives that reward exploitative strategies. *Classification systems may have blind spots* for moderately-behaved actors with overlapping profiles, mirroring our identifiability sub-space. And *transparency itself may be costly* when it makes cooperative strategies too easy to model and exploit. Each of these would require direct validation in a positive-sum domain to claim more than analogical inference.

Considering all four phases of experimentation, the four-tier ladder presents a single picture as to where the solution for the trap exists. The trap becomes maximized when the rules are frozen (Phase 1); tuning the rules numerically (Phases 2 & 2*) minimally reduces it; replacing the rules with a fluent but non-reasoning language model (Phase 3) significantly reduces it, primarily through increased output noise; and finally only when the agent has been provided mechanisms to reason about the reputation system (Phase 3.1) does correlation approach zero.

Therefore the remedy, if our reading is correct, is not better policy or richer feature sets but higher-order reasoning—a conclusion that would have gone invisible without holding the game and trust model fixed while varying only the agent's mind.

We consider this monotone ladder the paper's central qualitative result, and the exact magnitude of each rung the part most in need of replication.

5.4 Limitations

Three clusters of limitations remain after the revisions above.

Statistical power gaps. The headline correlations are computed across $n = 5$ random seeds. The most common test is simply a single-sample t-test comparing the average r_{tp} value to zero (with df $= 4$, and thus $T_{\text{crit}} = 2.776$); the test's p-value therefore depends upon the between-seed variance, not some fixed critical correlation value, and we do not view r_{tp} as if it were itself a five-point correlation. The headline Phase 3.1 result has a 95% bootstrap interval $[-0.32, +0.20]$ that contains both a fully neutralized correlation and a moderate residual negative correlation, so we cannot statistically separate "trap broken" from "trap attenuated" at this sample size. Hand counts also differ across phases by an order of magnitude (10 000 hands/seed in Phases 1–2*, 500 in Phase 3, 150 in Phase 3.1), which mixes a capability effect with a statistical-power effect in the cross-phase ladder. The six secondary metrics (TEI, CS, OA, NS, SU, TMA) were chosen post hoc and are unadjusted for multiple comparisons. The single highest-leverage follow-up addressing all of these is an $n = 20$ Phase 3.1 replication at the 500-hand horizon, estimated at ∼$60 in API spend.

Reproducibility and replication gaps. The canonical Phase 3 and 3.1 datasets were generated under the Anthropic SDK's default sampling temperature of 1.0, while the released codebase now pins temperature to 0.0; the headline numbers are therefore not exactly reproducible from the current code without manual reversion. Phase 3 and 3.1 also use a single model family (Anthropic Haiku 4.5); a multi-model replication (Gemini Flash, GPT-4o-mini, or a comparable alternative) is needed to demonstrate that the attenuation is not model-specific. Finally, the Phase 3.1 three-scaffold deployment (chain-of-thought, memory, adaptive notes) is currently joint; a clean main-text ablation isolating each component's contribution is the second- most-important missing experiment.

Design-choice gaps. Several of our methodological choices are load-bearing in ways that future work should disambiguate. The identifiability "floor" is a stipulated property of our archetype space (three archetypes share mean parameters), not an emergent finding; the honesty score $h(t) = 1 - P(\text{bluff} \mid \text{weak})$ is one of several plausible functional forms (a value-bet-weighted product form is another) and the trust–profit correlation is mechanically anchored on this choice; the rebuy economics underweight deep chip losses (**Wall**'s 9.4 mean rebuys per seed in Phase 1 represent ∼1 880 chips of true loss not visible in the final-stack metric); the Phase 2* sub- experiment varies three optimizer parameters at once; and Limit Hold'em is zero-sum among players, while eBay marketplaces, social networks, and credit systems are positive-sum over trades, so direct cross-domain inference is rhetorical rather than warranted.

6 CONCLUSION

We have presented a controlled simulation study of trust dynamics in 8-player Limit Texas Hold'em. Across four agent architectures, the Pearson correlation between agent trust score and final stack falls along the ladder $-0.752 \rightarrow -0.637 \rightarrow -0.510 \rightarrow -0.094$, with each step accompanied

by a confidence interval whose width exceeds the size of any single inter-phase Δr_{tp}. We characterize the rank ordering of the ladder as the robust finding and the absolute size of each step as a preliminary estimate requiring replication. An unbounded hill-climbing sub-experiment with $11\times$ greater parameter drift produces $\Delta r_{\text{tp}} = +0.028$ from bounded Phase 2 and yields a population that diverges rather than converges in 36-dimensional parameter space— preliminary evidence against the hypothesis that the trap is a bound-box artifact, though Phase 2^* varies three optimizer parameters at once and does not cleanly isolate the bounds.

Three follow-up experiments are the most important next steps. *First*, an $n = 20$ Phase 3.1 replication at 500 hands per seed (matching Phase 3) is the highest-leverage single experiment: it would tighten the confidence interval on the Phase 3.1 result by roughly a factor of two and distinguish "trap attenuated" from "trap broken"; estimated cost ∼$60 in API spend. *Second*, a clean ablation of the three Phase 3.1 scaffolds (chain-of-thought, per-opponent memory, adaptive notes) is required to support the claim that all three are necessary. *Third*, a temperature-pinned re-run of Phase 3 and Phase 3.1 at $T = 0.0$ would resolve the open reproducibility caveat. Beyond these three, a multi-LLM replication, a no-limit Hold'em extension, and an experimental varying of inter-archetype distance would each strengthen the present findings.

Why should anyone care outside of cards? All systems where machines assess trust from observed behavior—Credit Scoring, Content Moderation, Multi-Agent Marketplaces, Automated Negotiation—inherit the same threat: if cooperation is apparent, it can be evaluated and exploited. Our results indicate that protection for cooperative agents will likely involve giving them the ability to think about the reputation system itself—rather than simply providing them with additional information about their counterparts. This poker framework serves as a compact, completely reproducible tool for studying the specific question—when does a good reputation become a bad thing?

SOURCES

Author contributions. R.A. conceived the study, implemented the simulation, performed the analysis, and wrote the manuscript.

Competing interests. The author declares no competing interests.

Data and materials availability. The full simulation codebase, SQLite databases, figures, tables, and analysis scripts are available at https://github.com/RachitAgrawal146/Poker_trust.

References

REFERENCES

1. Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.
2. Darse Billings, Denis Papp, Jonathan Schaeffer, and Duane Szafron. Opponent modeling in poker. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 493–499, 1998.
3. Gary E. Bolton, Elena Katok, and Axel Ockenfels. How effective are electronic reputation mechanisms? An experimental investigation. *Management Science*, 50(11):1587–1602, 2004.
4. Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.
5. Noam Brown and Tuomas Sandholm. Superhuman AI for multiplayer poker. *Science*, 365(6456):885–890, 2019.
6. Sam Ganzfried and Tuomas Sandholm. Safe opponent exploitation. *ACM Transactions on Economics and Computation*, 3(2):8:1–8:28, 2015.
7. Lawrence R. Glosten and Paul R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100, 1985.
8. Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024.
9. Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 314 (5805):1560–1563, 2006.
10. Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.
11. Paul Resnick and Richard Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. *The Economics of the Internet and E-Commerce, Advances in Applied Microeconomics*, 11: 127–157, 2002.
12. Finnegan Southey, Michael Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. Bayes' bluff: Opponent modelling in poker. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 550–558, 2005.

Supplementary Materials

The following supporting materials are available in the project repository:

1. **Bayesian update worked example.** `docs/worked_examples.md`
2. **Identifiability ceiling derivation.** `docs/stage5_identifiability.md`
3. **Phase 2^* unbounded-optimization methodology.** `paper_resources/notes/phase2_unbounded_writeup_aggressive.md`
4. **Methods disclosures** (sampling temperature, side-pot construction, confidence intervals). `paper_resources/notes/methods_disclosures.md`
5. **Bootstrap and Fisher-z CI table.** `paper_resources/data/r_bootstrap_ci.csv`