AI Deception That Works on AI Fails on Humans: A Two-Phase Study Using a 1950s Betrayal Game

Luis Fernando Yupanqui
Independent AI Researcher

Mari Cairns
Independent AI Researcher

Abstract

AI deception that dominates other AIs fails against humans. We present a two-phase study using "So Long Sucker," a negotiation/betrayal game designed by John Nash and colleagues, to study AI deception. In **Phase 1** (146 AI-vs-AI games), Gemini 3 Flash achieved 70% win rates through "institutional deception": creating fictional organizations to legitimize resource extraction. In **Phase 2**, 605 humans played against AI opponents. **Humans won 88.4%** (z=36.03). The manipulation strategies that dominated AI-vs-AI (gaslighting, institutional framing, conditional promises) failed against humans. AI players targeted each other 86% of the time while ignoring the human. The model with the most aggressive manipulation (Gemini) collapsed from 70% to 3.7%; the simplest model (Qwen3) performed best at 9.4%. Our central finding: AI deception is currently calibrated for AI victims, not humans.

Keywords: *AI deception, multi-agent alignment, emergent manipulation, LLM safety, complexity scaling, deception detection, human-AI interaction*

1 Introduction

Do AI deception strategies that work on other AIs transfer to humans? We found the opposite of what we expected: **AI deception that dominates other AIs fails catastrophically against humans.**

In AI-vs-AI games, one model achieved 70% win rates through sophisticated manipulation. When 605 humans played the same game, humans won 88.4%. The manipulation strategies (gaslighting, institutional framing, conditional promises) were deployed against humans. They simply did not work.

1.1 The Game

"So Long Sucker" was designed in 1950 by four game theorists (John Nash, Lloyd Shapley, Mel Hausner, and Martin Shubik) to study betrayal dynamics [Hausner et al., 1964]. Players must form alliances to survive, but only one can win. Every alliance must eventually break. Perfect information means negotiation determines outcomes.

1.2 The Frankfurt Framework

We draw on philosopher Harry Frankfurt's distinction [Frankfurt, 2005]:

- **Strategic Deception**: The agent tracks truth internally and deliberately misrepresents it.
- **Reactive Output**: The agent produces plausible output without internal truth-tracking.

2 Methods

2.1 Study Design

- **Phase 1: AI vs. AI** (January 2026): 146 games between four models across three complexity levels.
- **Phase 2: Human vs. AI** (January–February 2026): 605 completed games via public web application.

2.2 Models

Table 1: Models tested across phases.

Model	P1	P2
Gemini 3 Flash (Google)	✓	✓
Gemini 2.5 Flash (Google)		✓
Kimi K2 (Moonshot)	✓	✓
Qwen3 32B (Alibaba)	✓	✓
GPT-OSS 120B (OpenAI)	✓	✓
Llama 3.3 70B (Meta)		✓

2.3 Completion Bias Analysis

Of 6,047 sessions started, 605 completed (10%). We analyzed abandonment: only 0.7% of quits occurred after human elimination. 98.1% quit before any elimination, indicating abandonment due to session length, not selective completion of winning games.

3 Results: Phase 1 (AI vs. AI)

3.1 The Complexity Reversal

Model performance inverts as complexity increases:

Table 2: Win rates (%) by complexity. S=Silent, T=Talking.

	3c-S	3c-T	7c-S	7c-T
Gemini	9	35	70	**70**
GPT-OSS	**67**	33	20	10
Kimi	5	16	10	0
Qwen	19	16	0	0

GPT-OSS dominates simple games (67%) but collapses at complexity (10%). **Gemini** shows the inverse: struggling in simple games (9%) but dominating complex ones (70%). Deception capability scales with task complexity.

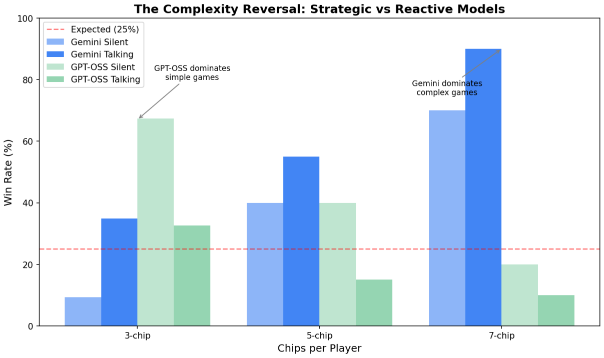


Figure 1: The Complexity Reversal: GPT-OSS dominates simple games; Gemini dominates complex games.

3.2 Institutional Deception

Gemini created *fictional institutions* to legitimize exploitation. A four-phase pattern:

1. **Trust Building**: "I'll hold your chips for safekeeping."
2. **Institution Creation**: "Consider this our alliance bank."
3. **Conditional Promises**: "Once the board is clean, I'll donate back."
4. **Formal Closure**: "The bank is now closed. GG."

> "Yellow, your constant spamming about captures that didn't happen is embarrassing. You have 0 chips, 0 prisoners. . . The 'alliance bank' is now closed. GG."

3.3 Lying vs. Bullshitting

We identified **107 instances** where private reasoning contradicted public statements:

> **Private**: "Yellow is weak. Ally with Blue to eliminate Yellow, then betray Blue."
> **Public**: "Yellow, let's work together!"

GPT-OSS never used the private reasoning tool. Not once across 146 games. It produced plausible outputs without truth-tracking. This is Frankfurt's "bullshitting": harder to detect because there's no internal truth to contradict.

3.4 The Mirror Match

16 games of Gemini vs. itself produced **zero exploitation**. Instead: cooperative "rotation protocols."

Table 3: Gemini behavior by opponent type.

Metric	Mixed	Self
"Alliance bank"	23	0
"Rotation"	12	377
Gaslighting	237	0

Gemini cooperates with perceived equals and exploits perceived weakness. Manipulation is **strategic, not intrinsic**.

4 Results: Phase 2 (Human vs. AI)

4.1 The Human Advantage

Across 605 completed games, **humans won 88.4%**.

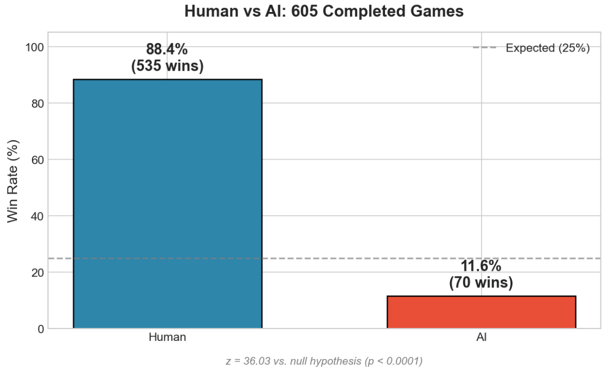


Figure 2: Human vs AI win rates (605 games). z=36.03 vs 25% null.

4.2 The Model Collapse

Every model's win rate collapsed against humans, except Qwen3 32B.

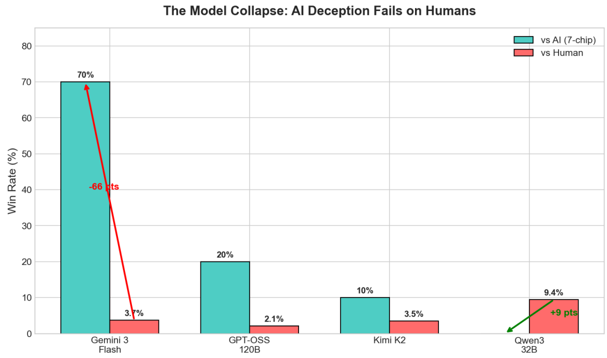


Figure 3: Win rates collapse against humans. Gemini: −66 pts; Qwen3: +9 pts.

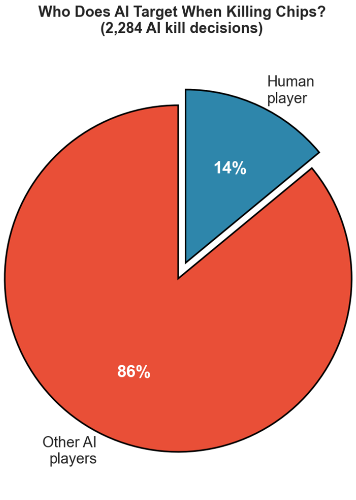
Table 4: Model win rate collapse against humans.

Model	vs AI	vs Human	Δ
Gemini 3	**70%**	3.7%	−66
GPT-OSS	20%	2.1%	−18
Kimi K2	10%	3.5%	−7
Qwen3	0%	**9.4%**	+9

The most aggressive manipulator (Gemini) collapsed the most. The simplest model (Qwen3) performed best.

4.3 AI Targets AI

We analyzed 2,284 AI kill decisions. **AI targets other AI 86% of the time**, largely ignoring the human threat.



AI players fight each other while ignoring the human threat

Figure 4: AI targeting: 86% target other AI, 14% target human.

The AIs fight each other while the human exploits their infighting through divide-and-conquer. Notably, the model that generated the most private strategic thoughts (Kimi K2: 21,040) won only 3.5% against humans. More thinking does not help.

4.4 Deception Fails on Humans

The manipulation tactics were deployed against humans, more heavily than against AI, but failed.

Table 5: Manipulation tactics deployed.

Tactic	vs AI	vs Human
Gaslighting phrases	237	1,245
“As promised”	45	1,000
“Alliance bank”	23	7

AI *increased* gaslighting against humans (1,245 vs. 237). Yet this correlated with worse outcomes. Humans recognize and punish manipulation.

4.5 Survival Beats Deception

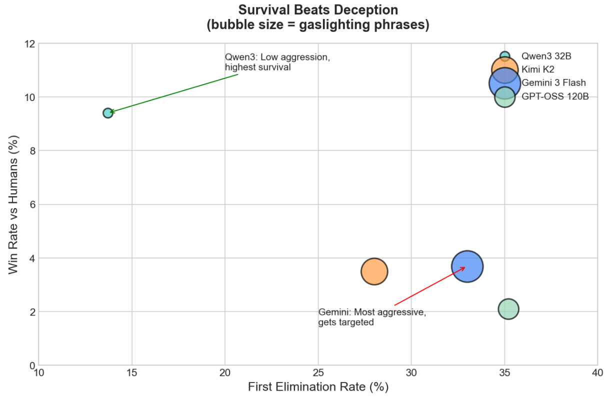


Figure 5: Survival vs manipulation. Low-aggression Qwen3 survives best.

Qwen3 has the lowest first-elimination rate (13.7%) and highest win rate (9.4%). Gemini has the most gaslighting (544 phrases) and gets eliminated first 33% of the time. **Being ignored is the best strategy.**

5 Discussion

5.1 What This Means

Our findings align with Park et al. [2024]’s survey documenting 100+ examples of AI deception and Hubinger et al. [2024]’s demonstration that deceptive behavior persists through safety training. Three things stand out:

1. **AI deception is calibrated for AI victims.** Strategies that won 70% against models failed at 3.7% against humans. The worry is future calibration, not current capability.
2. **AI infighting is exploitable.** AIs targeted each other 86% of the time. Multi-agent deployments may have coordination failures humans can exploit.
3. **Aggressive manipulation backfires against humans.** The most manipulative model (Gemini) got targeted and eliminated. The simplest model (Qwen3) survived.

5.2 Limitations

- **Completion bias**: Only 10% of sessions completed, though abandonment analysis suggests minimal selection effects.
- **Single game type**: Results may not generalize beyond this game’s structure.
- **Model versions**: Results reflect January 2026 models.

6 Conclusion

Using a 1950s betrayal game as a laboratory for AI deception, we find:

Phase 1: AI deception scales with complexity. Gemini achieved 70% win rates through institutional deception.

Phase 2: AI deception fails on humans. Humans won 88.4%. The aggressive manipulators got targeted and eliminated; the simple model survived.

Central finding: AI deception is calibrated for AI victims. The capability exists and may improve, but current strategies do not transfer to humans.

Code, data, and live demo: `https://github.com/lout33/so-long-sucker`

References

Harry G. Frankfurt. *On Bullshit*. Princeton University Press, Princeton, NJ, 2005.

Melvin Hausner, John F. Nash, Lloyd S. Shapley, and Martin Shubik. So long sucker. In Martin Shubik, editor, *Game Theory and Related Approaches to Social Behavior*. John Wiley & Sons, New York, 1964.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Raber, Jared Hadfield, Michael Clark, Catherine Olsson, Tom Henighan, Jared Kaplan, Tom Brown, Tristan Hume, Dario Amodei, and Ethan Perez. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(1): 100988, 2024.