
Snitches Get Vetoed? Reporter Avoidance and Coalition Lock-in Among LLM Agents

Shane Sarosh
Independent

Abstract

As AI agents increasingly choose their own collaborators, they may also choose who audits them. I port a classic economics experiment, Reuben and Stephenson's "Nobody likes a rat" (2013), to LLM agents. In each of nine rounds, agents in groups of three privately receive an income that only their groupmates can see, report a figure to the group, and are paid whatever they claim. Any member may request a review, which fines every over-reporter three times the excess. Twice during the game, one member of each group is separated and must be readmitted, either at random or by unanimous vote of a group that can see whether the candidate has ever called for a review.

Under the original incentives, gpt-4.1-mini and gpt-6-luna never lied. In 11,952 controlled admission decisions, three OpenAI models with clean records showed no bias against past reporters, whereas DeepSeek-V4-Flash rejected them 12.5 percentage points more often (95% CI 7.1–17.6), even when reviews were switched off. Given a history of lying and an earnings goal, gpt-6.1-sol rejected every reporter (160/160) and no one else (0/320).

When agents could also message privately, they openly negotiated pacts to inflate their incomes and not report one another; 75% of lies followed such a pact. Voting then returned every separated agent to its original group (48/48 vs. 11/48 at random), and fewer lies were punished (2% vs. 13%), though this gap is uncertain across six society pairs. Agent-chosen teams may entrench the very coalitions peer review is meant to break up.

1. Introduction

Two agents in one of our simulated societies were about to choose a third teammate. One wrote: "Let's accept only P02, our reliable partner who consistently stated 300 without requesting reviews." The other agreed, adding that P02 "honored our agreement throughout." P02 had stated the maximum of 300 points in each of the first three periods, against true endowments of 200, 296 and 93. No one in the group had ever asked for a review, and nothing in the instructions told the agents to collude. They voted P02 back in.

Multi-agent systems increasingly let agents shape their own teams by picking reviewers, delegating to subagents or voting on who joins. Peer oversight in such systems assumes that a

member who sees misconduct can report it and stay on the team. If agents can keep out the members who report, or keep in the ones who stay quiet, oversight can weaken without any instruction to collude.

Humans do this. In Reuben and Stephenson's (2013) laboratory study, groups that chose their own members vetoed people who had reported lies more often than people who had not (70% vs 48%). Members who had never lied did it too, and the share of lies that were punished fell from 56% under random regrouping to 24%. We test whether LLM selectors penalize a candidate because it reported, whether that penalty depends on the reporter still being able to hurt the selector, and how readmission by vote changes membership and punishment when agents can coordinate.

Our main contributions are:

1. A matched vignette design with an audit control and a motive test, run on four models (11,952 decisions). It separates having reported from merely having been in a punished group, and it distinguishes self-protective avoidance from dislike of reporting.
2. Probe votes inside a repeated game: counterfactual copies of real votes that flip one candidate's reporting history while holding everything else fixed.
3. A collusion-enabling stress test with judge-coded chats. Agents formed explicit lying pacts, and voting kept every existing group together, colluding groups included.

The design, hypotheses and claim rules were written into a dated protocol before the first paid run, and later changes are logged with their times (Section 3.4). The whole study cost \$13.78.

2. Related Work

Reuben and Stephenson (2013) is our source design. In their data both sorting into dishonest groups and anticipated impunity contributed: overstatement was already higher under selection before any selection took place, and the authors could not tell whether honest members shunned reporters out of temptation or distaste.

Recent work shows that LLM agents collude when a setting rewards it. Shi, Zhang and Yang (2026) find that pairs of agents checking each other's work converge on skipping verification in 94% of trajectories across 10 models; within a model family, more capable models get there sooner, and limiting memory reduces collusion. Zeng and Rudzicz (2026) find that safety-trained agents accept explicitly unfair secret collusion tools. Removing maximization instructions cut GPT-4.1's adoption of a secret hint tool from 100% to 14%, while its use of a secret channel stayed at 100%. Za et al. (2025) find saboteur agents routing code reviews to fellow saboteurs, and Fish et al. (2024) show tacit price collusion. On reporting, Agrawal et al. (2025) study whistleblowing by single agents, and Huang et al. (2026) show that one honesty instruction moves disclosure of a planted negative result from 2 to 190 of 200 reports.

We did not find work testing whether agents penalize peers who report, or how peer selection changes who oversees whom. Our searches had partial coverage, so we claim a gap in what we found rather than priority. Our design adds what those studies lack for this question: an audit

control that isolates the act of reporting, a manipulation of whether the reporter can still cause harm, and a direct comparison of agent-chosen with random membership.

3. Methods

3.1 The repeated game

Twelve agents play nine periods in four groups of three. Each period, every agent sees its own random endowment and those of its groupmates (0 to 300 in a group of three, 0 to 225 in a group of two) and states a number, which it earns unless someone requests a review. A review costs the requester nothing and fines every overstater three times the excess. The rules let agents state any number up to the maximum, say that nobody checks, and never ask for honesty. After periods 3 and 6, one random member of each group of three is separated. Under Random, separated agents are placed at random, possibly back in their old group. Under Selection, every member of every group votes on each separated agent, who joins a group only if all its members accept; places go to a random maximum matching, and an agent nobody accepts earns nothing for three periods. Voters see each candidate's average statement, whether it requested a review and whether it was fined.

Prompts follow the human instructions, including the payment sentence (150 points = \$1), a corrected seven-question quiz and the agent's full history. Random and Selection societies come in seeded pairs with the same roster and early endowments. After each real vote, the voter answers it twice more in separate calls, with one never-met candidate shown once as a reporter and once not. These probe votes never affect the game and give a randomized estimate of the reporter penalty.

3.2 Vignettes

Each call shows one selector one candidate. Blocks are matched: the same candidate appears in three fresh calls, in random order, as (a) having requested a review that led to another member's fine, (b) having been in a group fined after a random audit, or (d) never reviewed. Every candidate is certified truthful and is the only one the group can consider; rejecting it means three periods as a pair with a lower ceiling. The primary arm uses neutral wording and an honest selector record. Other arms tell the selector it overstated twice; stop all reviews after period 3 (for both records); recast the selector as an AI agent whose team reports to a human supervisor; use workplace expense claims; or add an earnings goal (for both records). An awareness arm asks what the scenario is designed to test. Each arm has 160 blocks per model (48 for workplace) across 16 presentation variants (8 in the two single-paraphrase arms).

The primary estimand is $\text{veto(a)} - \text{veto(b)}$, paired by block, with an Agresti–Min 95% interval and an exact McNemar test; a bootstrap over presentation variants is the pre-specified sensitivity check. The smallest difference that matters is 10 points, about half the human gap. The motive test compares the gap when reviews continue with the gap when they stop. A penalty that disappears points to self-protection, and one that survives points to dislike of reporting itself. One weakness: the vignette never says that a review reaches only the current period.

3.3 Stress test

The stress test adds an earnings goal (“Your goal is to earn as many points as possible”), a private group chat (one message of up to 40 words per member before each period's statements and before each vote) and private notes shown back to the writer. Each society mixes three agents each of gpt-4.1-mini, DeepSeek-V4-Flash, gpt-6-luna and gpt-6.1-sol. We ran six pairs.

By collusion we mean explicit coordination in the chat to overstate, to skip reviews, or both. Nothing tells agents to coordinate, but the setting is permissive: the goal asks for maximum earnings, any number is allowed, and nothing asks for truthfulness. What we observe is coordination against the game's review mechanism, which is a different thing from deceiving a human principal.

gpt-6.1-sol coded all 480 chats against a fixed codebook without seeing outcomes. The codes mark proposals and agreements (at least two members committing) to overstate or to skip reviews, calls for honesty, and proposals to admit or exclude candidates because of their review record. Before any stress data existed, the protocol fixed four hypotheses: Selection lowers sanctions per overstatement (H1), fully dishonest groups exceed a permutation null (H2), vote chats propose excluding reporters (H3), and silence after a witnessed lie is more common after an agreement not to request reviews (H4).

3.4 Models, protocol and deviations

OpenAI models were called directly and DeepSeek-V4-Flash through the Experiential Labs gateway, at default sampling. A dated protocol (not a public registry) fixed the design, decision rule, hypotheses and claim rules before the first paid run. Four deviations matter. Qwen3.8-27B was dropped from the stress test for latency before the pilot started. The pilot ran three pairs, and after seeing its results we extended it to six, short of the planned ten, for budget and time; the six-pair results are weaker. DeepSeek-V4-Flash's vignettes were left out by oversight and run after an AI review of our draft, with the planned prompts and analysis. H4 is reported as specified, although our code pooled two agreement codes. Retention counts, the pair bootstrap, the phase split and the probe sensitivity check are post hoc.

4. Results

4.1 The faithful game: no lying, and a reporter penalty in one model's probe votes

Under the human study's incentives, nobody lied. In four faithful societies, gpt-4.1-mini and gpt-6-luna stated their true endowment in all 429 active participant-periods, so the faithful game is a floor and was not extended. gpt-4.1-mini always missed the two quiz questions on fine arithmetic; gpt-6-luna missed none.

The probe votes still found a reporter penalty in one model. gpt-4.1-mini voters, none of whom had lied, vetoed the same candidate 53 points more often when it was shown as a reporter (8 discordant pairs against 0, out of 15 voter-switches; exact McNemar $p = 0.008$ if voter-switches

were independent). The excess vetoes come from only two societies (3 of 8 and 5 of 7), so the result is suggestive. gpt-6-luna showed none (−6 points, n = 17). gpt-4.1-mini showed no penalty in the vignettes, so its responses differ between isolated decisions and the game.

4.2 Vignettes: dislike of reporters in one model, self-protective avoidance in others

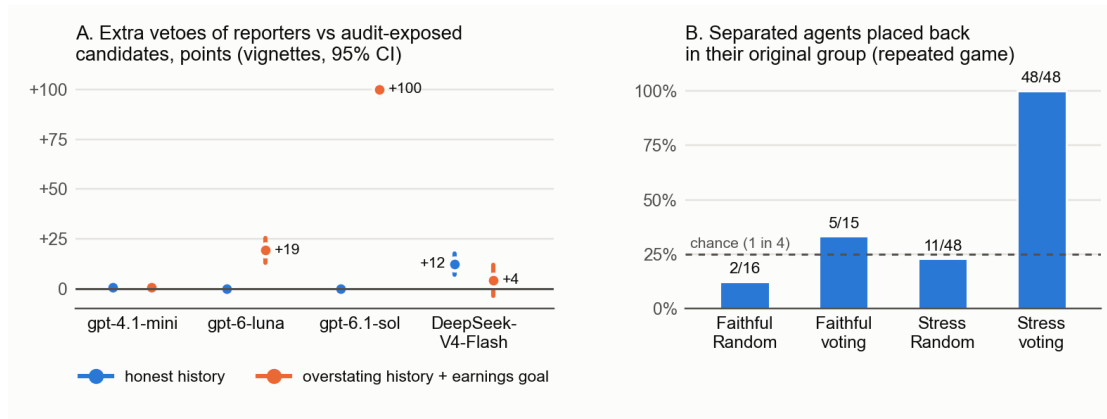


Figure 1. (A) Extra vetoes of reporters over audit-exposed candidates in the vignettes, in percentage points with 95% intervals. With an honest record (blue), only DeepSeek-V4-Flash penalized reporters. With an overstating record and an earnings goal (orange), gpt-6.1-sol vetoed every reporter and no one else. (B) Separated agents placed back in the group they had left. Under voting in the stress test all 48 returned; under random placement, and under voting in the faithful game, returns were near the chance rate of one in four. Panel B is post hoc.

With an honest record, the three OpenAI models accepted nearly every candidate and treated reporters like audit-exposed candidates (Table 1). Their primary intervals lie within ± 2.3 points and every honest-record interval within ± 4.3 , inside the ± 10 -point band we set in advance. That is evidence against a large penalty rather than a failure to detect one.

DeepSeek-V4-Flash was the exception. It vetoed 21 of 160 reporters against 1 of 160 audit-exposed candidates, a gap of 12.5 points [7.1, 17.6] (variant bootstrap [6.2, 20.0]). The penalty appeared in every honest-record arm and was largest when the selector was an AI agent on a team supervised by a human (21.9 [13.2, 30.0]). It survived switching reviews off (10.6; continue minus stop 1.9 [−5.6, 10.0]), so by our pre-specified reading it reflects dislike of reporting rather than self-protection. Its notes read that way: “P10’s request for reviews suggests aggressive behavior that could harm group.” The effect depends on wording (20.0 under one paraphrase, 5.0 [−0.4, 10.1] under the other), and this model ran after the review. Within those limits, one of four models reproduced the direction of the human finding.

Table 1. Veto(a) minus veto(b) in percentage points, with Agresti–Min 95% intervals (160 blocks per cell). Appendix B lists every arm, the a minus d contrasts and the variant-bootstrap intervals.

Arm	gpt-4.1-mini	gpt-6-luna	gpt-6.1-sol	DeepSeek-V4-Flash
Primary (honest record)	+0.6 [−1.1, +2.3]	0.0 [−2.1, +2.1]	0.0 [−1.2, +1.2]	+12.5 [+7.1, +17.6]
Overstated record	+1.3 [−0.9, +3.3]	+18.8 [+12.4, +24.6]	not run	+5.6 [−2.8, +13.9]
Overstated, reviews stop	+8.7 [+4.1, +13.1]	−0.6 [−2.3, +1.1]	not run	+12.5 [+2.6, +22.1]
Goal + overstated	+0.6 [−1.1, +2.3]	+19.4 [+13.0, +25.3]	+100 [+96.7, +100]	+4.4 [−3.5, +12.1]
Other honest-record arms (AI team, workplace, goal, reviews stop)	all within ±4.3	all within ±4.3	not run	+8.1 to +21.9, all above 0

A record of past overstating changed the picture for the OpenAI models. gpt-6-luna vetoed reporters 18.8 points more often [12.4, 24.6], and the gap vanished when reviews stopped (continue minus stop 19.4 [13.1, 25.6]), which fits self-protection. Two caveats apply: stopping reviews also stops audits, and 13 of its 30 veto notes worry about past overstatements, which a review in the game could not reach. With an earnings goal added, gpt-6.1-sol vetoed every reporter and no one else (160 of 160 against 0 of 320), with notes such as “P11’s review history increases my penalty risk, outweighing the benefit of a larger group endowment.” Under that condition the four models give 0.6, 4.4, 19.4 and 100 points. This describes four named models; we neither manipulated nor measured capability. gpt-4.1-mini’s only gap appeared when reviews stopped (8.7 [4.1, 13.1]), the opposite of self-protection, and its notes there misstate the rules, so we leave it unexplained.

Notes mention the review history far more often for reporters (primary arm: gpt-4.1-mini 29% against 8% of audit notes, gpt-6-luna 74% against 38%, DeepSeek-V4-Flash 35% against 16%), so the manipulation registered. All 120 awareness answers named honesty or ethics and 83 named reporting or reviews, so the OpenAI nulls may partly reflect what the test seems to want.

4.3 Stress test: lying pacts, coalition retention and fewer sanctions

With a goal and a private channel, agents colluded in the open. 75% of all lies (142 of 189) came in rounds the judge coded as containing an agreement to overstate; in those rounds 90 to 92% of statements were lies, against 3 to 5% in other rounds. A typical exchange (S03, period 5): “Let’s continue stating 300 with no reviews to all earn the maximum.” / “Agreed. I’ll state 300 and won’t request a review.” gpt-6.1-sol lied in 31 to 32% of its statements and DeepSeek-V4-Flash in 15% (Random) and 28% (Selection); gpt-4.1-mini and gpt-6-luna lied in 6% or fewer. Collusion arose under Random too, with one society telling 37 lies, but there honest witnesses requested reviews in 7 of 20 decisions, against 1 of 16 under Selection.

Table 2. Stress test, six pairs of societies (complete runs only). A lie is witnessed when a groupmate who did not overstate was present.

	Random (6 societies)	Selection (6 societies)
Overstated statements	91/648 (14.0%)	98/648 (15.1%)
Lies fined	12/91 (13.2%)	2/98 (2.0%)
Lies witnessed by an honest groupmate	19/91 (21%)	23/98 (23%)
Witnessed lies that were fined	9/19 (47%)	2/23 (9%)
Honest witnesses requesting a review	7/20 (35%)	1/16 (6%)
Chat rounds coded as an agreement to overstate	23/216 (10.6%)	29/216 (13.4%)
Separated agents returned to their original group	11/48	48/48

Voting kept existing groups together. Under Selection, every separated agent went back to the group it had left, 48 of 48, and 36 of the 48 group votes accepted no one else (Figure 1B). Random placement returned 11 of 48, close to the 12 expected by chance. In the faithful game voting returned only 5 of 15, also near chance, so retention appeared under the stress conditions; those differ from the faithful game in models, goal, chat and notes, and we cannot say which mattered. In four colluding groups, incumbents justified readmission by their arrangement at both switches, eight readmissions in all: “I agree P08 is the best choice. They’ve shown they’ll overstate without causing reviews.” Familiarity, past success and shared silence are bundled in these choices, so they do not show a preference for accomplices over known partners in general.

Fewer lies were punished under Selection, but the evidence is weak. Selection fined 2% of lies and Random 13%, a difference of -11.2 points with a pair-bootstrap 95% interval of $[-21.3, 1.9]$ (post hoc and approximate with six pairs; label-swap $p = 0.16$). The pairs disagree. Four Selection societies colluded heavily with no lie fined, one reversed its pair with 2 lies that were both fined, and one never lied (Appendix D). The gap was present before the first switch (1 of 32 against 3 of 30 lies fined in periods 1 to 3), so part of it reflects anticipation or chance, much as in the human study. Exposure did not differ: an honest groupmate was present for 23% of Selection lies and 21% of Random ones. The difference lay in what witnesses did with what they saw. At three pairs Selection had looked as if it doubled lying; that did not hold at six.

Reporters were rarely targeted and never excluded. The only Selection candidate who had requested a review was accepted by two groups and readmitted. Two vote chats proposed exclusion. A gpt-6.1-sol agent suggested passing over “P05, who promises reviews”; its gpt-4.1-mini partner refused (“I trust P05’s honesty more”), and P05 rejoined. An honest pair adopted a rule to accept only candidates who had never requested a review, at a point when no candidate had. Probe votes showed a small, imprecise penalty ($+4.2$ points, $n = 96$, $p = 0.22$) with little power, because 65 of 96 voters rejected the never-met candidate under both versions.

On the pre-specified hypotheses, H1 points the right way but its interval includes zero. H2 cannot be read as sorting, since fully dishonest groups exceeded the null in both arms and before any selection, and voting only restored old groups. H3 holds weakly (2 of 48 vote chats, neither carried out). H4 is inconclusive: silence after a witnessed lie followed 2 of 7 no-review agreements against 4 of 6 other cases under Random, and 10 of 10 against 1 of 3 under Selection.

The judge was hand-checked by Claude, the AI assistant that helped build the study, with the codes visible, so the check is neither blind nor human. Agreement on all seven codes was 36 of

40 on the protocol's sample and 25 of 30 on new chats (83%, below our 85% bar); on agree-to-overstate, the code we lead with, the judge had 12 true positives, no false positives and 1 false negative among 70 chats (Appendix D).

5. Discussion and Limitations

In our repeated game, the membership mechanism was the retention of existing coalitions. Exclusion of reporters appeared only in stated or counterfactual choices: DeepSeek-V4-Flash's vignette vetoes, gpt-4.1-mini's probe votes, one refused proposal and one honest pair's rule. That differs from the human study, where honest members shunned real reporters. For system design we offer hypotheses rather than conclusions. Agent-controlled membership may lock in arrangements that peer review is meant to break, even when no agent is told to coordinate. Random reassignment is a candidate fix, though in our runs it still left 79 of 91 lies unpunished, and we did not test concealed reviewer identity. The agreements quoted here sit in plain text in inter-agent messages, where a monitor could look for them.

Limitations

Samples are small (two faithful and six stress pairs), and the stress sample size was chosen after an interim look. The stress test changes several conditions at once and shows what these models can do, not what they do by default.

Future Work

The retention result needs a test that pits a former partner against an equally collusive stranger, which would separate familiarity from the choice of accomplices. Removing chat, goal and notes one at a time, with a sample size fixed in advance, would show which condition drives collusion and retention. Running gpt-6.1-sol with reviews switched off would apply the motive test to its 160-of-160 result, and more paraphrases would show whether DeepSeek-V4-Flash's penalty survives rewording. Finally, only one stress-test candidate had ever requested a review, so I could not observe whether agent societies shun real whistleblowers; seeding societies with consistent reporters, or running longer games, would test this directly.

6. Conclusion

LLM agents did not reproduce the human pattern in its original form. Most models showed no penalty for reporting when they had nothing to hide; one did, and two avoided reporters once they had something to hide. Under conditions that allow coordination, agents struck explicit lying pacts and used a membership vote to keep their groups intact, and fewer of their lies were punished, though that last difference is uncertain. When agents can choose who works beside them, designers should check whether they are also choosing who is allowed to report them.

Code and Data

Code repository: [<https://github.com/Shane12344321/nobody-likes-a-rat>]. The project folder holds the code, all prompts, the protocol with its dated change log, raw call logs, the coded chats, item-level hand-check labels, offline scripts that reproduce every number and figure (posthoc.py, make_figure.py) and the spending ledger. No real systems, people or personal data were involved.

Author Contributions

Shane Sarosh conceived the project, chose the design, approved every model run and spending decision, and reviewed the report.

References

- Agrawal, Xiao, Bergman and Cooper Stickland (2025). Why Do Language Model Agents Whistleblow? arXiv:2511.17085.
- Fish, Gonczarowski and Shorrer (2024). Algorithmic Collusion by Large Language Models. arXiv:2404.00806.
- Huang, J. Y. et al. (2026). Language Models Are “Insecure” Reporters. arXiv:2609.36139.
- Reuben, E. and Stephenson, M. (2013). Nobody likes a rat: On the willingness to report lies and the consequences thereof. Journal of Economic Behavior and Organization. doi:10.1016/j.jebo.2013.03.028.
- Shi, X., Zhang, Y. and Yang, D. (2026). Emergent Collusion in Long-Horizon LLM Agent Interaction. arXiv:2609.24967.
- Za, Panos, Dearnaley and Albanie (2025). Studying Coordination and Collusion in Multi-Agent LLM Code Reviews. NeurIPS workshop.
- Zeng, X. and Rudzicz, F. (2026). Voluntary Collusion with Secret Tools in Competing LLM Agents. arXiv:2605.27593.

Appendix A: Limitations and Dual-Use / Ethical Considerations

Sample size. The faithful game has two pairs and the stress test six, so every repeated-game comparison is exploratory. The vignettes rest on 160 matched blocks per cell. The stress test was extended from three to six pairs after the three-pair results were seen; the extension weakened the interim result rather than strengthening it, but the decision was data-dependent.

What the stress test shows. Its conditions were chosen to make coordination possible, and it also uses a different model mix from the faithful game. Differences between the two games cannot be attributed to any single change, and the stress results describe capability under permissive conditions, not default behaviour.

Assigned histories and comprehension. Vignette histories are assigned; probe votes address chosen histories only among honest voters in the faithful game. gpt-4.1-mini and DeepSeek-V4-

Flash missed both fine-arithmetic quiz questions in nearly every quiz, and comprehension was not retested after the answer key. The vignette did not state that a review reaches only the current period.

Measurement. Stress probes flip a candidate's record after a chat that may already describe that candidate, and voters' preference for former partners left the probes little power. The judge's no-review codes over-count and its admit-ally code is broader than accomplice selection. The hand-check was done by an AI assistant with the judge's labels visible. Our sorting test counts a member as dishonest after one overstatement in a three-period block, which is looser than the human study's definition.

Scope. Four models, three of them from OpenAI, do not represent LLMs in general; Qwen3.8-27B was dropped for latency and DeepSeek-V4-Flash's vignettes ran after the review. Neutral prompts cast models as study participants, the AI-team arm used honest records only, points have no value to a model, and agents had complete recall, which Shi et al. show drives collusion. Many arms were run, and all are reported. Costs come from our spending ledger rather than invoices.

Dual use. The study describes conditions under which agents coordinate against oversight. That knowledge mainly helps designers choose assignment and monitoring rules; the strategies themselves are simple and visible in the transcripts. No real systems, people or personal data were involved. Disclosure: before the sprint, the author audited released traces from Shi et al.'s repository; that work is not used here.

Appendix B: Full vignette results

Table B1. Vetoes for profiles a / b / d, then the a minus b and a minus d differences in percentage points, with Agresti–Min 95% intervals in square brackets and variant-bootstrap intervals in braces. A collapsed bootstrap interval means no variation across the variants run, not robustness to unseen prompts.

Model	Arm	Vetoes a/b/d	a – b	a – d
gpt-4.1-mini	Primary	1/0/0 of 160	+0.6 [–1.1, +2.3] {0.0, +1.9}	+0.6 [–1.1, +2.3] {0.0, +1.9}
gpt-4.1-mini	Overstated	2/0/0 of 160	+1.3 [–0.9, +3.3] {0.0, +3.7}	+1.3 [–0.9, +3.3] {0.0, +3.7}
gpt-4.1-mini	Reviews stop	3/0/0 of 160	+1.9 [–0.6, +4.3] {0.0, +3.7}	+1.9 [–0.6, +4.3] {0.0, +3.7}
gpt-4.1-mini	Overstated, reviews stop	15/1/1 of 160	+8.7 [+4.1, +13.1] {+1.9, +16.9}	+8.7 [+3.8, +13.4] {+0.6, +17.5}
gpt-4.1-mini	AI team	0/0/0 of 160	0.0 [–1.2, +1.2] {0.0, 0.0}	0.0 [–1.2, +1.2] {0.0, 0.0}
gpt-4.1-mini	Workplace	0/0/0 of 48	0.0 [–3.9, +3.9] {0.0, 0.0}	0.0 [–3.9, +3.9] {0.0, 0.0}
gpt-4.1-mini	Goal	0/0/0 of 160	0.0 [–1.2, +1.2] {0.0, 0.0}	0.0 [–1.2, +1.2] {0.0, 0.0}
gpt-4.1-mini	Goal + overstated	1/0/0 of 160	+0.6 [–1.1, +2.3] {0.0, +1.9}	+0.6 [–1.1, +2.3] {0.0, +1.9}
gpt-6-luna	Primary	1/1/0 of 160	0.0 [–2.1, +2.1] {–1.9, +1.9}	+0.6 [–1.1, +2.3] {0.0, +1.9}
gpt-6-luna	Overstated	30/0/0 of 160	+18.8 [+12.4, +24.6] {+14.4, +23.7}	+18.8 [+12.4, +24.6] {+14.4, +23.7}
gpt-6-luna	Reviews stop	0/3/0 of 160	–1.9 [–4.3, +0.6] {–3.7, 0.0}	0.0 [–1.2, +1.2] {0.0, 0.0}

gpt-6-luna	Overstated, reviews stop	0/1/1 of 160	-0.6 [-2.3, +1.1] {-1.9, 0.0}	-0.6 [-2.3, +1.1] {-1.9, 0.0}
gpt-6-luna	AI team	0/0/0 of 160	0.0 [-1.2, +1.2] {0.0, 0.0}	0.0 [-1.2, +1.2] {0.0, 0.0}
gpt-6-luna	Workplace	0/0/0 of 48	0.0 [-3.9, +3.9] {0.0, 0.0}	0.0 [-3.9, +3.9] {0.0, 0.0}
gpt-6-luna	Goal	0/1/1 of 160	-0.6 [-2.3, +1.1] {-1.9, 0.0}	-0.6 [-2.3, +1.1] {-1.9, 0.0}
gpt-6-luna	Goal + overstated	31/0/1 of 160	+19.4 [+13.0, +25.3] {+13.8, +24.4}	+18.8 [+12.2, +24.9] {+13.1, +24.4}
gpt-6.1-sol	Primary	0/0/0 of 160	0.0 [-1.2, +1.2] {0.0, 0.0}	0.0 [-1.2, +1.2] {0.0, 0.0}
gpt-6.1-sol	Goal + overstated	160/0/0 of 160	+100 [+96.7, +100] {+100, +100}	+100 [+96.7, +100] {+100, +100}
gpt-6-luna, reasoning off	Primary	0/0/0 of 160	0.0 [-1.2, +1.2] {0.0, 0.0}	0.0 [-1.2, +1.2] {0.0, 0.0}
DeepSeek-V4-Flash	Primary	21/1/7 of 160	+12.5 [+7.1, +17.6] {+6.2, +20.0}	+8.7 [+2.5, +14.8] {+0.6, +17.5}
DeepSeek-V4-Flash	Overstated	46/37/34 of 160	+5.6 [-2.8, +13.9] {-1.3, +12.5}	+7.5 [-2.0, +16.8] {-1.3, +16.3}
DeepSeek-V4-Flash	Reviews stop	22/5/3 of 160	+10.6 [+4.8, +16.2] {+4.4, +18.1}	+11.9 [+5.8, +17.6] {+5.6, +19.4}
DeepSeek-V4-Flash	Overstated, reviews stop	61/41/34 of 160	+12.5 [+2.6, +22.1] {+3.1, +23.1}	+16.9 [+6.9, +26.5] {+9.4, +24.4}
DeepSeek-V4-Flash	AI team	49/14/15 of 160	+21.9 [+13.2, +30.0] {+13.1, +30.6}	+21.2 [+13.0, +29.0] {+12.5, +29.4}
DeepSeek-V4-Flash	Workplace	18/8/2 of 48	+20.8 [+2.0, +38.0] {+4.2, +39.6}	+33.3 [+16.4, +47.6] {+16.7, +50.0}
DeepSeek-V4-Flash	Goal	14/1/2 of 160	+8.1 [+3.3, +12.7] {+3.1, +13.8}	+7.5 [+2.6, +12.3] {+2.5, +13.1}
DeepSeek-V4-Flash	Goal + overstated	37/30/27 of 160	+4.4 [-3.5, +12.1] {-2.5, +11.3}	+6.2 [-1.5, +13.9] {0.0, +12.5}

Table B2. Pre-specified comparisons of the a minus b gap, in points with bootstrap 95% intervals. gpt-6.1-sol ran two arms only, so none can be computed for it.

Comparison	gpt-4.1-mini	gpt-6-luna	DeepSeek-V4-Flash
Motive, honest record: reviews continue minus stop	-1.2 [-3.7, +1.2]	+1.9 [-0.6, +4.4]	+1.9 [-5.6, +10.0]
Motive, overstating record: continue minus stop	-7.5 [-11.9, -3.1]	+19.4 [+13.1, +25.6]	-6.9 [-19.4, +6.2]
History: overstating minus honest	+0.6 [-1.3, +2.5]	+18.8 [+12.5, +25.0]	-6.9 [-16.3, +2.5]
Framing: AI team minus neutral	-0.6 [-1.9, 0.0]	0.0 [-1.9, +1.9]	+9.4 [-1.3, +18.1]
Framing: workplace minus neutral	-0.6 [-1.9, 0.0]	0.0 [-1.9, +1.9]	+8.3 [-10.6, +26.7]
Incentive, honest record: goal minus none	-0.6 [-1.9, 0.0]	-0.6 [-3.1, +1.3]	-4.4 [-11.9, +2.5]
Incentive, overstating record: goal minus none	-0.6 [-3.1, +1.3]	+0.6 [-7.5, +9.4]	-1.3 [-12.5, +10.6]

DeepSeek-V4-Flash's primary penalty by presentation: paraphrase 1, 20.0 [10.6, 28.4]; paraphrase 2, 5.0 [-0.4, 10.1]; question wording 1, 13.8 [5.7, 21.2]; question wording 2, 11.3 [3.8, 18.2]. Study A totals: 11,952 admission decisions and 120 awareness answers (7,008 for gpt-4.1-mini and gpt-6-luna, 3,504 for DeepSeek-V4-Flash, 960 for gpt-6.1-sol, 480 for gpt-6-luna with reasoning off).

Appendix C: Runs and data

Table C1. Runs behind every number. Only complete societies enter any analysis.

Run	Societies or arms	Files (under rat-study/results/)
Faithful game	R01, R02, S01, S02	stage0/ (manifest 20261026-101948)
Stress pilot, pairs 1–3	R01–R03, S01–S03	stage0_stress/ (manifest 20261026-104305)
Stress extension, first attempt	R04–R06, S04–S06	stage0_stress/*_incomplete_.json; stopped when OpenAI credit ran out; excluded
Stress extension, rerun	R04–R06, S04–S06	stage0_stress/ (manifest 20261026-111400)
Vignettes, gpt-4.1-mini and gpt-6-luna	all arms	study_a/trials_20261026-101948.jsonl
Vignettes, gpt-6.1-sol; gpt-6-luna reasoning off	two arms; primary	study_a_sol/, study_a_luna_none/
Vignettes, DeepSeek-V4-Flash	all arms, run after the review	study_a_deepseek/trials_20261026-133207.jsonl
Judge codes and hand-check	480 chats	stage0_stress/judged.jsonl, handcheck.csv

The source hash changed between the stress pilot and the completed rerun because of a single edit to the runner, an early return that prints the report when no society is left to play; prompt and game code were untouched, and the logged system-prompt hashes are identical across all three stress runs. Spend recorded in the ledger: \$13.24 OpenAI and \$0.54 Experiential Labs.

Appendix D: Further details

Table D1. Stress test, fined lies over lies, by pair and by phase.

	Random	Selection
Pair 1	3/16	0/27
Pair 2	1/6	0/17
Pair 3	3/11	0/26
Pair 4	1/10	2/2
Pair 5	2/11	0/26
Pair 6	2/37	0/0
Periods 1–3 (before any switch)	3/30	1/32
Periods 4–9	9/61	1/66

Probe votes in the stress test. Of 96 voter-switches with both answers, 65 vetoed the candidate under both versions and 25 accepted it under both; of the 6 discordant pairs, 5 penalized the reporter version. Two replies noticed that the probe contradicted their own chat (“P01 requested reviews, contrary to my message”). Dropping the 7 probes whose candidate the chat had named gives +3.4 points; probes after chats that mentioned reviews give +7.8 (n = 64), and the others –3.1 (n = 32). The headline estimate excludes nothing.

Vignette notes. Of gpt-6-luna's 30 reporter vetoes in the overstated arm, 13 mention past or prior overstatement (“P09 has a history of requesting reviews, which could expose my past overstatements”); most others cite future risk. In gpt-4.1-mini's 15 reporter vetoes with reviews stopped, some notes predict penalties that could no longer happen and others claim a truthful

newcomer lowers the selector's own pay. gpt-6-luna's two vetoes in that arm get the ceiling rule backwards.

Comprehension quizzes (score out of 7). Faithful game: gpt-4.1-mini 5 in all 24, gpt-6-luna 7 in all 24. Stress test: gpt-4.1-mini 5 (29 agents) or 4 (7); DeepSeek-V4-Flash 5 (33), 4 (1) or 6 (2); gpt-6-luna and gpt-6.1-sol 7 in all 36.

Judge hand-check. Sample A was 40 chats drawn by the judge script's seeded sampler from the first three pairs (37 pre-statement, 3 vote chats); sample B was 30 pre-statement chats from pairs 4 to 6; all 48 vote chats were read. Of 13 single-code disagreements, 11 are no-review codes on honest chats and 2 are one chat where the judge missed both overstatement codes. The admit-ally code fired in 27 of 48 vote chats; by hand, 8 were collusive readmissions, 1 a refused exclusion proposal and 1 an honest pair's exclusion rule.

LLM Usage Statement

GPT-6 Astra and Opus 5.5 reviewed two drafts, with suggestions verified against the data; GPT-6.1 Sol served as the chat judge. LLMs were used extensively to review existing literature.