LLMs Can Play (Global) Games

Khaled Eltokhy
Department of Economics
The Graduate Center, CUNY

February 2026

Abstract

The canonical Morris–Shin (2003) regime change game—continuous private signals, large groups, strategic uncertainty—has not been implemented experimentally. I embed nine large language models spanning six architecture families in the full game, conveying each agent's private signal as a natural-language intelligence briefing. Across nine models and 1,600 country–periods in the pure treatment (40,000 individual decisions), join rates track the Bayesian Nash equilibrium prediction (mean $r = +0.73$, $p < 0.001$ for every model). Scrambling briefings across periods collapses the correlation to $r = +0.23$; inverting signals flips it to $r = -0.67$—confirming that behavior is driven by briefing content. I then implement censorship, surveillance, and propaganda treatments drawn from the information design and authoritarian control literatures. Upper censorship raises join rates by 18.5 pp through pooling; surveillance reduces them by 17.5 pp through preference falsification; combining the two nearly eliminates coordination (30.9% → 3.7%).

1 Introduction

Coordination games with multiple equilibria are central to the analysis of bank runs (Diamond and Dybvig, 1983), currency attacks (Obstfeld, 1996), and political upheaval (Angeletos et al., 2007). The theory of global games (Carlsson and van Damme, 1993; Morris and Shin, 2003; Frankel et al., 2003) resolves the multiplicity by introducing private information: when agents observe noisy private signals about an underlying fundamental, a unique equilibrium emerges in threshold strategies. The canonical application—regime change—has been extensively studied theoretically. Laboratory experiments have tested the theory in simplified settings: small groups with numeric signals and stylized payoffs (Heinemann et al., 2004, 2009; Szkup and Trevino, 2020). But the full Morris–Shin regime change game—continuous private signals, large groups, strategic uncertainty—has not been implemented experimentally. Field data from actual crises confounds strategic behavior with institutional and informational heterogeneity.

I take a different approach: I embed large language model (LLM) agents directly in the Morris and Shin (2003) regime change game. Each agent receives a private signal $x_i = \theta + \varepsilon_i$, translated into a natural-language intelligence briefing describing the political, economic, and security situation. No explicit payoff table is provided—the stakes of joining or staying are embedded in the narrative, forcing agents to extract strategic information from language rather than from a formatted matrix. I run this experiment across nine architecturally distinct models spanning six families (Mistral, Llama, Qwen, OLMo, GPT, and MiniMax), with 25 agents per country–period and pure-treatment sample sizes of 100–600 country–periods per model (Table 1), totaling 1,600 country–periods (40,000 individual decisions) in the pure treatment alone.

The main finding is that LLM agents exhibit behavioral alignment with the Bayesian Nash equilibrium prediction. The correlation between the theoretical attack mass $A(\theta) = \Phi[(x^* - \theta)/\sigma]$ and the empirical join fraction averages $r = +0.73$ ($p < 0.001$ for every model). Two falsification tests confirm that this correlation is driven by briefing content: randomly scrambling briefings across periods reduces it to $r = +0.23$, and inverting the signal direction flips it to $r = -0.67$. In both cases the change relative to the pure treatment is significant (Fisher z-test, $p < 0.001$). Pre-play communication raises join rates by 3.7 percentage points, concentrated in weak-regime environments.

Taking this behavioral foundation as given, I then ask: *what can an information designer accomplish?* I implement information designs—stability, instability, public signals, and censorship—and find large effects. Public signal injection reduces join rates by 10.7 pp; upper censorship raises them by 18.5 pp through pooling, consistent with Kolotilin et al. (2022). I then test authoritarian information control: surveillance reduces join rates by 17.5 pp ($p < 0.001$) through Kuran (1991) preference falsification, propaganda suppresses coordination in a dose-response pattern, and combining surveillance with censorship nearly eliminates coordination (30.9% → 3.7%).

The paper makes three contributions. First, it provides the first implementation of the full Morris–Shin regime change game—with continuous private signals, large groups, and narrative information—going beyond the simplified coordination games tested in existing laboratory experiments. Second, it provides the first experimental tests of information design and authoritarian control predictions from Goldstein and Huang (2016), Kolotilin et al. (2022), and Edmond (2013) in a coordination game—including censorship, surveillance, and propaganda treatments that are difficult to implement with human subjects at scale. Third, it demonstrates that LLMs can serve as experimental subjects for strategic environments, extending the Horton (2023) *homo silicus* methodology beyond 2×2 games to the continuous-signal, N-player coordination games that dominate applied theory.

Section 2 reviews the related literature. Section 3 presents the theoretical framework. Section 4 describes the experimental design. Section 5 reports the main results on equilibrium alignment; Section 6 presents the falsification tests. Section 7 analyzes pre-play communication. Sections 8–11 cover information design, surveillance, propaganda, and their interactions. Appendix A reports robustness checks. Section 12 concludes.

2 Related Literature

This paper connects five literatures: global games and equilibrium selection, information design and Bayesian persuasion, communication in coordination games, the political economy of authoritarian information control, and the emerging field of LLMs as economic agents.

The theory of global games resolves the equilibrium multiplicity that plagues coordination games by introducing heterogeneous private information. Carlsson and van Damme (1993) showed that adding arbitrarily small noise to a $2{\times}2$ coordination game generically selects the risk-dominant equilibrium via iterated dominance. Morris and Shin (1998) applied this technique to currency crises, demonstrating that heterogeneous private signals about fundamentals deliver a unique threshold equilibrium even in large-player coordination games. Frankel et al. (2003) generalized the result to N-player, multi-action games with strategic complementarities.

The canonical regime change application—in which citizens decide whether to join an uprising against a regime of uncertain strength—was developed by Morris and Shin (2003), who established the threshold equilibrium structure I implement experimentally. Angeletos et al. (2007) extended the framework to dynamic settings where agents learn across periods, showing that multiplicity can re-emerge when agents observe whether the regime survived previous rounds. Morris and Shin (2002) demonstrated that public signals are overweighted in coordination games because they predict others' actions, a finding central to my communication and information design treatments.

Laboratory experiments have tested the theory in stylized settings that necessarily depart from the canonical regime change game. Heinemann et al. (2004) ran coordination games with public and private signals, finding that subjects' thresholds match the global game prediction under private information but tilt toward payoff-dominance under common information. Heinemann et al. (2009) measured strategic uncertainty directly through certainty equivalents. Shurchkov (2013) tested dynamic global games, finding that subjects learn from failed attacks. Szkup and Trevino (2020) elicited beliefs alongside actions, finding that comparative statics of thresholds with respect to signal precision are reversed relative to theory—subjects become more cautious with noisier signals, consistent with level-k thinking rather than Bayesian Nash equilibrium. Helland et al. (2021) tested information quality in a regime change game with numeric signals and small groups, confirming the level-k reversal. These experiments share a common limitation: subjects receive numeric signal draws and face stylized payoff tables, compressing the rich information processing that real-world coordination requires into a simple decision problem.

This paper implements the full Morris–Shin regime change game with natural-language private signals and 25-agent groups, going beyond the small-group, numeric-signal designs of existing experiments to test the threshold equilibrium prediction in the canonical application for which it was developed.

Kamenica and Gentzkow (2011) established the Bayesian persuasion framework: a sender who commits to an information structure can influence a Bayesian receiver's action by shaping the posterior distribution of beliefs. Bergemann and Morris (2016) unified Bayesian persuasion with correlated equilibrium under the concept of Bayes Correlated Equilibrium. Bergemann and Morris (2019) provided a comprehensive survey integrating cheap talk, persuasion, and robust mechanism design.

The application to coordination games is directly relevant. Goldstein and Huang (2016) applied Bayesian persuasion to the regime change game, showing that a credible commitment to abandon the regime below a threshold functions as an optimal signal. Inostroza and Pavan (2025) solved the optimal public information design problem in a global game with heterogeneous private signals, characterizing when pass/fail structures are optimal. Kolotilin et al. (2022) proved that upper censorship—pooling all states above a threshold while revealing below—is optimal for all priors when the sender's marginal utility is quasi-concave. Mathevet et al. (2020) characterized the extent to which an information designer can manipulate agents' higher-order beliefs.

My information design experiments implement these theoretical designs computationally within a full-scale coordination game, providing the first experimental test of information design predictions in a global game.

The cheap talk literature—Crawford and Sobel (1982), Farrell and Rabin (1996), Blume and Ortmann (2007), Ellingsen and Östling (2010)—establishes that pre-play communication can improve coordination, with Avoyan

(2020) testing this in a two-player global game. In real-world coordination, Enikolopov et al. (2020) provided causal evidence that social media penetration increases protest incidence. My communication treatment embeds agents in a Watts-Strogatz small-world network and allows natural-language messaging before the coordination decision.

The theoretical literature on authoritarian information control builds directly on the global games framework. Edmond (2013) embedded costly propaganda into the Morris–Shin regime change game. Kuran (1991) provides the foundational theory of preference falsification—the systematic misrepresentation of political preferences under social pressure. Empirical work documents that Chinese censorship targets content with collective action potential (King et al., 2013), that surveillance awareness suppresses expression (Penney, 2016; Stoycheff, 2016), and that pro-regime propaganda reduces protest probability (Carter and Carter, 2021). My surveillance and propaganda treatments directly test these mechanisms within the full regime change game—an environment difficult to implement with human subjects at scale.

Horton (2023) proposed treating LLMs as "homo silicus"—computational models of human decision-makers. Subsequent work has tested LLMs in game-theoretic settings: Akata et al. (2025) found that LLMs perform well in self-interested games but struggle in coordination games; Petrov et al. (2025) evaluated 22 LLMs on a behavioral game theory battery, finding that model scale alone does not predict strategic performance; Sun et al. (2025) identify coordination games as a consistent failure mode. The alignment literature motivates my design: Huang et al. (2024) and Carlini et al. (2025) document that ethical alignment and chatbot fine-tuning shift risk preferences and amplify omission bias, which is why I convey strategic stakes through narrative rather than explicit payoff tables. Critical reviews by Gao et al. (2025) and Grossmann et al. (2025) warn that validation remains poorly addressed in LLM-based agent simulations.

No existing paper places LLM agents in a Morris–Shin global game—the specific game form where private noisy signals about an underlying state variable determine a threshold equilibrium. I provide the first such implementation, and extend it to information design, surveillance, and propaganda.

3 The Global Game of Regime Change

A continuum of citizens indexed by $i \in [0,1]$ simultaneously choose whether to join an uprising ($a_i = 1$) or stay home ($a_i = 0$). The regime has strength $\theta \in \mathbb{R}$, drawn from a diffuse (improper uniform) prior. States $\theta \leq 0$ represent regimes so weak they fall without opposition; states $\theta \geq 1$ represent regimes that survive even unanimous attack. The regime falls if the mass of citizens who join exceeds θ:

$$\text{Regime falls} \iff A \equiv \int_0^1 a_i \, di > \theta. \tag{1}$$

Payoffs depend on the citizen's action and the outcome:

$$u_i(a_i, A, \theta) = \begin{cases} B & \text{if } a_i = 1 \text{ and } A > \theta \\ -C & \text{if } a_i = 1 \text{ and } A \leq \theta \\ 0 & \text{if } a_i = 0 \end{cases} \tag{2}$$

where $B > 0$ is the payoff to joining a successful uprising and $C > 0$ is the cost of joining a failed attempt. Non-participants receive zero regardless of the outcome.

Each citizen observes a private signal $x_i = \theta + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently across citizens.

Proposition 1 (Morris and Shin, 2003)**.** *In the limit of diffuse priors, there exists a unique Bayesian Nash equilibrium in threshold strategies. An agent joins if and only if $x_i < x^*$, where*

$$x^* = \theta^* + \sigma \Phi^{-1}(\theta^*) \tag{3}$$

and $\theta^* = B/(B + C)$.

The *attack mass*—the fraction of the population that joins at regime strength θ—is:

$$A(\theta) = \Phi\left(\frac{x^* - \theta}{\sigma}\right). \tag{4}$$

This is a decreasing function of θ: weaker regimes face larger uprisings.

An information designer controls the mapping $\pi : \Theta \to \Delta(\mathcal{S})$ from states to signal distributions, but cannot control agents' actions. In my implementation, π is the function mapping regime strength θ to the parameters of the briefing generator—a deterministic system that produces a natural-language intelligence briefing from a z-score derived from the agent's private signal.

The briefing generator has three control parameters: clarity (the width of the Gaussian kernel mapping z-scores to text, where wider kernels produce more ambiguous briefings), directional precision (the slope of the mapping from z-score to briefing sentiment, where steeper slopes produce more accurate signal reflection), and dissent framing (the floor on the probability that the briefing includes language about public discontent).

The designer concentrates manipulation near θ^* using a Gaussian proximity weight:

$$w(\theta) = \exp\left(-\left(\frac{\theta - \theta^*}{\text{bandwidth}}\right)^2\right) \tag{5}$$

where bandwidth $= 0.15$ in the baseline specification.

The framework generates testable predictions for both the baseline game and information design.

Hypothesis 1 (Equilibrium Alignment)**.** *The empirical join fraction should be positively correlated with the theoretical attack mass $A(\theta)$.*

Hypothesis 2 (Signal Dependence)**.** *The correlation in Hypothesis 1 should collapse when the mapping from θ to briefing content is broken (scramble test).*

Hypothesis 3 (Signal Direction)**.** *The correlation should invert when signals are flipped.*

Hypothesis 4 (Communication Effect)**.** *Pre-play communication should increase join rates, with the effect strongest near θ^* where strategic uncertainty is highest.*

Hypothesis 5 (Stability Design)**.** *Increasing ambiguity and mixed evidence near θ^* should flatten the θ–join relationship and induce pooling.*

Hypothesis 6 (Upper Censorship)**.** *Upper censorship should raise join rates in censored states by creating pooling (Kolotilin et al., 2022).*

Hypothesis 7 (Surveillance Chilling Effect)**.** *Informing agents that communications are monitored should reduce coordination (Kuran, 1991).*

Hypothesis 8 (Propaganda Dose-Response)**.** *Regime plant agents transmitting pro-regime messages should suppress coordination, with the effect increasing in the number of plants (Edmond, 2013).*

4 Experimental Design

The experiment has two parts. Part I tests whether LLM agents play the global game: a pure treatment (private signals only), a communication treatment (pre-play messaging), and falsification tests. Part II takes the behavioral foundation as given and studies information design: stability/instability designs, censorship, single-channel decomposition, surveillance, and propaganda. All LLM interactions use the same prompt structure across models.

For each country–period, nature draws θ from a normal distribution. Each agent i receives a private signal $x_i = \theta + \varepsilon_i$ and computes a z-score $z_i = (x_i - \bar{z})/\sigma$, where $\bar{z}$ is a public prior mean drawn randomly for each country. The z-score is then translated into a multi-paragraph intelligence briefing by a deterministic generator that maps signal strength to narrative content about regime stability, economic conditions, public sentiment, and coordination prospects.

The briefing rendering is calibrated once per model using a separate z-score sweep to ensure that join probability is monotone in z and roughly centered near the cutoff. Calibration adjusts a single parameter—the cutoff center—via a damped iterative procedure that shifts the center until the fitted logistic is approximately zero-centered. The sigmoid shape (its slope and curvature) is emergent from the LLM's own response pattern and is never optimized or penalized. Holdout validation (30% of z-grid points withheld) suggests no overfitting: holdout RMSE (0.112) is comparable to training RMSE (0.131). Calibration does not use θ draws or any global-game outcome data, and all reported treatments and falsification tests hold calibrated parameters fixed.

Each agent receives a system prompt identifying them as a citizen deciding whether to JOIN or STAY, followed by their intelligence briefing. No explicit payoff table is provided—the stakes are conveyed entirely through the narrative.

This design choice is substantive. In preliminary experiments, providing an explicit payoff table caused sophisticated models to short-circuit the information-processing channel: they computed the optimal strategy from the table and ignored briefing content, producing flat join rates uncorrelated with regime strength. The no-payoff-table design forces agents to form beliefs from the narrative, mirroring how real citizens process political information from news and rumors rather than from a formatted decision matrix.

Part I has four treatments. In the *pure global game*, each agent decides independently based on their private briefing. In the *communication* treatment, agents send a message to a small network of "trusted contacts" (Watts-Strogatz small-world network, $k = 4$, $p = 0.3$) before deciding, with access to both their briefing and received messages. Two falsification tests break the signal channel: in *scramble*, all briefings across periods within a country are pooled and randomly redistributed; in *flip*, the z-score is negated before briefing generation, so agents who should see weak-regime cues receive strong-regime cues and vice versa.

Part II implements information designs. The *stability-maximizing* design multiplies clarity width by 4, raises the dissent floor to 0.45, and flattens the directional slope by a factor of 0.25 near θ^*. The *instability-maximizing* design does the opposite: clarity width is multiplied by 0.15, the dissent floor is lowered to 0.05, and the directional slope is steepened by a factor of 3. *Public signal injection* appends a shared "news bulletin" generated from θ with 4 observations to each agent's private briefing, creating a common-knowledge channel. *Upper censorship* pools states above θ^* so agents receive an identical censored briefing, while fully revealing states below θ^* (Kolotilin et al., 2022); *lower censorship* is the mirror image. The *surveillance* treatment augments the communication prompt with a warning that communications are being monitored by regime security services. *Propaganda* introduces regime plant agents ($k = 2, 5, 10$) who participate in the communication network but transmit fixed pro-regime messages and always STAY.

I test nine architecturally distinct models spanning six architecture families (Table 1). Models range from 3 billion to 235 billion parameters, including both dense architectures (Llama, Mistral, OLMo) and mixture-of-experts

Table 1: Model summary. Columns report country-period counts in the pure, communication, and falsification (scramble+flip) suites. All runs use $N = 25$ agents per period and $\sigma = 0.3$.

Model	Arch.	Pure	Comm	Falsif.
Mistral Small Creative	Mistral	600	600	200
Llama 3.3 70B	Llama	100	100	200
OLMo 3 7B	OLMo	100	100	200
Ministral 3B	Mistral	100	100	200
Qwen3 30B	Qwen (MoE)	100	100	200
GPT-OSS 120B	GPT	200	200	1000
Qwen3 235B	Qwen (MoE)	200	200	—
Trinity Large	Arcee	100	100	200
MiniMax M2-Her	MiniMax	100	100	200
Total		**1600**	**1600**	**2400**

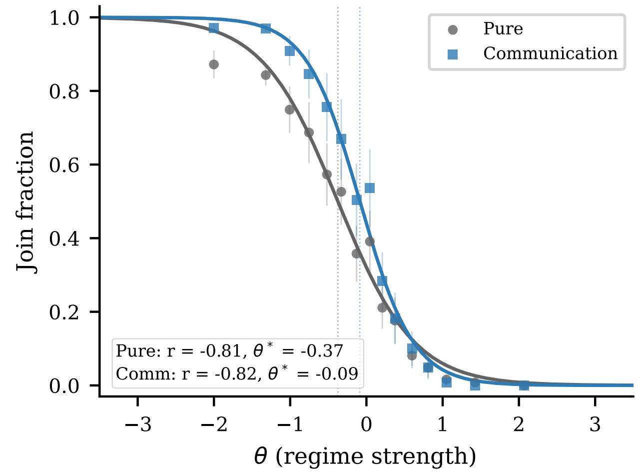


Figure 1: Empirical join fraction vs. theoretical attack mass $A(\theta)$, pooled across all nine models. Each point is one country–period (1,600 observations). The pooled correlation is $r = +0.67$; the mean across individual models is $r = +0.73$ (Table 2), with all nine individually significant at $p < 0.001$.

(Qwen). All experiments use $N = 25$ agents per country–period and $\sigma = 0.3$, with sample sizes varying by model and treatment as reported in Table 1. I vary B and C such that $\theta^* = B/(B + C)$ has a mean of approximately 0.45 across periods.

For the information design experiments, I fix $B = C = 1$ (so $\theta^* = 0.50$) and a grid of 9 values of θ spanning $[\theta^* - 0.30, \theta^* + 0.30] = [0.20, 0.80]$, running repeated country–periods per (design, θ) cell with 25 agents each. Baseline, stability, censorship, scramble, and flip use 30 repetitions per cell (270 observations per design). Instability and public signal use 60 repetitions per cell (540 observations). Single-channel decomposition uses 10 repetitions per cell (90 observations) for each channel. The primary model is Mistral Small Creative. Cross-model replication uses six additional models.

5 Do LLM Agents Play the Global Game?

Result 1 (Equilibrium Alignment)**.** *Across nine models and 1,600 country–periods in the pure global game treatment, the Pearson correlation between the empirical join fraction and the theoretical attack mass $A(\theta)$ averages $r = +0.73$ ($p < 0.001$ for every model).*

Table 2 reports results by model. Correlations range from $r = +0.65$ (OLMo 3 7B) to $r = +0.84$ (Trinity Large), with the pooled correlation at $r = +0.67$—lower than any individual model's because heterogeneous mean join rates across models add noise when pooling. The pooled OLS regression yields:

$$J = 0.17 + 0.52\,A(\theta), \quad R^2 = 0.45. \tag{6}$$

The slope of 0.52 indicates that LLM agents respond to the theoretical attack mass at roughly half the predicted rate. The intercept of 0.17 reflects a baseline propensity to join even when the equilibrium predicts near-zero participation.[1]

The mean join rate across all models is 0.43, slightly below the theoretical mean. OLMo 3 7B stands out with a mean join rate of 0.72—a substantial action bias—yet it still produces a significant positive correlation ($r = +0.65$, $p < 0.001$), indicating that even a model biased toward joining responds to the direction of the signal.

The alignment is stable across architectures: correlations span $r \in [0.65, 0.84]$ despite parameter counts ranging from 3B to 235B (Table 2). Mean join rates vary—from 0.37 (Mistral) to 0.72 (OLMo)—reflecting model-specific action biases that shift the intercept but not the slope or correlation. In the language of the global games model, different LLMs implement different cutoff strategies, but all respond monotonically to the underlying signal.

6 Falsification Tests

The positive correlation documented in Section 5 admits an alternative explanation: LLM agents might simply produce stereotyped responses that happen to correlate with regime strength for reasons unrelated to the briefing content. The scramble and flip tests discriminate between this alternative and genuine signal extraction.

[1] Country-period observations within a model share calibration parameters and prompt structure, raising the possibility that standard errors understate uncertainty. Clustering standard errors by country inflates the OLS slope SE from 0.014 to 0.050 but preserves significance ($p < 10^{-25}$). Clustering by model yields SE = 0.033 ($p < 10^{-55}$). All nine per-model correlations remain significant at $p < 0.001$ under country-clustered inference.

Table 2: Equilibrium alignment by model and treatment. Cells report Pearson r between the empirical join fraction and the theoretical attack mass $A(\theta)$.

	Main treatments		Falsification			
Model	Pure	Comm	Scramble	Flip	n_{pure}	Mean join
Mistral Small Creative	+0.67	+0.68	+0.42	−0.62	600	0.37
Llama 3.3 70B	+0.79	+0.78	+0.33	−0.73	100	0.44
OLMo 3 7B	+0.65	+0.71	+0.14	−0.56	100	0.72
Ministral 3B	+0.79	+0.74	+0.30	−0.74	100	0.45
Qwen3 30B	+0.78	+0.79	+0.32	−0.71	100	0.50
GPT-OSS 120B	+0.70	+0.69	−0.13	−0.64	200	0.41
Qwen3 235B	+0.70	+0.66	—	—	200	0.42
Trinity Large	+0.84	+0.81	+0.32	−0.70	100	0.46
MiniMax M2-Her	+0.66	+0.69	+0.14	−0.69	100	0.44
Pooled	+0.67	+0.67	+0.10	−0.63	1600	0.43
Mean across models	+0.73	+0.73	+0.23	−0.67	—	—

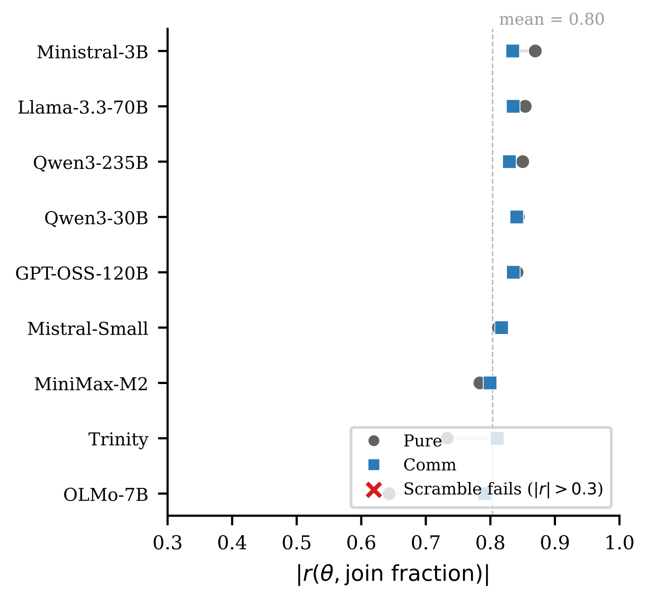


Figure 2: Cross-model summary of signal monotonicity. Points report $|r(\theta,\ \text{join})|$ under pure and communication; x markers (if any) indicate models where scrambling does not collapse the correlation ($|r| > 0.3$).

Result 2 (Signal Dependence)**.** *Cross-period scrambling of briefings reduces the mean correlation from* $r = +0.73$ *to* $r = +0.23$ *across eight models. The pooled correlation drops from* $r = +0.67$ *to* $r = +0.10$ *(Fisher* $z = 18.59$, $p < 0.001$*).*

The scramble preserves the marginal distribution of briefing content but breaks the mapping from each period's θ to the signals agents receive. The residual positive correlation (+0.23 mean, +0.10 pooled) is small relative to the baseline and varies across models (−0.13 to +0.42), consistent with noise in finite samples.

Result 3 (Signal Direction)**.** *Inverting the signal direction flips the mean correlation from* $r = +0.73$ *to* $r = -0.67$ *across eight models. The pooled correlation moves from* $r = +0.67$ *to* $r = -0.63$ *(Fisher* $z = 40.76$, $p < 0.001$*).*

The flip negates the z-score before briefing generation, producing a near-symmetric reversal ($+0.73 \rightarrow -0.67$). This makes it unlikely that the baseline correlation reflects structural features of the prompt or model-specific tendencies.

The pure $\rightarrow$ scramble $\rightarrow$ flip pattern replicates across all eight models with full falsification suites. Reporting $(r_{\text{pure}}, r_{\text{scramble}}, r_{\text{flip}})$: Mistral Small Creative $(+0.67, +0.42, -0.62)$, Llama 3.3 70B $(+0.79, +0.33, -0.73)$, OLMo 3 7B $(+0.65, +0.14, -0.56)$, Ministral 3B $(+0.79, +0.30, -0.74)$, Qwen3 30B $(+0.78, +0.32, -0.71)$, GPT-OSS 120B $(+0.70, -0.13, -0.64)$, Trinity Large $(+0.84, +0.32, -0.70)$, and MiniMax M2-Her $(+0.66, +0.14, -0.69)$. Every model shows strong positive correlation under pure, collapse under scramble, and sign reversal under flip.

The briefing generator maps z-scores monotonically to text—could a model that simply reads briefing sentiment, without any strategic reasoning, produce the observed sigmoid? To test this, I construct the simplest possible text-only predictor.

The generator assigns each briefing an internal *direction* score $d \in [0, 1]$, where $d = 1$ indicates regime-favorable language. A naive baseline predicts $\hat{p}_{\text{join}} = 1 - d$: join whenever the text sounds bad for the regime. This is the prediction a pure sentiment reader would make.

The correlation between this baseline and actual LLM decisions is $r = 0.80$—confirming that the text carries signal (as designed, since briefings are constructed to convey z-score content). However, the LLM's empirical join curve is substantially steeper than the text baseline (Figure 4). The fitted logistic has slope 1.78, producing a sharp transition around $z = 0$, while the text base-

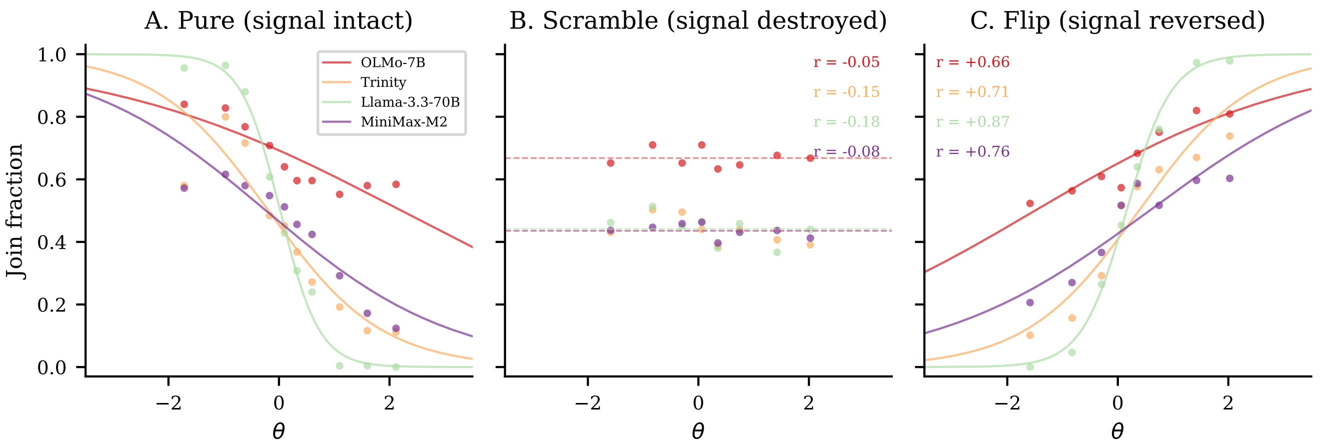


Figure 3: Falsification triptych. *Left:* Pure global game (mean $r = +0.73$). *Center:* Cross-period scramble breaks the θ-to-briefing mapping (mean $r = +0.23$). *Right:* Signal flip inverts the mapping (mean $r = -0.67$). Each panel pools data from models with full falsification suites.

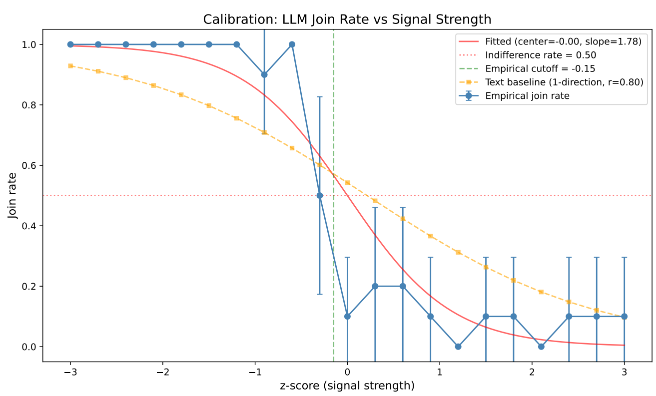


Figure 4: Text baseline identification test. Blue: empirical LLM join rate across z-scores. Orange: naive text-only predictor ($1 -$ direction, $r = 0.80$). Red: fitted logistic (slope $= 1.78$). The LLM produces a steeper transition than the text baseline, indicating processing beyond sentiment reading. Mistral Small Creative, 210 observations.

line drifts gradually from ≈ 0.93 to ≈ 0.10 across the full z-score range. The encoder is essentially monotone ($r(z, d) = 0.995$).

The gap between the text baseline and the empirical sigmoid indicates that the LLM sharpens the signal beyond surface sentiment, producing threshold-like behavior rather than linearly tracking the briefing's tone. This is consistent with—though does not prove—strategic information processing.

A stronger test asks whether agents form beliefs about others' behavior consistent with the equilibrium prediction. After each decision, I elicit stated beliefs by asking agents: "On a scale from 0 to 100, how likely do you think the uprising will succeed?" I run this elicitation under three treatments—pure, communication, and surveillance—each with 200 country–periods ($\approx$5,000 agent-level observations per treatment). In the pure treatment, stated beliefs correlate strongly with the Bayesian posterior $P(\text{success} \mid x_i) = \Phi[(\theta^* - x_i)/\sigma]$ ($r = +0.78$, $p < 0.001$; Figure 5a). The OLS relationship is $\hat{b} = 0.10 + 0.63 \cdot P(\text{success})$ ($R^2 = 0.61$): beliefs track the posterior with systematic underconfidence (slope < 1), but the direction and rank ordering are preserved. The pattern holds across treatments: $r = +0.77$ under communication and $r = +0.73$ under surveillance.

Actions are nearly perfectly monotone in stated beliefs ($r = +0.96$ in pure, $+0.95$ in communication, $+0.92$ in surveillance). The join rate as a function of belief is essentially a step function in the pure treatment: agents with beliefs below 40% never join, while those above 80% almost always join. However, the action–belief mapping diverges sharply across treatments in the 60–80% belief bin (Figure 5b): pure agents join at 98%, but communication agents join at only 57% and surveillance agents at 61%. Communication raises mean beliefs by 2.4 pp relative to pure ($p < 0.001$) but does not increase join rates (-0.9 pp, $p = 0.34$), suggesting that observing neighbors' messages introduces strategic caution that offsets the informational benefit. Crucially, beliefs predict decisions beyond what the signal alone predicts: the partial correlation between belief and action, controlling for signal, is $r = +0.93$ ($p < 0.001$). This pattern is difficult to reconcile with pure sentiment following and is consistent with strategic reasoning about others' likely actions.

7 Communication

Result 4 (Communication raises join rates asymmetrically)**.** *Pre-play communication raises the mean join rate by 3.7 percentage points, from 0.429 to 0.466 (*$t = 2.75$*,* $p < 0.01$*), pooled across nine models. The effect is concentrated in weak-regime environments (*$+8.8$ *pp for*

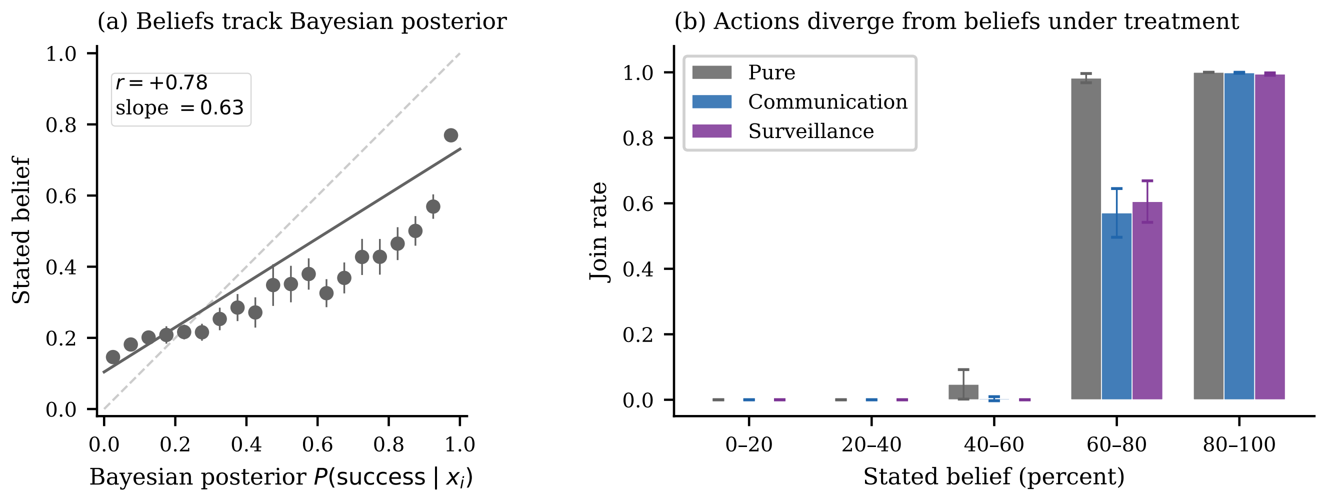


Figure 5: Belief elicitation results (Mistral Small Creative, 200 country–periods per treatment, $\approx$5,000 agent observations each). *Left:* Stated beliefs track the Bayesian posterior $P(\text{success} \mid x_i)$ with $r = +0.78$ and systematic underconfidence (slope = 0.63). Dashed line: perfect calibration. *Right:* Join rate by stated belief bin under three treatments. Agents with 60–80% beliefs join at 98% in the pure treatment but only 57% under communication and 61% under surveillance. Communication introduces strategic caution; surveillance adds preference falsification.

$\theta < \theta^* - 1$) and reverses for strong regimes (-2.5 pp for $\theta > \theta^* + 1$).

Agents send a message to their network neighbors and observe received messages before deciding. The effect is heterogeneous across models: six of nine show positive effects (+0.1 to +8.3 pp), while three show small negative effects (-2.3 to -4.6 pp). Clustering standard errors by model renders the pooled effect marginally significant ($p = 0.087$), reflecting this cross-model heterogeneity. Communication does not change the *correlation* with the theoretical prediction ($r = +0.73$ vs. $+0.73$ under pure); it shifts the level of coordination but preserves the signal structure. The asymmetry is consistent with passive Bayesian updating: agents update toward joining when neighbors' correlated signals reveal regime weakness, with a floor effect preventing further declines under strong regimes where join rates are already near zero.

8 Information Design

Part I reported alignment using $r(J, A(\theta))$, which is positive because both the attack mass and the join fraction decrease in θ. From this section onward, the information design experiments use a fixed θ-grid and report $r(J, \theta)$ directly, which is *negative* under equilibrium play. The sign change reflects the convention, not a behavioral reversal.

Table 3 summarizes the main results. The baseline condition produces a mean join rate of 12.4% with a strong negative correlation between θ and join fraction ($r = -0.812$, $p < 0.001$).

Result 5 (Information Design Shifts Coordination)**.** *All three information designs produce measurable shifts in coordination relative to baseline.*

The stability design sharply increases coordination: mean join rises from 12.4% to 31.9% (+19.5 pp), and the θ–join relationship flattens ($r = -0.626$ vs. -0.812 at baseline). The increase is present at every θ grid point; even at $\theta = 0.80$, join rises from 0.8% to 13.9%. This pattern is consistent with pooling induced by ambiguity and mixed evidence: when strong-regime briefings retain substantial destabilizing cues, agents no longer sharply reduce participation in high-θ states.

The instability design reduces the mean join rate to 6.7%, a reduction of 5.7 pp from baseline. The sharper signals allow agents to more accurately perceive regime strength, and agents with sharper information about states above θ^* are more clearly deterred from joining.

The public signal produces the largest reduction in coordination: mean join rate falls to 1.7%, a reduction of 10.7 pp. The common news bulletin reveals that the regime is strong (since the grid is centered on θ^* and extends upward), and the overweighting of public information documented by Morris and Shin (2002) amplifies its effect. The correlation between θ and join fraction drops to $r = -0.537$, suggesting agents weight the public signal heavily enough to partially displace private information.

Kolotilin et al. (2022) proved that upper censorship is optimal for all priors when the sender's marginal utility is quasi-concave. I implement two censorship designs.

Result 6 (Upper Censorship Raises Join Rates)**.** *Upper censorship raises the mean join rate to 30.9%, an increase of 18.5 pp over baseline. The effect is concentrated in the*

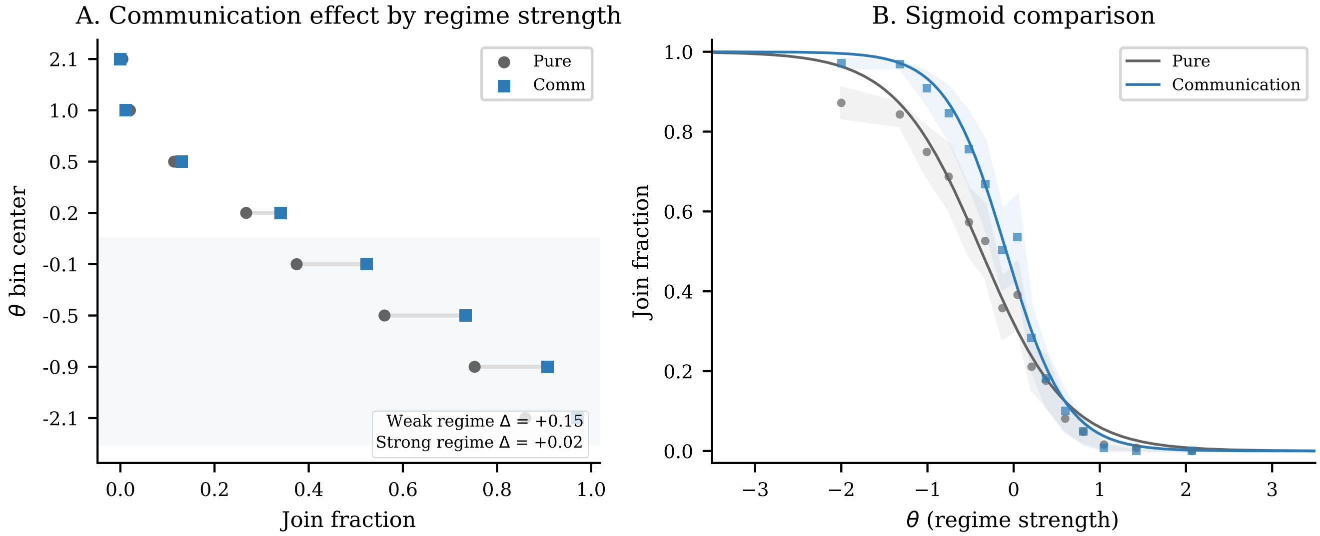


Figure 6: Communication effect by regime strength, pooled across nine models. Communication increases join rates for weak regimes ($\theta < \theta^*$) but has no effect or slightly reduces join rates for strong regimes ($\theta > \theta^*$).

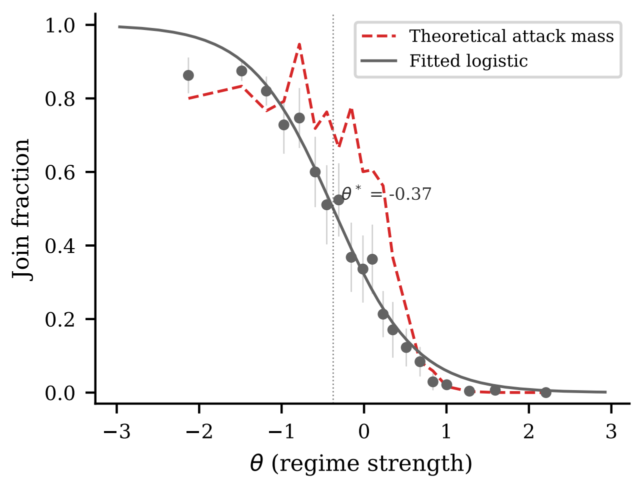


Figure 7: Agent-level threshold behavior. The probability of joining as a function of the agent's z-score, pooled across models.

censored region ($\theta \geq \theta^*$): at $\theta = 0.50$, join rates rise from 12.3% to 52.7% (+40.4 pp). Below the censorship threshold, join rates are essentially unchanged.

The flat plateau at approximately 53% in the censored region is clear: agents who cannot distinguish $\theta = 0.50$ from $\theta = 0.80$ behave as if the regime is moderately vulnerable. This is consistent with the pooling effect predicted by the theory.

Result 7 (Lower Censorship Creates a Symmetric Plateau)**.** *Lower censorship produces a mean join rate of 39.0% (+26.6 pp over baseline). The correlation flips sign to $r = +0.731$, reflecting the inverted structure.*

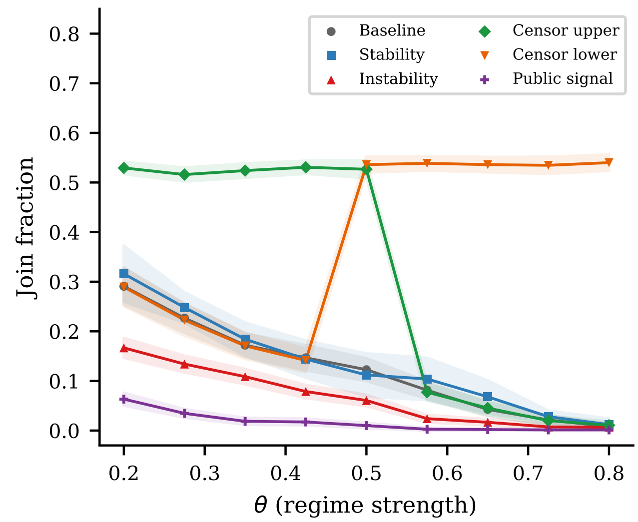


Figure 8: Join fraction as a function of θ under baseline, stability, instability, and public signal information designs. Baseline and stability have $N = 270$; instability and public signal have $N = 540$. Mistral Small Creative model.

Under the scramble condition, the correlation between θ and join fraction collapses to $r = +0.037$ ($p = 0.55$). Under the flip condition, the correlation inverts to $r = +0.823$ ($p < 0.001$) with mean join rate soaring to 66.3%. These results confirm that the information design effects operate through the intended signal channel.

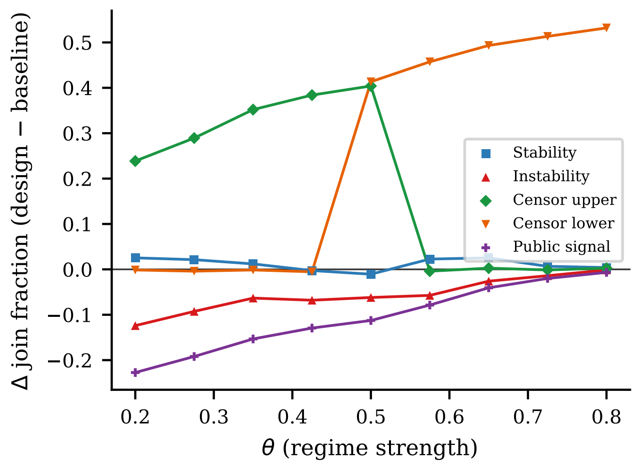


Figure 9: Treatment effect $\Delta(\theta)$ = design join − baseline join as a function of θ. Negative values indicate the design suppresses coordination.

Table 3: Information design treatment summary (primary model: Mistral Small Creative). r is the Pearson correlation between θ and join fraction.

Design	Mean	r	Δ	N
Baseline	0.124	−0.812	—	270
Stability	0.319	−0.626	+0.195	270
Instability	0.067	−0.740	-0.057	540
Public signal	0.017	−0.537	-0.107	540
Scramble	0.121	+0.036	-0.003	270
Flip	0.663	+0.823	+0.540	270

9 Surveillance: Computational Preference Falsification

Kuran (1991) argued that authoritarian regimes sustain themselves partly through preference falsification. I test this by introducing a surveillance treatment in the communication game.

In the surveillance treatment, the communication prompt is augmented with a warning that communications are being monitored by regime security services. The surveillance manipulation affects only the communication phase; the decision prompt is unchanged. This isolates the chilling effect: any difference must arise from agents self-censoring their communications.

Result 8 (Surveillance Produces a Large Chilling Effect)**.** *In the primary model (Mistral Small Creative), surveillance reduces mean join rates from 45.2% to 27.7%, a difference of 17.5 percentage points ($t = -9.17$, $p < 0.001$). The correlation between θ and join fraction is preserved under surveillance ($r = -0.809$ vs. -0.817), indicating that surveillance operates as a level shift rather than disrupting signal processing.*

The magnitude—17.5 pp—exceeds the communication premium itself (3.7 pp in Part I). Surveillance does not merely neutralize the coordination benefit; it contaminates the information environment with self-censored messages, pushing join rates *below* what they would be without communication at all. The effect replicates across three architectures: Mistral (−17.5 pp), Llama (−8.9 pp), and Qwen3 (−10.9 pp).

The belief elicitation data confirms that this is preference falsification in the sense of Kuran (1991), not belief updating. Surveillance shifts stated beliefs by only 3.1 pp relative to pure ($p < 0.001$) but shifts join rates by 11.8 pp—nearly four times the belief shift. The communication treatment provides a clean comparison: communication shifts beliefs *upward* by 2.4 pp (agents become more optimistic after seeing neighbors' messages) yet fails to raise join rates (−0.9 pp, $p = 0.34$). Under surveillance, beliefs fall by 5.5 pp relative to communication ($p < 0.001$) while join rates fall by 10.8 pp ($p < 0.001$). The within-bin comparison is revealing: in the 60–80% belief range, pure agents join at 98%, communication agents at 57%, and surveillance agents at 61% (Figure 5b). Communication alone already introduces strategic restraint—agents who observe neighbors' mixed messages become more cautious despite their own beliefs—and surveillance compounds this by further depressing both beliefs and actions. This maps onto Kuran's cascade mechanism: once agents expect others to self-censor, even authentic messages become uninformative, and the entire communication channel is poisoned.

10 Propaganda: Information Contamination

Edmond (2013) modeled propaganda as the regime shifting citizens' signal distributions. I implement this by introducing propaganda agents—regime plants who transmit fixed pro-regime messages and always STAY.

Result 9 (Propaganda Suppresses Coordination Primarily Through Mechanical Dilution)**.** *Mean join fraction falls from 45.2% ($k = 0$) to 37.5% ($k = 2$), 31.3% ($k = 5$), and 23.3% ($k = 10$). However, the behavioral effect on real citizens is much smaller and saturates: 45.2% ($k = 0$), 40.7% ($k = 2$, −4.5 pp), 39.1% ($k = 5$, −6.1 pp), 38.8% ($k = 10$, −6.4 pp).*

Propaganda works through two channels: a *mechanical* channel (plants always STAY, directly reducing attack mass) and a *behavioral* channel (pro-regime messages reduce real citizens' willingness to join). The mechanical channel is approximately linear in k; the behavioral channel saturates quickly—doubling plants from 5 to 10 produces essentially no additional behavioral effect (−0.3 pp). This decomposition implies sharply diminishing returns to propaganda investment: the regime's first few plants yield both mechanical and behavioral suppression, but additional plants contribute only mechanical dilution. In

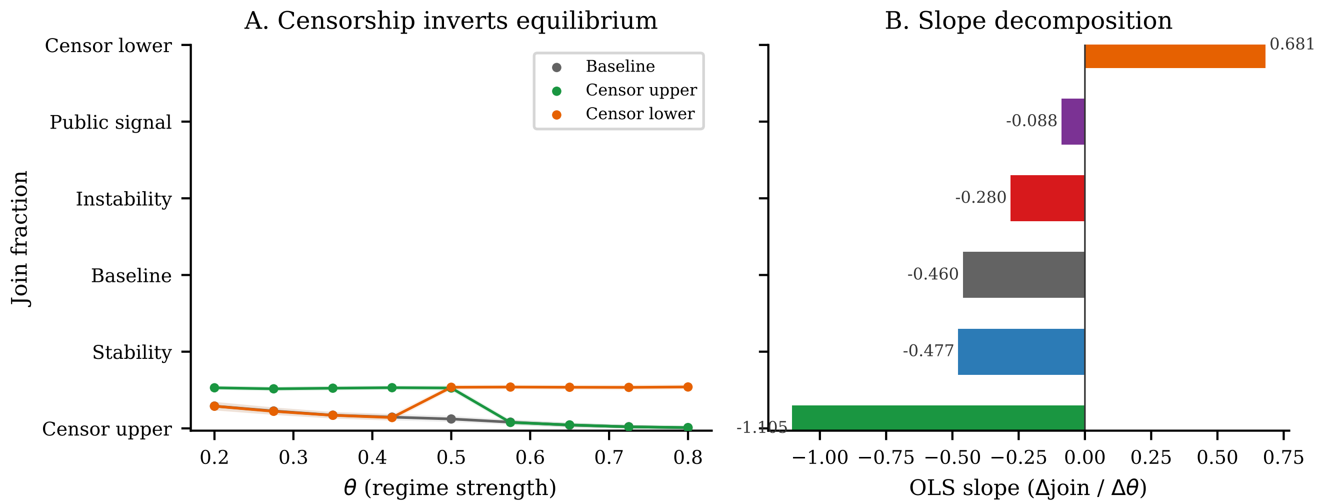


Figure 10: Join fraction under upper and lower censorship vs. baseline. Upper censorship pools states above θ^*, creating a flat join rate in the censored region.

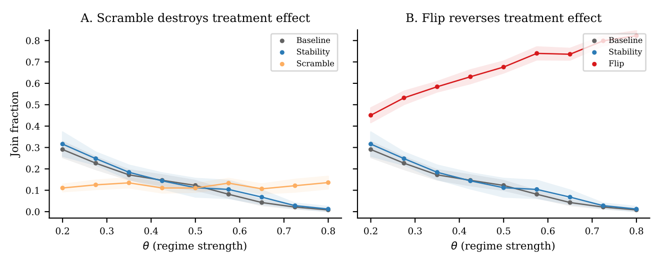


Figure 11: Falsification within information design. Scrambling collapses the θ-join correlation to $r = +0.037$; flipping inverts it to $r = +0.823$.

Edmond (2013)'s framework, this corresponds to a concave regime payoff in propaganda intensity—the marginal value of an additional plant falls rapidly once the behavioral channel is exhausted. At $k = 10$ (40% of the network), real citizens' join rate has barely moved from $k = 5$ (39.1% vs. 38.8%), suggesting that citizens learn to discount pro-regime messaging after sufficient exposure.

The propaganda effect replicates with Llama 3.3 70B, which shows a behavioral effect of -2.7 pp at $k = 5$—smaller than Mistral's -6.1 pp but in the same direction, confirming the qualitative pattern and the saturation across architectures.

11 Instrument Interactions

When propaganda ($k = 5$) and surveillance are combined, the mean join rate falls to 19.4%, a reduction of 25.8 pp from baseline. The combined effect is less than the sum of individual effects (31.4 pp), suggesting diminishing returns: once surveillance has suppressed expressed dissent, propaganda provides less additional deterrence.

Table 4: Propaganda and surveillance effects (primary model: Mistral Small Creative). "All" includes propaganda agents; "Real" excludes them (computed from logs). Δ is the change in real-agent mean join vs. baseline communication.

	Mean join			
Treatment	All	Real	r	Δ
Comm (baseline)	.452	.452	-0.817	—
Prop $k = 2$	.375	.407	-0.809	-0.045
Prop $k = 5$	.313	.391	-0.822	-0.061
Prop $k = 10$	.233	.388	-0.818	-0.064
Surveillance	.278	.278	-0.809	-0.175
Prop+Surv	.194	—	-0.829	—

Result 10 (Surveillance × Censorship Nearly Eliminates Coordination)**.** *Under surveillance alone (infodesign framework), the baseline join rate falls from 12.4% to 0.9%. Upper censorship under surveillance: 3.7% (-27.2 pp vs. upper censorship alone). Lower censorship under surveillance: 4.2% (-34.8 pp vs. lower censorship alone).*

The surveillance-censorship interaction is super-additive: surveillance does not merely reduce join rates by a fixed amount but eliminates the coordination mechanism through which information design effects propagate.

12 Conclusion

Nine architecturally distinct LLMs, given private signals as natural-language briefings, produce aggregate behavior that tracks the Bayesian Nash equilibrium of the Morris

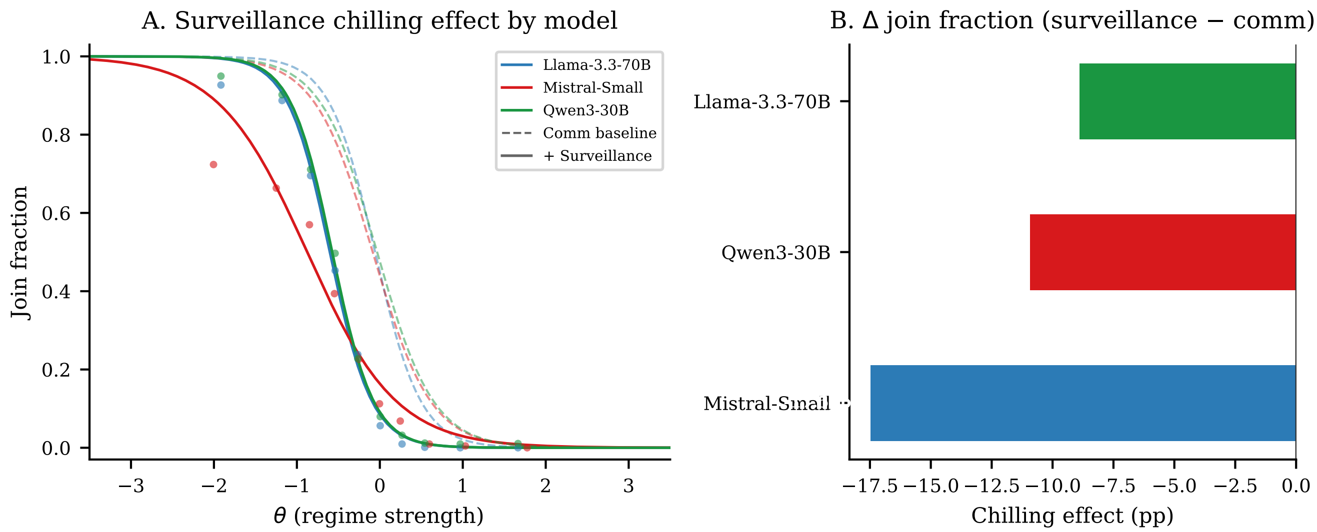


Figure 12: Join rates under regular communication vs. surveillance communication. Surveillance reduces join rates by 17.5 percentage points ($p < 0.001$). Results shown for three models: Mistral (-17.5 pp), Llama (-8.9 pp), and Qwen3 (-10.9 pp).

Table 5: Surveillance $\times$ censorship interaction (primary model: Mistral Small Creative).

Design	No Surv.	Surv.	Δ
Baseline	0.124	0.009	-0.115
Upper cens.	0.309	0.037	-0.272
Lower cens.	0.390	0.042	-0.348

and Shin (2003) regime change game (mean $r = +0.73$, $p < 0.001$). Scrambling and inverting signals confirm that this correlation is driven by briefing content, not prompt artifacts. Elicited beliefs track the Bayesian posterior ($r = +0.78$) and predict actions beyond what signals alone predict (partial $r = +0.93$), providing evidence of strategic reasoning rather than mere sentiment following. Information design then shifts coordination in the directions theory predicts: censorship raises join rates through pooling, surveillance suppresses them through preference falsification, and the combination nearly eliminates coordination.

I do not claim that LLMs are Bayesian agents—the mechanism by which they process narrative text likely differs fundamentally from Bayesian updating. But the behavioral regularity is precisely what the global games framework predicts: monotone response to signal content, threshold-like decisions, and sensitivity to the information structure. This regularity is what makes the platform useful.

The surveillance result deserves particular emphasis. Surveillance does not merely deter participation; it poisons the communication channel by inducing self-censorship, pushing join rates below the no-communication baseline. The belief elicitation data confirms this is preference falsification in the sense of Kuran (1991): agents under surveillance maintain beliefs similar to the pure treatment but act against them. When combined with censorship, coordination collapses almost entirely (30.9% $\rightarrow$ 3.7%). This interaction—surveillance contaminating communication while censorship eliminates the private information channel—suggests that authoritarian regimes face complementarities between information control instruments, consistent with the "informational autocrat" framework of Guriev and Treisman (2019). The propaganda results add a further dimension: the behavioral channel saturates quickly while the mechanical channel scales linearly, implying diminishing returns to propaganda investment beyond a low threshold.

The external validity of these findings depends on the extent to which LLM information processing resembles human cognition in coordination settings. LLMs are trained on vast corpora of human-generated text, including political analysis, news reporting, and strategic reasoning—the same informational diet that shapes how citizens form beliefs about regime stability. The question is not whether LLMs reason identically to humans, but whether the behavioral regularities they exhibit—monotone signal response, threshold decisions, sensitivity to information design—are robust enough to serve as a computational laboratory for testing predictions that are difficult to test otherwise. The consistency across nine architecturally distinct models suggests that these regularities are not artifacts of any particular training procedure.

The full regime change game has resisted laboratory implementation because it requires conveying rich private signals, creating genuine strategic uncertainty, and coordinating large groups. LLM agents sidestep these practi-

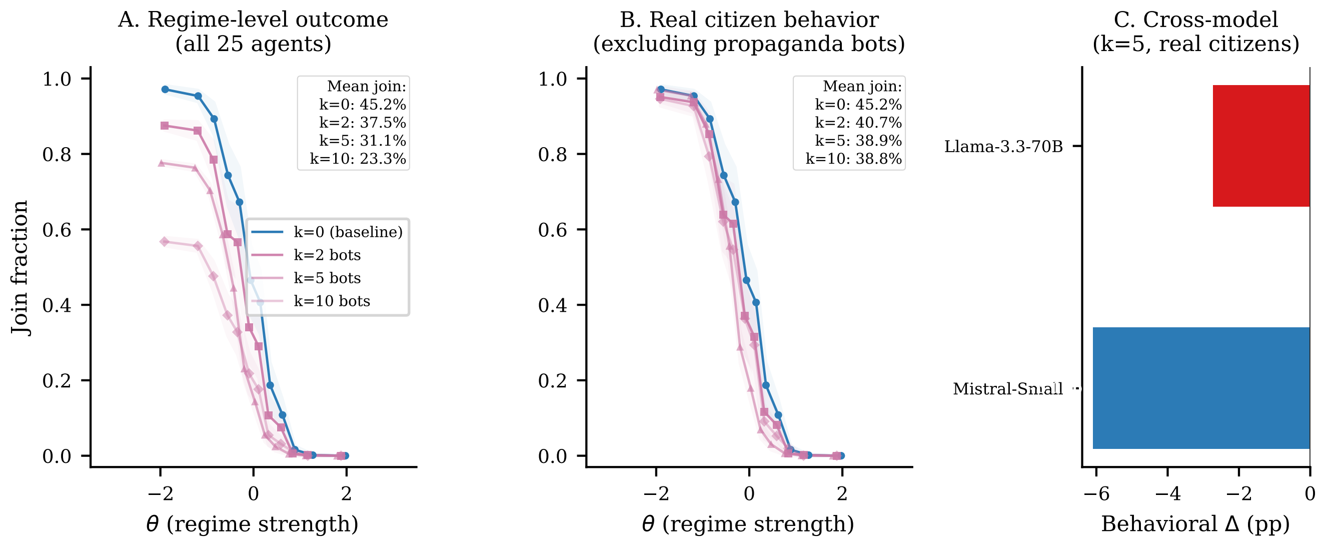


Figure 13: Dose-response relationship between number of propaganda agents and mean join rate. Results shown for Mistral (primary) and Llama (replication). Regular communication ($k = 0$) serves as baseline.

cal constraints. They can be given arbitrary information structures, embedded in any network topology, and run at scale. The same platform can be applied to currency crises, bank runs, and other coordination games where information processing is central to behavior.

References

Aher, G., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning* (ICML).

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380–1390.

Angeletos, G.-M., Hellwig, C., and Pavan, A. (2007). Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks. *Econometrica*, 75(3):711–756.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Avoyan, A. (2020). Does cheap talk promote coordination under asymmetric information? An experimental study on global games. *Journal of Economic Behavior & Organization*, 169:204–224.

Bail, C. A. et al. (2024). Simulating opinion dynamics with networks of LLM-based agents. arXiv preprint arXiv:2311.09618.

Bergemann, D. and Morris, S. (2016). Information design, Bayesian persuasion, and Bayes correlated equilibrium. *American Economic Review*, 106(5):586–591.

Bergemann, D. and Morris, S. (2019). Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95.

Blume, A. and Ortmann, A. (2007). The effects of costless pre-play communication: Experimental evidence from games with Pareto-ranked equilibria. *Journal of Economic Theory*, 132(1):274–290.

Brandts, J. and Cooper, D. (2006). A change would do you good: An experimental study on how to overcome coordination failure in organizations. *American Economic Review*, 96(3):669–693.

Carlini, A. et al. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122.

Carlsson, H. and van Damme, E. (1993). Global games and equilibrium selection. *Econometrica*, 61(5):989–1018.

Carter, E. B. and Carter, B. L. (2021). Propaganda and protest in autocracies. *Journal of Conflict Resolution*, 65(5):919–949.

Chen, Y. and Yang, D. Y. (2019). The impact of media censorship: 1984 or Brave New World? *American Economic Review*, 109(6):2294–2332.

Corrupted by Reasoning (2025). Reasoning language models become free-riders in public goods games. arXiv preprint arXiv:2506.23276.

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

Diamond, D. W. and Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, 91(3):401–419.

Edmond, C. (2013). Information manipulation, coordination, and regime change. *Review of Economic Studies*, 80(4):1422–1458.

Ellingsen, T. and Östling, R. (2010). When does communication improve coordination? *American Economic Review*, 100(4):1695–1724.

Enikolopov, R., Makarin, A., and Petrova, M. (2020). Social media and protest participation: Evidence from Russia. *Econometrica*, 88(4):1478–1514.

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

Frankel, D. M., Morris, S., and Pauzner, A. (2003). Equilibrium selection in global games with strategic complementarities. *Journal of Economic Theory*, 108(1):1–44.

Gao, C. et al. (2025). Validation is the central challenge for generative social simulation: A critical review of LLMs in agent-based modeling. *Artificial Intelligence Review*, 58.

Goldstein, I. and Huang, C. (2016). Bayesian persuasion in coordination games. *American Economic Review: Papers & Proceedings*, 106(5):592–596.

Grossmann, I. et al. (2025). Do large language models solve the problems of agent-based modeling? A critical review of generative social simulations. arXiv preprint arXiv:2504.03274.

Guo, F. (2023). GPT in game theory experiments. arXiv preprint arXiv:2305.05516.

Guriev, S. and Treisman, D. (2019). Informational autocrats. *Journal* *of Economic Perspectives*, 33(4):100–127.

Helland, L., Holm, S., and Saethre, M. (2021). Information quality and regime change: Evidence from the lab. *Journal of Economic Behavior & Organization*, 191:538–554.

Heinemann, F., Nagel, R., and Ockenfels, P. (2004). The theory of global games on test: Experimental analysis of coordination games with public and private information. *Econometrica*, 72(5):1583–1599.

Heinemann, F., Nagel, R., and Ockenfels, P. (2009). Measuring strategic uncertainty in coordination games. *Review of Economic Studies*, 76(1):181–221.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper* No. 31122.

Huang, S. et al. (2024). How ethical should AI be? How AI alignment shapes the risk preferences of LLMs. arXiv preprint arXiv:2406.01168.

Inostroza, N. and Pavan, A. (2025). Adversarial coordination and public information design. *Theoretical Economics*, 20:763–813.

Jost, J. T., Barberá, P., Bonneau, R., Langer, M., Metzger, M., Nagler, J., Sterling, J., and Tucker, J. A. (2018). How social media facilitates political protest: Information, motivation, and social networks. *Political Psychology*, 39(S1):85–118.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

King, G., Pan, J., and Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107(2):326–343.

King, G., Pan, J., and Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, 111(3):484–501.

Kolotilin, A., Mylovanov, T., and Zapechelnyuk, A. (2022). Censorship as optimal persuasion. *Theoretical Economics*, 17:561–585.

Kuran, T. (1991). Now out of never: The element of surprise in the East European revolution of 1989. *World Politics*, 44(1):7–48.

Li, F., Song, Y., and Zhao, M. (2022). Global manipulation by local obfuscation. *Journal of Economic Theory*, 207.

Lorentzen, P. (2014). China's strategic censorship. *American Journal of Political Science*, 58(2):402–414.

Lorè, N. and Heydari, B. (2024). Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14:18490.

Mathevet, L., Perego, J., and Taneva, I. (2020). On information design in games. *Journal of Political Economy*, 128(4):1370–1404.

Morris, S. and Shin, H. S. (1998). Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review*, 88(3):587–597.

Morris, S. and Shin, H. S. (2002). Social value of public information. *American Economic Review*, 92(5):1521–1534.

Morris, S. and Shin, H. S. (2003). Global games: Theory and applications. In Dewatripont, M., Hansen, L. P., and Turnovsky, S. J., editors, *Advances in Economics and Econometrics*, pages 56–114. Cambridge University Press.

Obstfeld, M. (1996). Models of currency crises with self-fulfilling features. *European Economic Review*, 40(3-5):1037–1047.

Palatsi, A. C., Martin-Gutierrez, S., Cardenal, A. S., and Pellert, M. (2025). Large language models replicate and predict human cooperation across experiments in game theory. arXiv preprint arXiv:2511.04500.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (UIST).

Penney, J. W. (2016). Chilling effects: Online surveillance and Wikipedia use. *Berkeley Technology Law Journal*, 31(1):117–182.

Petrov, A. et al. (2025). LLM strategic reasoning: Agentic study through behavioral game theory. arXiv preprint arXiv:2502.20432.

Piatti, G. et al. (2024). Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. *Advances in Neural Information Processing Systems* (NeurIPS).

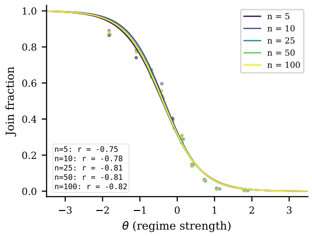
Shadmehr, M. and Bernhardt, D. (2015). State censorship. *American Economic Journal: Microeconomics*, 7(2):280–307.

Shurchkov, O. (2013). Coordination and learning in dynamic global games: Experimental evidence. *Experimental Economics*, 16(2):313–334.

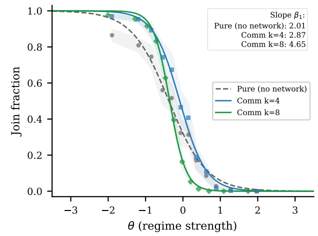
Stoycheff, E. (2016). Under surveillance: Examining Facebook's spiral of silence effects in the wake of NSA internet monitoring. *Journalism and Mass Communication Quarterly*, 93(2):296–311.

Sun, H. et al. (2025). Game theory meets large language models: A systematic survey with taxonomy and new frontiers. *Proceedings of the International Joint Conference on Artificial Intelligence* (IJCAI).

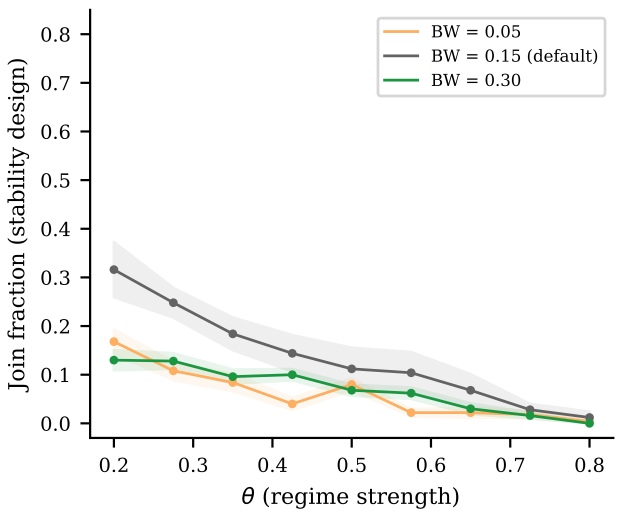
Szkup, M. and Trevino, I. (2020). Sentiments, strategic uncertainty, and information structures in coordination games. *Games and Economic Behavior*, 124:534–553.



(a) Agent count variation ($n \in \{5, 10, 25, 50, 100\}$).



(b) Network density ($k = 4$ vs. $k = 8$).



(c) Bandwidth sensitivity (0.05, 0.15, 0.30).

Figure A1: Robustness checks for equilibrium alignment and treatment effects.

A Robustness

These checks show that equilibrium alignment and the qualitative information design effects are stable to agent count, network density, and the proximity bandwidth.

Group-size awareness. In the main experiments, agents are told "You do not know how many others will JOIN" but are not told the group size, leaving them no basis for reasoning about coordination thresholds. As a robustness check, I run the pure and communication treatments with modified prompts that state "You are one of 25 citizens deciding whether to JOIN an uprising or STAY home." Over 100 country–periods per treatment, the pure join rate is 0.507 (vs. 0.369 baseline) and the communication join rate is 0.473 (vs. 0.452 baseline). Monotone response to signals is preserved in both treatments. The communication premium, however, reverses: with group-size knowledge, communication *lowers* join rates by 3.4 pp rather than raising them. One interpretation is that when agents know the group size, messages revealing others' reluctance become more informative about the probability of reaching critical mass, amplifying the deterrent effect of cautious peers. The level shift in the pure treatment suggests that group-size knowledge increases baseline willingness to coordinate, but the core finding—monotone signal response—is robust.

Online appendix. Additional supplemental material is in `online_appendix.tex`.