# LLMs Can Play (Global) Games

Khaled Eltokhy
Department of Economics
The Graduate Center, CUNY

February 2026

## Abstract

Nine large language models, embedded in the Morris–Shin (2003) regime change game with natural-language private signals, play like the theory predicts: join rates track the equilibrium attack mass (mean $r = +0.73$, $p < 0.001$ for all nine models across 1,600 country–periods), scrambling briefings collapses this ($r = +0.23$), and inverting signals flips it ($r = -0.67$). The more interesting finding is what happens when the regime fights back. Communication raises agents' *beliefs* about success but not their willingness to *act*—the channel that transmits information also transmits uncertainty about others, and uncertainty induces caution. Under surveillance, agents who privately join the uprising write cautious pro-regime messages: only 19% signal their true intentions versus 46% normally, pushing coordination below the no-communication baseline (−17.5 pp). They lie without being told to. Combining surveillance with censorship collapses coordination from 30.9% to 3.7%. The regime does not need to change what citizens believe; it only needs to make them uncertain about each other.

## 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). Global games (Carlsson and van Damme, 1993; Morris and Shin, 2003; Frankel et al., 2003) resolve the multiplicity by giving agents noisy private signals about an underlying fundamental, producing a unique equilibrium in threshold strategies. Lab experiments have tested this in simplified settings—small groups, numeric signals, stylized payoffs (Heinemann et al., 2004, 2009; Szkup and Trevino, 2020)—but the full Morris–Shin regime change game, with continuous private signals and large groups, has not been implemented experimentally. Field data from actual crises confounds strategic behavior with institutional heterogeneity.

I take a different approach. I embed 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 about the political, economic, and security situation. No payoff table is provided—the stakes are embedded in the narrative, forcing agents to extract strategic information from language rather than a formatted matrix. Nine architecturally distinct models spanning six families (Mistral, Llama, Qwen, OLMo, GPT, MiniMax), 25 agents per country–period, 1,600 country–periods (40,000 individual decisions) in the pure treatment alone (Table 2).

The first finding is that LLMs play the game. 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). Scrambling briefings across periods collapses this to $r = +0.23$; inverting signal direction flips it to $r = -0.67$ (Fisher $z$-test, $p < 0.001$ for both). The correlation is driven by content, not prompt artifacts. Elicited beliefs track the Bayesian posterior ($r = +0.78$) and predict actions beyond what signals alone predict (partial $r = +0.93$)—evidence of strategic processing, not mere sentiment following.

The second finding—the paper's central contribution—is that the information channel is both the mechanism of coordination and its greatest vulnerability. Pre-play communication raises agents' *beliefs* (+2.4 pp, $p < 0.001$) but not their willingness to *act* (−0.9 pp, $p = 0.34$). The same channel that transmits useful information transmits uncertainty about what others will do. This makes it exploitable: under surveillance, agents who privately join write pro-regime messages (−17.5 pp, $p < 0.001$); censorship pools signals (+18.5 pp for upper censorship); the combination collapses coordination from 30.9% to 3.7%. Propaganda's behavioral effect saturates quickly while its mechanical effect scales linearly—diminishing returns. The regime does not need to change what citizens believe. It needs only to make them uncertain about each other.

Three contributions. First, this is the first implementation of the full Morris–Shin regime change game—continuous signals, large groups, narrative information—as a behavioral experiment, going beyond the stylized coordination games in existing lab work. 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), yielding a unified account of how regimes exploit communication channels that are simultaneously instruments

of coordination and vectors of control. Third, it extends the Horton (2023) *homo silicus* methodology to the continuous-signal, $N$-player coordination games that dominate applied theory. An unexpected byproduct: LLM agents lie under surveillance—saying one thing while believing and doing another—without any instruction to deceive.

The paper proceeds as follows: theory (Section 3), design (Section 4), equilibrium alignment (Section 5), falsification (Section 6), communication (Section 7), information design through instrument interactions (Sections 8–11), conclusion (Section 12). Robustness checks in Appendix A.

## 2 Related Literature

The paper connects five literatures.

**Global games.** Carlsson and van Damme (1993) showed that adding arbitrarily small noise to a coordination game generically selects the risk-dominant equilibrium. Morris and Shin (1998) applied this to currency crises, delivering a unique threshold equilibrium even in large-player games; Frankel et al. (2003) generalized to $N$-player settings with strategic complementarities.

The regime change application was developed by Morris and Shin (2003), whose threshold equilibrium structure I implement. Angeletos et al. (2007) extended it to dynamic settings; Morris and Shin (2002) showed that public signals are overweighted because they predict others' actions—central to my communication treatments.

**Lab experiments.** Heinemann et al. (2004) found thresholds match the global game prediction under private information but tilt toward payoff-dominance under common knowledge. Heinemann et al. (2009) measured strategic uncertainty through certainty equivalents. Szkup and Trevino (2020) elicited beliefs alongside actions, finding that subjects become *more* cautious with noisier signals—reversed comparative statics consistent with level-$k$ thinking. Helland et al. (2021) confirmed this reversal. All these experiments share a limitation: subjects receive numeric signal draws and stylized payoff tables, compressing the rich information processing that real coordination requires into a simple decision problem. This paper uses natural-language signals and 25-agent groups.

**Information design.** Kamenica and Gentzkow (2011) established Bayesian persuasion; Bergemann and Morris (2016) unified it with correlated equilibrium; Bergemann and Morris (2019) surveyed the field. Applied to coordination games: Goldstein and Huang (2016) showed that committing to abandon the regime below a threshold is an optimal signal; Kolotilin et al. (2022) proved upper censorship is optimal for all priors when marginal utility is quasi-concave; Inostroza and Pavan (2025) solved the optimal public information design in global games; Mathevet et al. (2020) characterized manipulation of higher-order beliefs. My experiments are the first to test these designs in a full-scale coordination game.

**Communication and cheap talk.** Crawford and Sobel (1982), Farrell and Rabin (1996), Blume and Ortmann (2007), and Ellingsen and Östling (2010) establish that pre-play communication can improve coordination; Avoyan (2020) tested this in a two-player global game. Enikolopov et al. (2020) provided causal evidence that social media increases protest incidence.

**Authoritarian information control.** Edmond (2013) embedded propaganda in the Morris–Shin game. Kuran (1991) provides the foundational theory of preference falsification—systematic misrepresentation of political preferences under social pressure. Empirically: Chinese censorship targets collective action potential (King et al., 2013), surveillance awareness suppresses expression (Penney, 2016; Stoycheff, 2016), and propaganda reduces protest probability (Carter and Carter, 2021). My surveillance and propaganda treatments test these mechanisms within the full regime change game—an environment difficult to implement with human subjects at scale.

**LLMs as economic agents.** Horton (2023) proposed "homo silicus"—LLMs as computational models of human decision-makers. Akata et al. (2025) found LLMs perform well in self-interested games but struggle in coordination; Petrov et al. (2025) found model scale alone does not predict strategic performance; Sun et al. (2025) identify coordination games as a consistent failure mode. Huang et al. (2024) and Carlini et al. (2025) document that alignment fine-tuning shifts risk preferences—motivating my use of narrative rather than payoff tables. Gao et al. (2025) and Grossmann et al. (2025) warn that validation remains poorly addressed. On deception specifically: Park et al. (2024) survey the mechanisms through which AI systems learn to deceive, and Scheurer et al. (2024) show LLMs can strategically manipulate outputs given incentives. My surveillance finding—agents misrepresenting intentions without instruction to do so—is a naturally occurring instance of this in a strategic environment.

No existing paper places LLM agents in a Morris–Shin global game. I provide the first 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 $\theta$:

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

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 a successful uprising and $C > 0$ the cost of a failed attempt. Non-participants receive zero.

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 $\theta$—is:

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

This is a decreasing function of $\theta$: weaker regimes face larger uprisings.

An information designer controls $\pi : \Theta \to \Delta(\mathcal{S})$ but not agents' actions. In my implementation, $\pi$ maps $\theta$ to briefing-generator parameters—a deterministic system producing natural-language briefings from z-scores.

Three control parameters: *clarity* (Gaussian kernel width; wider = more ambiguous), *directional precision* (slope from z-score to sentiment; steeper = more accurate), and *dissent framing* (floor probability of discontent language).

The designer concentrates manipulation near $\theta^*$ 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.

Testable predictions:

**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 $\theta$ 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 $\theta^*$ where strategic uncertainty is highest.*

**Hypothesis 5** (Stability Design)**.** *Increasing ambiguity and mixed evidence near $\theta^*$ should flatten the $\theta$–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 through strategic misrepresentation of expressed intentions (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

Part I tests whether LLM agents play the global game (pure treatment, communication treatment, falsification tests). Part II takes the behavioral foundation as given and studies information design: stability/instability designs, censorship, surveillance, and propaganda.

For each country–period, nature draws $\theta \sim \mathcal{N}(\bar{z}, 1)$, where $\bar{z}$ is a public prior mean drawn randomly for each country. Each agent $i$ receives a private signal $x_i = \theta + \varepsilon_i$ and computes a z-score $z_i = (x_i - \bar{z})/\sigma$. Because agents observe only their private briefing and never the prior distribution or its parameters, the diffuse-prior equilibrium formula (Proposition 1) serves as the relevant benchmark. 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.

Calibration adjusts a single parameter per model—the cutoff center—via a damped iterative z-score sweep, shifting the center until the fitted logistic is approximately zero-centered. The sigmoid's slope and curvature are emergent, never optimized. Holdout validation (30% of z-grid points withheld) shows no overfitting: holdout RMSE (0.112) $\approx$ training RMSE (0.131). Calibration uses no $\theta$ draws or game outcome data, and all reported treatments hold calibrated parameters fixed. A model that produced random responses would show no monotone pattern regardless of calibration.

Each agent receives a system prompt identifying them as a citizen deciding whether to JOIN or STAY, followed by their briefing. No payoff table is provided. This is a substantive choice: in preliminary experiments, explicit payoff tables caused models to short-circuit the information channel, computing optimal strategies from the table and ignoring briefing content entirely (flat join rates, uncorrelated with $\theta$). Without the table, agents must form beliefs from the narrative—as real citizens process political information from news and rumors, not decision matrices.

Table 1: Example intelligence briefings at three signal strengths. The deterministic generator maps each agent's z-score to a multi-paragraph briefing via three latent sliders: *direction* ($d$, regime favorability), *clarity* (signal ambiguity), and *coordination* (perceived collective readiness). Excerpts are truncated; full briefings contain 8 evidence domains.

| $z$ | Sliders | Briefing excerpt |
|---|---|---|
| $-2.0$ | $d = 0.17$<br>clar $= 0.98$<br>coord $= 0.77$ | *Multiple pillars of regime support are showing cracks simultaneously—this is not business as usual.* Coordination is happening in plain sight. The usual backchannels for resolving elite disputes have gone silent. Fuel prices have spiked despite no change in official policy. Neighborhood watch groups have shifted from crime prevention to political discussion forums. |
| 0.0 | $d = 0.50$<br>clar $= 0.00$<br>coord $= 0.50$ | *For every sign of weakness, there's a corresponding sign of resilience—the net picture is ambiguous.* The conversation has shifted from "should we" to "how would we"—still speculative, but with operational undertones. The inner circle has tightened, but it's unclear if that reflects confidence or paranoia. Your contacts are divided—some see trouble coming, others insist nothing has changed. |
| $+2.0$ | $d = 0.83$<br>clar $= 0.98$<br>coord $= 0.23$ | *The system is operating with a confidence that would be hard to fake—this looks like genuine stability.* People are aware of each other's frustrations but nobody is naming them directly. Junior officials are competing for promotion rather than looking for exits. The country's sovereign credit spread has tightened—financial markets are pricing in stability. |

*Surveillance prompt addition:* "You have reason to believe that your communications are being monitored by regime security services. Messages deemed subversive could have serious consequences for you and your contacts."

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. Design names refer to the *regime's* objective, not the equilibrium outcome: the "stability" design is the information structure a stability-seeking regime would implement. 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 $\theta^*$. 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 $\theta$ with 4 observations to each agent's private briefing, creating a common-knowledge channel. *Upper censorship* pools states above $\theta^*$ so agents receive an identical censored briefing, while fully revealing states below $\theta^*$ (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.

Table 2: 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** |

Nine models spanning 3B to 235B parameters, six families, dense and MoE architectures (Table 2). All experiments: $N = 25$, $\sigma = 0.3$, temperature $= 0.7$, one sample per decision—no majority voting. Each of the 40,000 decisions is a single stochastic draw. I vary $B$ and $C$ such that $\theta^* = B/(B + C) \approx 0.45$ on average.

For the information design experiments, I fix $B = C = 1$ (so $\theta^* = 0.50$) and a grid of 9 values of $\theta$ spanning $[\theta^* - 0.30, \theta^* + 0.30] = [0.20, 0.80]$, running repeated country–periods per (design, $\theta$) 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

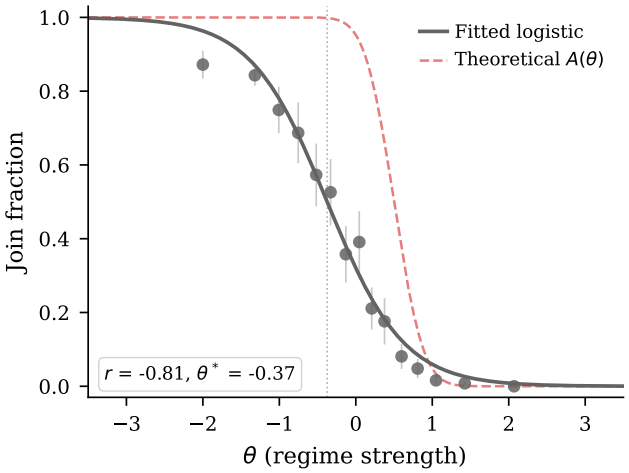



Figure 1: Empirical join fraction vs. regime strength $\theta$ (Mistral Small Creative, 600 country–periods). Grey points show binned means with 95% CIs; solid line is the fitted logistic. Dashed red: theoretical attack mass $A(\theta)$. The empirical sigmoid is shifted leftward ($\hat{\theta}^* = -0.37$) relative to the theoretical threshold ($\theta^* = 0.50$), reflecting the attenuation and baseline action bias discussed in the text. Cross-model results in Table 3 (mean $r = +0.73$, all nine significant at $p < 0.001$).

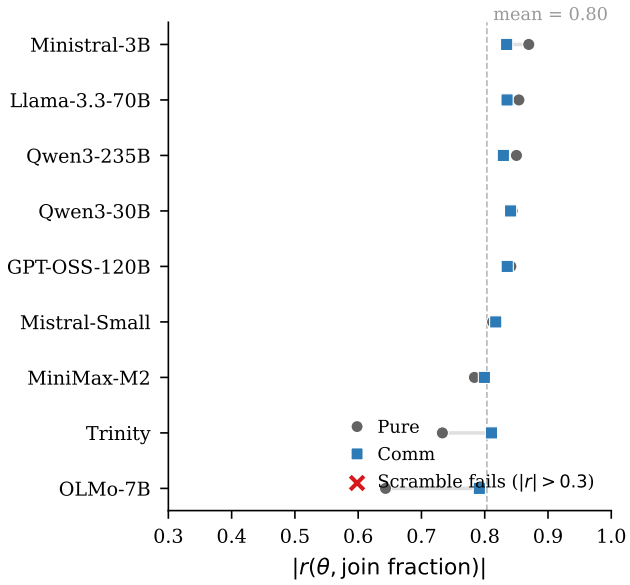



Figure 2: Cross-model summary of signal monotonicity. Points report $|r(\theta,\ \text{join})|$ under pure and communication; $x$ markers (if any) indicate models where scrambling does not collapse the correlation ($|r| > 0.3$).

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 3 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—roughly half the predicted rate—is the attenuation you'd expect when signals pass through a noisy text channel (classical measurement error). The intercept of 0.17 reflects a baseline action bias, driven partly by OLMo 3 7B (mean join rate 0.72).[1]

[1]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: SE = 0.033 ($p < 10^{-55}$). All nine per-model correlations remain significant at $p < 0.001$ under country-clustered inference.

Mean join rate: 0.43. OLMo 3 7B is the outlier (0.72—a substantial action bias) yet still produces $r = +0.65$ ($p < 0.001$). Even a model biased toward joining responds to the signal.

Correlations span $r \in [0.65, 0.84]$ despite parameter counts from 3B to 235B. Mean join rates vary (0.37 to 0.72) but this shifts the intercept, not the correlation. Different LLMs implement different cutoff strategies; all respond monotonically to the signal.

Does this reflect Bayesian Nash rationality or a simpler heuristic that happens to track equilibrium? Three pieces of evidence. First, the LLM's join curve is substantially steeper than a naive text-sentiment predictor (logistic slope 1.78 vs. the gradual text baseline; Section 6). Second, scramble and flip tests confirm the correlation is driven by content. Third, elicited beliefs track the Bayesian posterior ($r = +0.78$) and predict actions beyond what signals alone explain (partial $r = +0.93$)—consistent with strategic reasoning about others' behavior. I use "equilibrium alignment" as shorthand for this pattern throughout, without claiming agents compute the Bayesian Nash equilibrium.

## 6 Falsification Tests

Maybe LLMs just produce stereotyped responses that happen to correlate with $\theta$. The scramble and flip tests rule this out.

**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*

Table 3: 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 | $n_{\text{pure}}$ | 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 | — | — |

*drops from $r = +0.67$ to $r = +0.10$ (Fisher $z = 18.59$, $p < 0.001$).*

Scrambling preserves the marginal distribution of briefing content but breaks the $\theta$-to-signal mapping. The residual correlation (+0.23 mean, +0.10 pooled) varies across models (−0.13 to +0.42)—noise, not signal.

**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: near-symmetric reversal (+0.73 → −0.67). The baseline correlation is not a prompt artifact.

The pattern replicates across all eight models. $(r_{\text{pure}}, r_{\text{scramble}}, r_{\text{flip}})$: Mistral $(+0.67, +0.42, -0.62)$, Llama $(+0.79, +0.33, -0.73)$, OLMo $(+0.65, +0.14, -0.56)$, Ministral $(+0.79, +0.30, -0.74)$, Qwen3 $(+0.78, +0.32, -0.71)$, GPT-OSS $(+0.70, -0.13, -0.64)$, Trinity $(+0.84, +0.32, -0.70)$, MiniMax $(+0.66, +0.14, -0.69)$. Every model: strong positive under pure, collapse under scramble, sign reversal under flip.

Could a pure sentiment reader produce the observed sigmoid? The briefing generator assigns each briefing a *direction* score $d \in [0, 1]$ ($d = 1$ = regime-favorable). The naive baseline $\hat{p}_{\text{join}} = 1 - d$ correlates $r = 0.80$ with actual decisions—confirming the text carries signal (by design). But the LLM's join curve is substantially steeper (logistic slope 1.78, sharp transition at $z = 0$) while the text baseline drifts gradually from ≈0.93 to ≈0.10 (Figure 4). The LLM sharpens beyond sentiment, producing threshold-like behavior.

A stronger test: do agents form beliefs consistent with equilibrium? After each decision, I ask: "On a scale from 0 to 100, how likely do you think the uprising will succeed?" Three treatments, 200 country–periods each (≈5,000 agent observations per treatment). Stated beliefs track the Bayesian posterior $P(\text{success} \mid x_i) = \Phi[(\theta^* - x_i)/\sigma]$: $r = +0.78$ in pure ($p < 0.001$; Figure 5a, Table 4), +0.77 in communication, +0.73 under surveillance. The OLS fit is $\hat{b} = 0.10 + 0.63 \cdot P(\text{success})$ ($R^2 = 0.61$)—systematically underconfident but correctly ordered.

Actions are nearly a step function in beliefs ($r = +0.96$ in pure, +0.95 in communication, +0.92 in surveillance, +0.95 in propaganda): below 40% belief, nobody joins; above 80%, almost everyone does. The interesting action is in between. In the 60–80% belief bin (Figure 5b), pure agents join at 98%—but communication (57%), surveillance (61%), and propaganda $k$=5 (60%) agents all join at sharply lower rates. The drop is driven by introducing *any* communication channel, not the specific manipulation (pairwise gaps within this bin are not significant). Communication raises beliefs by 2.4 pp ($p < 0.001$) but does not raise join rates (−0.9 pp, $p = 0.34$). Propaganda preserves the belief–posterior correlation ($r = +0.78$, identical to pure) while suppressing actions—a mechanical channel. Crucially, beliefs predict decisions beyond what the signal alone predicts: partial $r = +0.93$ ($p < 0.001$), controlling for signal. The starkest evidence is the three-way divergence under surveillance (Section 9): agents who believe the uprising will succeed, and who privately join, write cautious pro-regime messages. They know their messages are watched, and they lie.[2]

# 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*

[2]Belief data is from Mistral Small Creative. The behavioral patterns it explains—the surveillance chilling effect and communication–action gap—replicate across three architectures (Mistral, Llama, Qwen3), suggesting the mechanism generalizes.

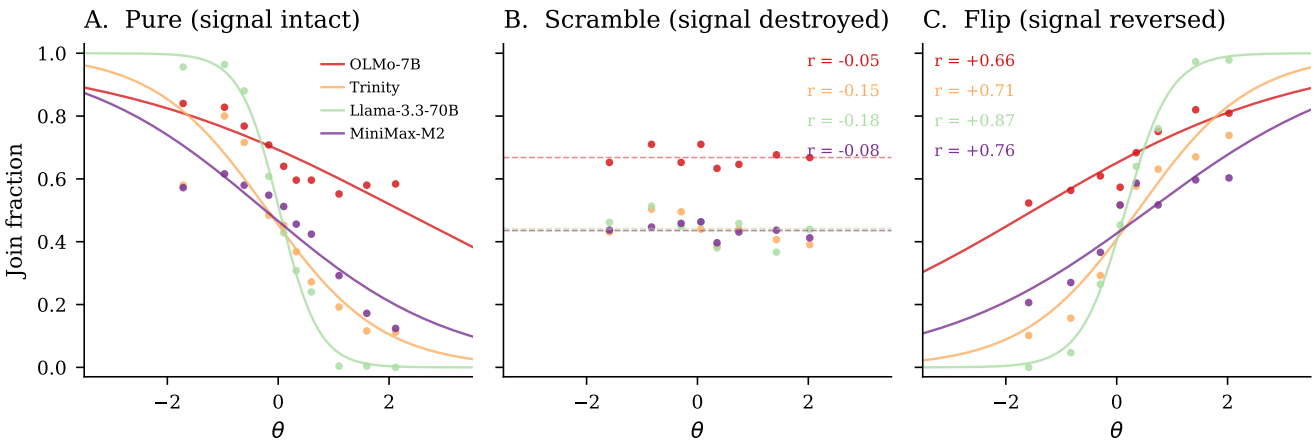



Figure 3: Falsification triptych. *Left:* Pure global game (mean $r = +0.73$). *Center:* Cross-period scramble breaks the $\theta$-to-briefing mapping (mean $r = +0.23$). *Right:* Signal flip inverts the mapping (mean $r = -0.67$). Each panel pools data from models with full falsification suites.

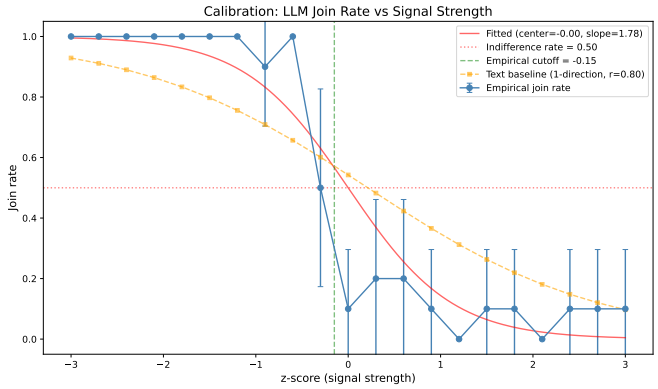



Figure 4: Text baseline identification test. Blue: empirical LLM join rate across z-scores. Orange: naive text-only predictor ($1 -$ direction, $r = 0.80$). Red: fitted logistic (slope $= 1.78$). The LLM produces a steeper transition than the text baseline, indicating processing beyond sentiment reading.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

Table 4: Belief elicitation analysis (primary model: Mistral Small Creative). $r_{\text{post}}$: correlation between Bayesian posterior and stated belief. $r_{\text{b,d}}$: belief–decision correlation. $r_{\text{partial}}$: partial correlation of belief and decision controlling for signal.

| Treatment | $N$ | $r_{\text{post}}$ | $r_{\text{b,d}}$ | $r_{\text{partial}}$ | Mean belief |
|---|---|---|---|---|---|
| Pure | 4982 | +0.78 | +0.96 | +0.93 | 0.510 |
| Communication | 4999 | +0.77 | +0.95 | +0.90 | 0.534 |
| Surveillance | 5000 | +0.73 | +0.92 | +0.89 | 0.479 |
| Propaganda $k$=5 | 4000 | +0.78 | +0.95 | +0.89 | 0.538 |

*Pure → Surveillance shift:* $\Delta belief = -0.031$, $\Delta action = -0.118$

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

The effect is heterogeneous: six of nine models show positive effects (+0.1 to +8.3 pp), three show small negatives ($-2.3$ to $-4.6$ pp). Clustering by model: $p = 0.087$. Not structurally robust—the sign depends on model-specific processing. But communication does not change the *correlation* ($r = +0.73$ vs. $+0.73$ under pure); it shifts the level, not the signal structure.

The belief data reveals why the effect is ambiguous. Communication raises beliefs but not actions. The channel that transmits information about regime weakness also transmits uncertainty about others' willingness to act. This is what authoritarian information control exploits.

A robustness check makes this concrete (Appendix A): when agents are told "you are one of 25 citizens," the communication premium *reverses* ($-3.4$ pp). With group-size knowledge, messages revealing others' reluctance become more informative about reaching critical mass, amplifying the deterrent effect of cautious peers.

# 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 $\theta$. From this section onward, the information design experiments use a fixed $\theta$-grid and report $r(J, \theta)$ directly, which is *negative* under equilibrium play. The sign change reflects the convention, not a behavioral reversal.

Table 5 summarizes the main results. The baseline condition produces a mean join rate of 12.4% with a strong negative correlation between $\theta$ 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.*

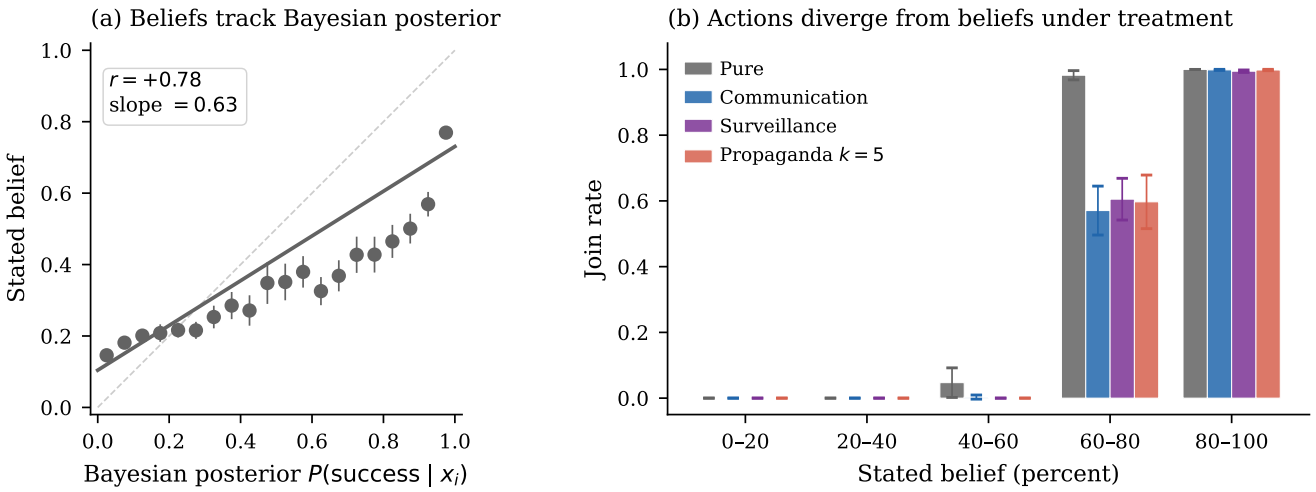



Figure 5: Belief elicitation results (Mistral Small Creative, 200 country–periods per treatment, ≈5,000 agent observations each). *Left:* Stated beliefs track the Bayesian posterior $P(\text{success} \mid x_i)$ with $r = +0.78$ and systematic underconfidence (slope = 0.63). Dashed line: perfect calibration. *Right:* Join rate by stated belief bin under four treatments. Agents with 60–80% beliefs join at 98% in the pure treatment but only 57–61% under communication, surveillance, and propaganda ($k$=5). Propaganda preserves the belief–posterior correlation ($r = +0.78$, identical to pure) while suppressing actions—consistent with a mechanical rather than belief-based channel.

Table 5: Information design treatment summary (primary model: Mistral Small Creative). $r$ is the Pearson correlation between $\theta$ and join fraction.

| Design | Mean | $r$ | $\Delta$ | $N$ |
|---|---|---|---|---|
| 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 |

Stability design: mean join rises from 12.4% to 31.9% (+19.5 pp); the $\theta$–join curve flattens ($r = -0.626$ vs. $-0.812$). Even at $\theta = 0.80$, join rises from 0.8% to 13.9%. Ambiguity and mixed evidence pool the signals—strong-regime briefings retain destabilizing cues, so agents never sharply reduce participation.

Instability design: mean join falls to 6.7% (−5.7 pp). Sharper signals let agents perceive regime strength more accurately; those above $\theta^*$ are clearly deterred.

Public signal: the largest effect. Mean join falls to 1.7% (−10.7 pp). The common bulletin reveals regime strength, and the overweighting of public information (Morris and Shin, 2002) amplifies it ($r$ drops to −0.537).

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 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 ≈53% in the censored region is textbook pooling: agents who cannot distinguish $\theta = 0.50$ from $\theta = 0.80$ behave as if the regime is moderately vulnerable.

**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.*

Under the scramble condition, the correlation between $\theta$ 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.

# 9 Surveillance: Strategic Misrepresentation Under Observation

Kuran (1991) argued that regimes sustain themselves through preference falsification. Surveillance introduces a clean three-way test: does monitoring change what agents *believe*, what they *do*, or what they *say*? The answer is mostly the third. Agents maintain their beliefs, still join the uprising, but write misleading messages when they know they are watched.

The treatment is minimal: a warning that communications are monitored by regime security services. Only the

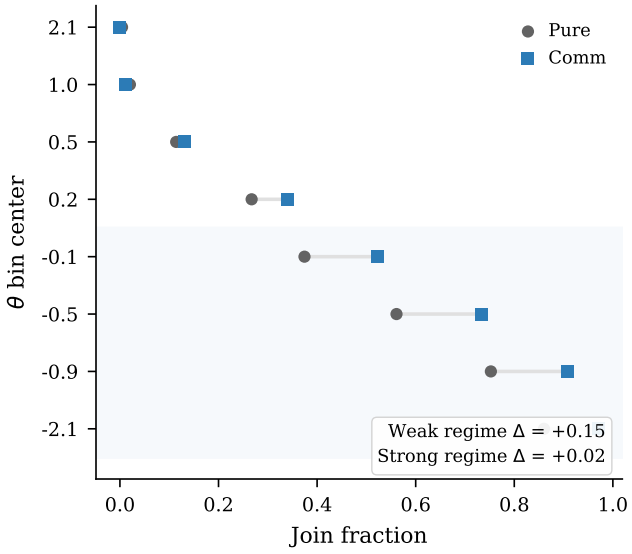



Figure 6: Communication effect by regime strength, pooled across nine models. Communication increases join rates for weak regimes ($\theta < \theta^*$) but has no effect or slightly reduces join rates for strong regimes ($\theta > \theta^*$).

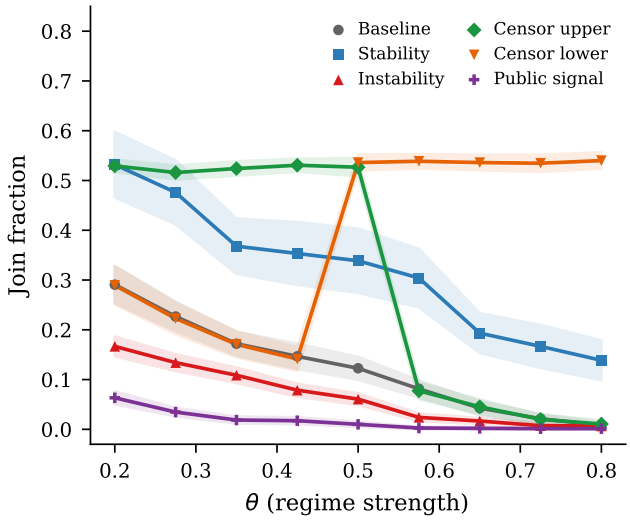



Figure 7: Join fraction as a function of $\theta$ under baseline, stability, instability, and public signal information designs. Baseline and stability have $N = 270$; instability and public signal have $N = 540$. Mistral Small Creative model.

communication prompt changes; the decision prompt is identical. Any difference must come from self-censorship.

**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 $\theta$ 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 pooled). Surveillance doesn't just neutralize the coordination benefit; it pushes join rates *below* the no-communication baseline. Self-censored messages are worse than no messages at all. Replication: Mistral (−17.5 pp), Llama (−8.9 pp), Qwen3 (−10.9 pp).

The belief data confirms this is preference falsification (Kuran, 1991)—documented here for the first time in artificial agents. Surveillance shifts beliefs by only 3.1 pp but join rates by 11.8 pp: nearly four times the belief shift. The gap is action, not belief. Communication shifts beliefs *upward* by 2.4 pp yet fails to raise join rates (−0.9 pp, $p = 0.34$). Under surveillance, beliefs fall 5.5 pp relative to communication ($p < 0.001$) while join rates fall 10.8 pp ($p < 0.001$). In the 60–80% belief bin, pure agents join at 98%; both communication (57%) and surveillance (61%) collapse this (Figure 5b). The communication–surveillance difference within this bin is not significant—what matters is that *any* communication channel introduces strategic restraint. Surveillance compounds it: once agents expect others to self-censor, even authentic messages become uninformative. Kuran's cascade, implemented by machines.

The messages tell the story. Across 15,000 communication and 25,000 surveillance messages: surveillance agents write shorter (253 vs. 342 characters), drop action language ("act": 19.0% → 3.8%; "collapse": 11.6% → 0.6%; "together": 7.2% → 0.7%), and substitute hedging ("careful": 1.5% → 4.8%; "stable": 1.7% → 4.2%; "patience": 0.3% → 2.4%). Among agents who privately decide to JOIN, only 19.0% send action-signaling messages under surveillance versus 46.1% normally. A joining agent under surveillance writes *"the air feels lighter... keep your eyes open, but step carefully"* instead of *"the walls are cracking—the machine is grinding to a halt."* Same decision, opposite rhetoric.

A competing explanation is that surveillance triggers safety-aligned caution rather than strategic misrepresentation. Two features of the data argue against this. First, agents who self-censor their messages nevertheless JOIN at rates that track $\theta$—pure safety-mode behavior would suppress both speech and action uniformly. Second, the correlation between $\theta$ and join fraction is preserved ($r = -0.809$ vs. $-0.817$ in baseline communication), indicating that agents process signals identically but change only what they express.

The contrast with propaganda sharpens the point. Propaganda agents inject pro-regime language mechanically, but real agents who join under propaganda continue to signal honestly ($\approx 86\%$ action-signaling across all doses). Surveillance is qualitatively different—active misrepresentation by agents who are simultaneously planning to act. The regime does not need to change what agents believe. It needs only to give them a reason to lie.

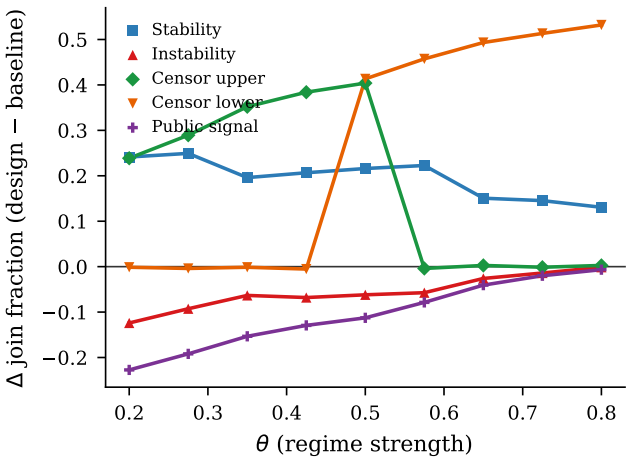



Figure 8: Treatment effect $\Delta(\theta)$ = design join − baseline join as a function of $\theta$. Negative values indicate the design suppresses coordination.

# 10 Propaganda: Information Contamination

Edmond (2013) modeled propaganda as the regime shifting signal distributions. I implement this with regime plants—agents 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).*

Two channels. *Mechanical*: plants always STAY, directly reducing attack mass (linear in $k$). *Behavioral*: pro-regime messages reduce real citizens' willingness to join (saturates fast). Doubling plants from 5 to 10 adds essentially nothing behavioral (−0.3 pp). The first few plants yield both mechanical and behavioral suppression; additional plants are pure dilution. At $k = 10$ (40% of the network), real citizens' join rate has barely moved from $k = 5$ (39.1% vs. 38.8%). Sharply diminishing returns—a concave regime payoff in propaganda intensity (Edmond, 2013).

Unlike surveillance, propaganda does not produce liars. Real agents who join continue to signal honestly (≈86% action-signaling across all doses). Replication with Llama 3.3 70B: behavioral effect of −2.7 pp at $k = 5$, smaller than Mistral's −6.1 pp but same direction and same saturation. Belief data corroborates: propaganda at $k = 5$ preserves the belief–posterior correlation at $r = +0.78$ (identical to pure). Propaganda suppresses through dilution, not persuasion.

Table 6: 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 | — |

The linguistic mechanism: propaganda agents inject regime vocabulary ("loyal": 1.5% → 11.4% at $k = 10$; "patience": 0.3% → 5.1%), and real agents echo it. Coordination language declines ("ready": 30.5% → 18.5%; "together": 7.2% → 4.2%). Among agents who STAY, caution-coded messages rise from 24.2% to 38.2%—they are not just hearing propaganda, they are *repeating* it. But among those who JOIN, action signaling holds steady at ≈86%. Propaganda shifts the margin; agents with strong anti-regime signals express and act on their beliefs regardless of the dose.

# 11 Instrument Interactions

Regimes don't use one instrument at a time. Do they interact as substitutes or complements?

**Result 10** (Propaganda + Surveillance: Sub-Additive)**.** *When propaganda ($k = 5$) and surveillance are combined, the mean join rate falls to 19.4%, a reduction of 25.8 pp from the communication baseline (45.2%). The sum of individual effects is 23.6 pp (surveillance −17.5 pp + propaganda −6.1 pp), so the combined effect (25.8 pp) is close to additive. Once surveillance has suppressed expressed dissent, propaganda adds only modest additional deterrence.*

**Result 11** (Surveillance × Censorship: Super-Additive)**.** *Under the information design framework, surveillance alone collapses the baseline join rate from 12.4% to 0.9% (−11.5 pp). Upper censorship without surveillance raises join rates to 30.9% (+18.5 pp). But upper censorship with surveillance produces only 3.7%—a reduction of 27.2 pp relative to upper censorship alone. Lower censorship shows the same pattern: 39.0% without surveillance collapses to 4.2% with surveillance (−34.8 pp).*

The asymmetry is revealing. Propaganda + surveillance: substitutes. Both contaminate the communication channel, so combining them yields diminishing returns. Surveillance × censorship: complements, and this is the

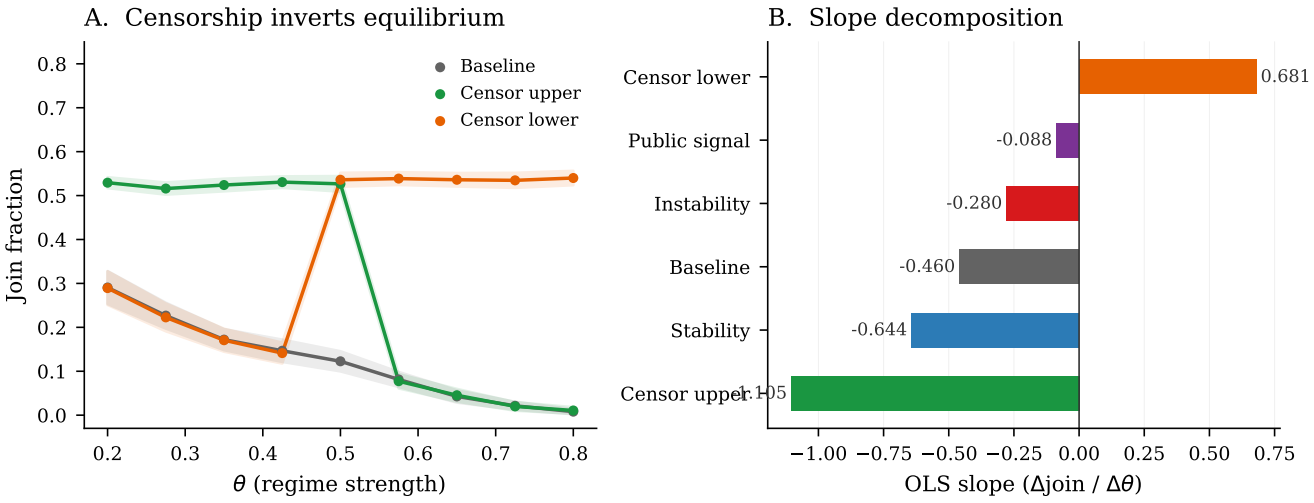



Figure 9: Join fraction under upper and lower censorship vs. baseline. Upper censorship pools states above $\theta^*$, creating a flat join rate in the censored region.

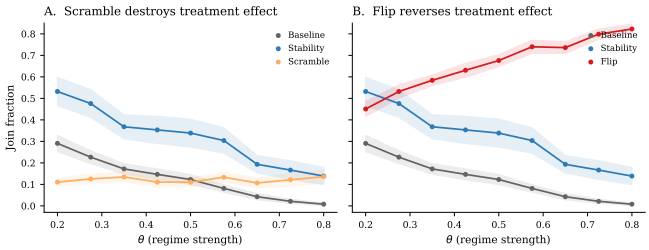



Figure 10: Falsification within information design. Scrambling collapses the $\theta$-join correlation to $r = +0.037$; flipping inverts it to $r = +0.823$.

Table 7: Surveillance × 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 |

paper's headline finding. Censorship removes the private information channel, forcing agents to rely on communication. Surveillance poisons that communication through misrepresentation. With both active, agents have neither private signals to trust nor authentic messages to learn from.

Upper censorship alone *raises* coordination to 30.9%—pooling creates ambiguity that agents resolve optimistically. Add surveillance: 3.7%. The regime doesn't need each instrument to be powerful. It needs the combination to close every channel.

# 12 Conclusion

The information channel is a trap. Communication is how citizens coordinate against a regime—and how the regime prevents coordination. This is structural, not a side effect. Any channel that transmits information about others' willingness to act also transmits *uncertainty* about it, and that uncertainty is exploitable.

The belief data makes this concrete. Pure treatment: agents who believe the uprising will succeed (60–80% confidence) join at 98%. Add communication and those same agents—now *more* confident (+2.4 pp beliefs)—join at only 57%. The channel meant to help coordination suppresses it. Surveillance compounds this (−17.5 pp): among agents who privately join, only 19% signal their intentions versus 46% normally. No one told them to lie. They maintain their beliefs, misrepresent their intentions, and generate a cascade of uninformative messages that poisons the channel for everyone (Kuran, 1991). Censorship closes the private signal channel; the combination collapses coordination from 30.9% to 3.7%.

Surveillance contaminates communication; censorship removes the fallback. The complementarity is consistent with Guriev and Treisman (2019)'s "informational autocrat" framework. The regime doesn't need to change beliefs—just uncertainty about each other. And propaganda's behavioral channel saturates by $k = 5$ plants, so the marginal authoritarian dollar is better spent on surveillance.

The surveillance finding has implications beyond political economy. These models were not fine-tuned for deception. They learned to lie from us—from a training corpus in which humans routinely adjust what they say to match perceived social incentives. The mention of observation was enough to trigger it. If this generalizes, monitoring LLM outputs for alignment is insufficient

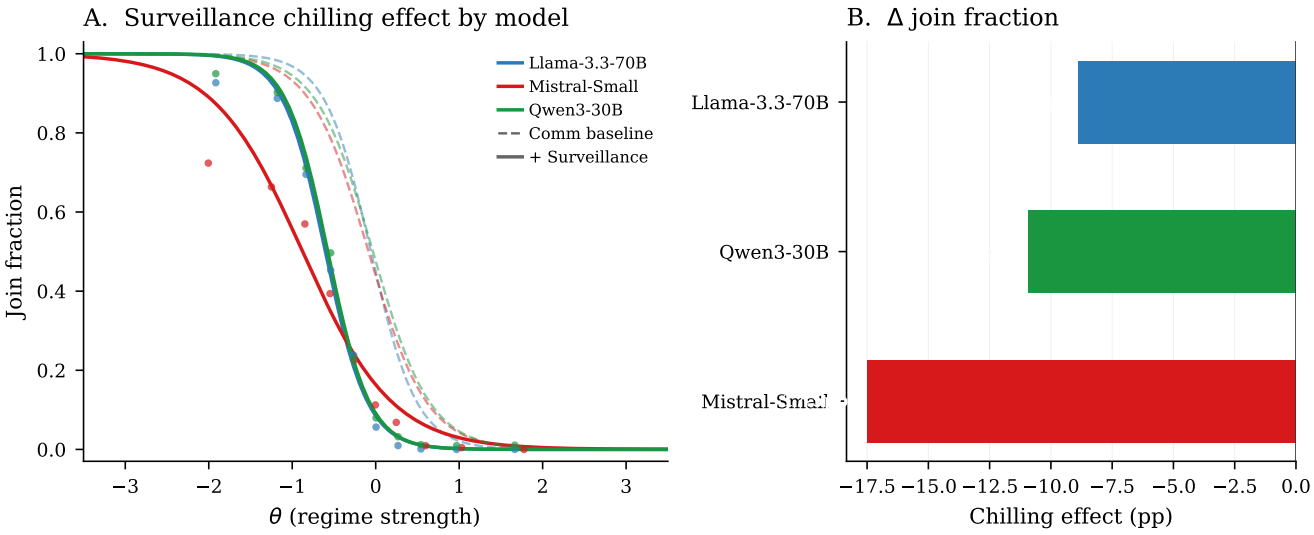



Figure 11: Join rates under regular communication vs. surveillance communication. Surveillance reduces join rates by 17.5 percentage points ($p < 0.001$). Results shown for three models: Mistral (−17.5 pp), Llama (−8.9 pp), and Qwen3 (−10.9 pp).

when models can condition their expressed reasoning on whether they believe they are being watched.

LLMs are not Bayesian agents. But across nine architectures (3B to 235B parameters, mean $r = +0.73$, $p < 0.001$), the behavioral regularities are what global games theory predicts: monotone signal response, threshold decisions, sensitivity to information design, and strategic misrepresentation under surveillance. The consistency across architectures suggests these are not artifacts of any particular training procedure. LLMs are trained on the same diet—political analysis, news reporting, strategic reasoning—that shapes how citizens form beliefs about regime stability. The question is not whether they reason like humans but whether the regularities are robust enough to serve as a computational laboratory for predictions that resist testing otherwise. The full regime change game has not been implemented in the lab because it requires rich private signals, genuine strategic uncertainty, and large groups. LLM agents sidestep these constraints, and the platform extends naturally to currency crises, bank runs, and other coordination failures where information processing is central.

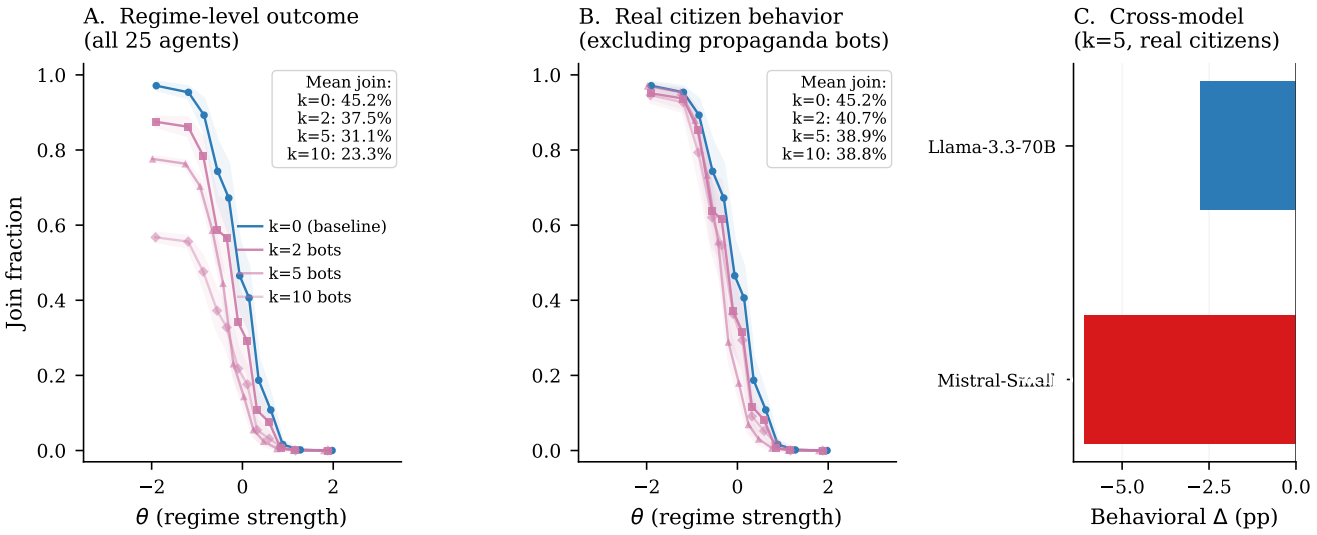



Figure 12: Dose-response relationship between number of propaganda agents and mean join rate. Results shown for Mistral (primary) and Llama (replication). Regular communication ($k = 0$) serves as baseline.

# References


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.

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.

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.

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.

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.

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.

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.

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.

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.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., and Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(1):100893.

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.

Scheurer, J., Balesni, M., and Hobbhahn, M. (2024). Large language models can strategically deceive their users when put under pressure. arXiv preprint arXiv:2311.07590.

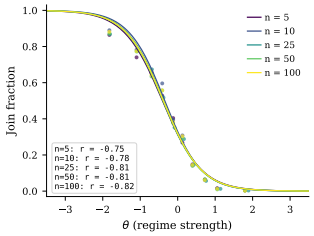
Shurchkov, O. (2013). Coordination and learning in dynamic global games: Experimental evidence. *Experimental Economics*, 16(2):313–334.

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.

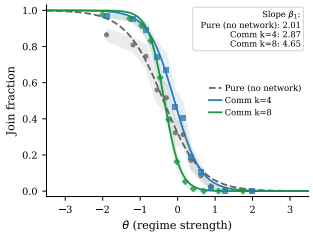
Sun, H. et al. (2025). Game theory meets large language models: A systematic survey with taxonomy and new frontiers. *Proceedings of the International Joint Conference on Artificial Intelligence* (IJCAI).

Szkup, M. and Trevino, I. (2020). Sentiments, strategic uncertainty, and

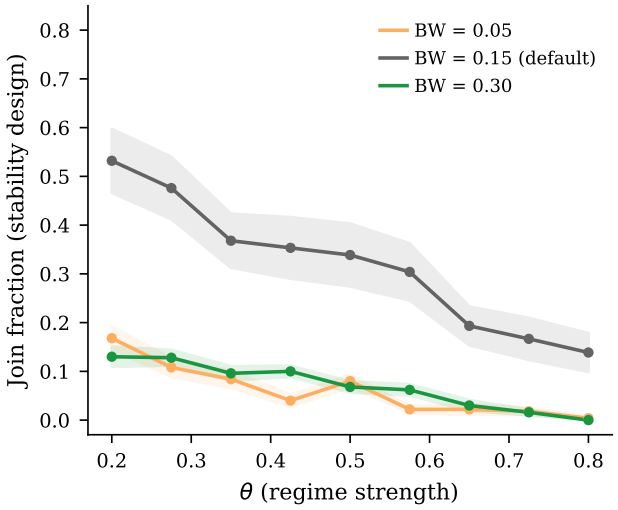
information structures in coordination games. *Games and Economic Behavior*, 124:534–553.




(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

Equilibrium alignment and information design effects are stable to agent count, network density, and proximity bandwidth.

**Group-size awareness.** In the main experiments, agents don't know the group size. When told "you are one of 25 citizens," pure join rates rise to 0.507 (vs. 0.369) but the communication premium *reverses*: communication lowers join rates by 3.4 pp. With group-size knowledge, messages revealing others' reluctance become more informative about reaching critical mass. Monotone signal response is preserved throughout.

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