Political Persuasion under Strategic Fact-Checking

Nektaria Glynia (UCY) Paris Trokkos (UCY & UB) Dimitrios Xefteris (UCY)

The 24th Conference on Research on Economic Theory and Econometrics
Ios, July 07–11, 2026

Introduction

Motivation: political persuasion

Political communication is strategic.

Political actors do not simply reveal information.

- They design strategically messages to shape voters' beliefs and actions;
- Communication can be sophisticated and targeted.
- The same issue can be framed differently for different voters.

Motivation: persuasion under scrutiny

Political persuasion takes place under scrutiny.

Claims may be checked by fact-checkers, journalists, media, or political rivals.

- A misleading claim can be refuted publicly.
- But if a claim survives scrutiny, it may become more credible.
- Voters therefore interpret both refutation and no refutation.
- Fact-checking disciplines persuasion, but it can also certify unrefuted messages.

Research Question: *When persuasion and fact-checking coexist in public view, how do politicians optimally design messages, and when do fact-checkers choose to intervene?*

This paper

- Bayesian persuasion with endogenous fact-checking.
- Persuader commits to a communication policy and anticipates that claims may be checked.
- Checker chooses whether to audit based on communication policy — a political rival, not a neutral/truth seeking checker. (Andreottola and Louis-Sidois, 2025; Louis-Sidois, 2025)
- Persuader maximises support for her preferred policy; Checker minimises this support.
- Heterogeneous voters (lenient vs. skeptical) who update from both refutations and non-refutations.

Related literature I

Bayesian persuasion — Kamenica and Gentzkow (2011)

- **Constrained information design:** Doval and Skreta (2024)*
- **Exogenous lie detection with commitment:** Ederer and Min (2022)*
- **Verification / falsification :** Venkatesh *et al.* (2025); Levkun (2022); Lukyanov and Safaryan (2025); Matyskova and Montes (2023); Yang (2024); Perez-Richet and Skreta (2022)

Political incentives of fact-checking: Andreottola and Louis-Sidois (2025)

Cheap talk — Crawford and Sobel (1982)

- **Lie detection:** Dziuda and Salas (2018); Balbuzanov (2019)
- **Lying costs / costly miscalibration:** Kartik *et al.* (2007); Kartik (2009); Guo and Shmaya (2021)

Related literature II

Other related strands

- **Partial commitment:** Lipnowski *et al.* (2022); Min (2021); Lin and Liu (2024)
- **Audience structure:** Laclau and Renou (2017); Gratton *et al.* (2025)
- **Empirical & Experimental:** Rentschler and Samad (2025)*; Pennycook *et al.* (2020)*

Our contribution

1. We endogenize the audit decision by a strategic fact checker, who *checks message–state consistency* rather than revealing the state. This generates a *certification channel*
2. We consider heterogeneous voters: lenient and skeptists–which is a crucial assumption
3. We show that auditing is selective: the Checker may abstain to avoid certifying claims for skeptical voters.
4. We show that strategic fact checking has no trivial welfare effects.

Model

Players and environment

States, prior, players:

- Two states $\omega \in \{A, B\}$, with $\Pr(\omega = A) = \mu$, $\Pr(\omega = B) = 1 - \mu$.
- Persuader P (wants B) and Checker C (wants A), interact in front of a continuum of voters.

Heterogeneous voters:

- Two types, masses h_1 and $h_2 = 1 - h_1$
- Type i supports B after history X iff $\Pr(\omega = B \mid X) \geq r_i$. Microfoundation
- Thresholds $0 < r_1 < r_2 < 1$: type 1 lenient, type 2 skeptical.

Default-A action: $\mu > 1 - r_1$:
At the prior *no* voter supports B: persuasion is essential, and arises only through belief updating from the signal (and the audit outcome).

The Persuader's communication policy

Information structure (full commitment, as in Kamenica and Gentzkow (2011)–KG)

$$\pi = (\pi_A, \pi_B), \qquad \begin{array}{c|cc} & s=a & s=b \\ \hline \omega = A & 1-\pi_A & \pi_A \\ \omega = B & 1-\pi_B & \pi_B \end{array} \qquad \pi_A, \pi_B \in [0,1], \quad \pi_B \geq \pi_A.$$

- $s = b$: pro-B message; $s = a$: pro-A message.
- $\pi_B := \Pr(s = b \mid \omega = B)$; $\pi_A := \Pr(s = b \mid \omega = A)$
- π_A is the distortion rate (noise): how often b is realised in the wrong state.
- The key variable will be: the relative distortion ratio π_A/π_B.

For realization s, the posterior is $\beta(B \mid s; \pi) = \dfrac{(1-\mu)\pi(s \mid B)}{(1-\mu)\pi(s \mid B) + \mu\pi(s \mid A)}$.

Misleading signals

Our definition aligns with the concept of deception in games by Sobel (2020).

Definition

We say that message s *tilts toward* B if it raises the posterior probability of B relative to the prior, and *tilts toward* A if it lowers it. Formally, the *informational tilt* of a realization s is

$$t(s;\pi) = \begin{cases} B & \text{if } \beta(B \mid s;\pi) > 1-\mu, \\ A & \text{if } \beta(B \mid s;\pi) < 1-\mu, \\ \varnothing & \text{if } \beta(B \mid s;\pi) = 1-\mu. \end{cases}$$

A message s is **misleading** if the message tilts away from the true state, i.e., if it moves beliefs in the wrong direction. The indicator for a misleading event is therefore

$$\text{Mislead}(s;\omega,\pi) = \mathbf{1}\{t(s;\pi) \neq \varnothing \text{ and } t(s;\pi) \neq \omega\}.$$

Fact-checking technology

The Checker's audit: The Checker commits to $\sigma : \{a, b\} \times \Pi \to [0, 1]$: the probability of attempting an audit. An attempt $m = 1$ produces a public outcome $\kappa \in \{0, 1\}$.

- $\kappa = 1$: refutation; $\kappa = 0$: no refutation.
- **No false positives:** a state-consistent message is never refuted.
- A *misleading* message is refuted with probability $1 - f$.
- $f \in (0, 1)$: the failure (false-negative) rate. $f = 1 \Rightarrow$ auditing uninformative $=$ KG.

Key structural feature: The audit is not state verification. It does not directly reveal whether the state is (A) or (B). Instead, it tests the consistency of the message–state pair: whether the realized message matches the realized state. A detected inconsistency refutes the message. But non-refutation is ambiguous: the message may be true, or the audit may simply have failed to detect the inconsistency.

Payoffs & objectives

- The public history is $X = (\pi, s, m, \kappa)$, where $\kappa = \varnothing$ if and only if $m = 0$.
- Let $g_i(X) = \mathbf{1}\{\beta(B \mid X) \geq r_i\}$ denote type i's voting decision.
- The Persuader's payoff is the B-vote share and the Checker's payoff is the A-vote share:

$$U_P(\pi, \sigma) = \sum_{i=1}^{2} h_i \, \mathbb{E}\big[g_i(X)\big], \qquad U_C(\pi, \sigma) = 1 - U_P(\pi, \sigma),$$

  where the expectation is taken over ω, the persuasion signal s, the audit attempt m, and the audit outcome κ.
- Formally, the Persuader chooses π anticipating the Checker's best response:

$$\pi^* \in \arg\max_{\pi \in \Pi} \sum_{i=1}^{2} h_i \, \mathbb{E}\big[g_i(X)\big], \qquad \sigma^*(\cdot, \pi) \in \arg\max_{\sigma \in [0,1]^2} U_C(\pi, \sigma).$$

The Checker's Problem

Solution concept and roadmap

Solution concept: Perfect Bayesian Equilibrium. We solve the game by backward induction:

- Characterize the Checker's best response (Prop. 1) $\Rightarrow$ the Persuader's constraint set.
- Solve the Persuader's problem (Prop. 2) $\Rightarrow$ equilibrium regimes characterisation in $(f, h_1) - \textit{plane}$.
- Welfare analysis

Behind the result: the certification channel

Posteriors after a pro-B message b

$$\beta_0^b(\pi) = \Pr(B \mid b) = \frac{(1-\mu)\pi_B}{(1-\mu)\pi_B + \mu\pi_A} \qquad \text{(no audit)}$$

$$\beta_1^b(\pi) = \Pr(B \mid b, \kappa = 0) = 0 \qquad \text{(refutation)}.$$

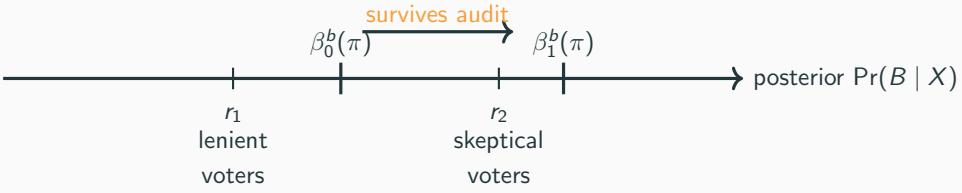
$$\beta_1^b(\pi) = \Pr(B \mid b, \kappa = 1) = \frac{(1-\mu)\pi_B}{(1-\mu)\pi_B + \mu f \pi_A} \geq \beta_0^b(\pi) \qquad \text{(survives audit)}.$$

- A surviving audit is *informative*: $\kappa = 0$ is more likely when b is state-consistent.
- Surviving scrutiny raises the posterior whenever $f < 1$ and $\pi_A > 0$.
- A non-refuted b can therefore become more credible than an unchecked b.

Why heterogeneous voters matter

Voters differ in how much evidence they require to support B:

Unchecked b: $\beta_0^b(\pi) \in [r_1, r_2)$ persuades only lenient voters.
Surviving audit: $\beta_1^b(\pi) > r_2$ can also persuade skeptical voters.



Core mechanism

With heterogeneous voters, the Checker trades off stopping persuasion among lenient voters against certifying the message to skeptical voters. When $r_1 = r_2$ no such trade off exists.

Checker's best response

Proposition 1

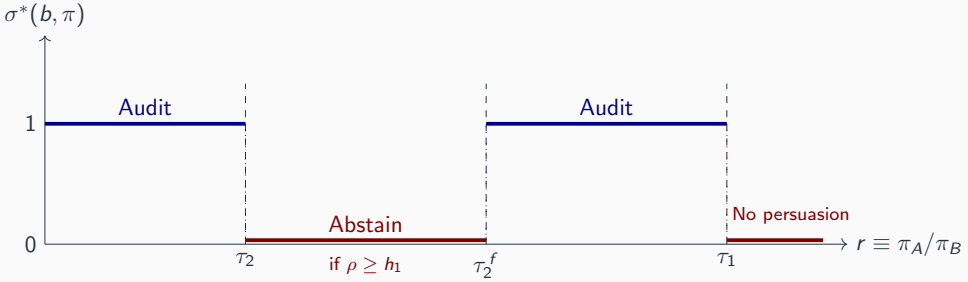
Define the persuasion thresholds $\beta_0^b \geq r_i \iff \frac{\pi_A}{\pi_B} \leq \tau_i$ and $\beta_1^b \geq r_i \iff \frac{\pi_A}{\pi_B} \leq {\tau_i}^f$.

distortion tolerance cutoff: $\tau_i := \dfrac{(1-\mu)(1-r_i)}{\mu r_i}, \qquad {\tau_i}^f := \dfrac{\tau_i}{f}, \qquad i \in \{1,2\}.$

After message $s = a$. The Checker weakly prefers to abstain: $\sigma^*(a,\pi) = 0$. (ii) After message $s = b$. Let $r := \pi_A/\pi_B$ being the relative distorion rate. Define audit survival ratio: $\rho(\pi_A,\pi_B) = \frac{(1-\mu)\pi_B + \mu f\,\pi_A}{(1-\mu)\pi_B + \mu\pi_A}$. Then the optimal audit decision after b is given by

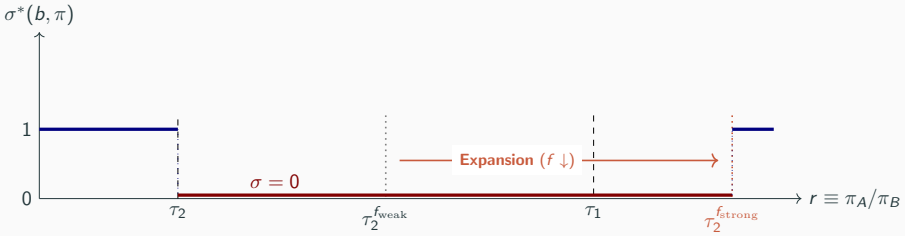
$$\sigma^*(b,\pi) = \begin{cases} 1, & r \leq \tau_2 \quad \text{(skeptics persuaded even without audit)}, \\ \mathbf{1}\{\, h_1 > \rho(\pi_A,\pi_B)\,\}, & \tau_2 < r \leq \min\{{\tau_2}^f, \tau_1\} \quad \text{(backfire zone)}, \\ 1, & {\tau_2}^f < r \leq \tau_1 \quad \text{(“safe audit”)}, \\ 0, & r > \tau_1 \quad \text{(no persuasion)}. \end{cases}$$

The Checker's strategy at a glance



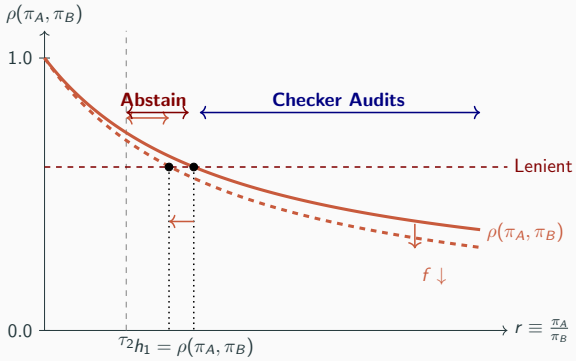
- In the backfire zone, only type 1 supports B without audit, but a *surviving* audit brings type 2 on board.
- The Checker abstains there iff the lenient mass is small: $h_1 \leq \rho(\pi_A, \pi_B)$.

The paradox of improvement (Corollary 1)



- Lower f makes survival a *stronger* signal: τ_2^f shifts right, enlarging the backfire region (weak regime) and erasing the safe-audit gap.
- A Checker with better tools is *more* constrained by the fear of validating the claim.
- Persuader can *increase* noise without triggering an audit

Two opposing forces & comparative statics



Improving fact-checking ($f \downarrow$).

- Extends the range of cases where a successful audit can change voters' beliefs.
- But it also lowers the survival probability of false claims (ρ), making unchecked claims more credible.

Comparative statics.

- Higher prior beliefs (μ) reduce the scope for strategic abstention.
- Greater voter skepticism (r_2) makes persuasion more difficult.

As state-inconsistency π_A rises, ρ falls below h_1 and the incentive to audit is restored.

Equilibrium

The Persuader's problem

Reduction (Lemma)

Under $\mu > 1 - r_1$ and $\pi_B \geq \pi_A$, the a-branch is payoff-irrelevant and the payoff is weakly increasing in π_B. Hence $\pi_B^* = 1$ and the choice reduces to $\pi_A = x \in [0, 1]$.

With $A(x) = (1 - \mu) + \mu x, \quad B(x) = (1 - \mu) + \mu f x, \quad \rho(x) = B(x)/A(x)$

After imposing the Checker's best response (Prop. 1), the Persuader's payoff can lie only on one of four reduced-form branches:

$$V(x) \in \Big\{ 0, \;\; B(x), \;\; h_1 B(x), \;\; \min\{h_1 A(x), B(x)\} \Big\}.$$

- Each branch is weakly increasing in $x \Rightarrow$ optimum at a region endpoint.
- Candidates in $[0, \min\{\tau_1, 1\}]$: $\tau_2, \; \tau_2^f, \; \tau_1$.
- Regime split: strong audit $f \leq K$ ($\tau_2^f \geq \tau_1$) vs. weak audit $f > K$.

Proposition 2 — strong-audit regime ($f \le K$)

Candidates collapse to $\{\tau_2, \tau_1\}$. Define

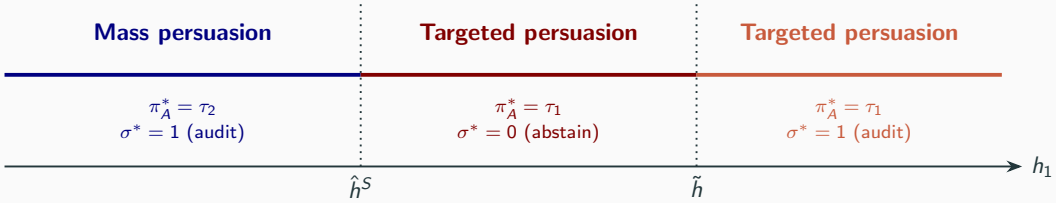
$$\hat{h}^S := \frac{B(\tau_2)}{A(\tau_1)} = \frac{r_1}{r_2}\big(r_2 + f(1-r_2)\big), \qquad \tilde{h} := \rho(\tau_1) = f + r_1(1-f).$$

Equilibrium

$$(\pi_A^*, \sigma^*) = \begin{cases} (\tau_2,\, 1), & h_1 \le \hat{h}^S \quad \text{(mass persuasion)}, \\ (\tau_1,\, 0), & \hat{h}^S < h_1 \le \tilde{h} \quad \text{(targeted, abstain)}, \\ (\tau_1,\, 1), & h_1 > \tilde{h} \quad \text{(targeted, audit)}. \end{cases}$$

- Persuader chooses between a *clean audited* signal and a *noisy unverified* one.
- Abstention arises only on an intermediate h_1-range: $0 < \hat{h}^S < \tilde{h} < 1$.

Strong-audit regime — equilibrium in h_1



When the Persuader plays $\pi_A = \tau_1$, the Checker abstains iff $h_1 \leq \rho(\tau_1) = \tilde{h}$; otherwise she audits.

Proposition 2 — weak-audit regime ($f > K$)

Now $\tau_2^f < \tau_1$; all three candidates matter. The comparison hinges on $f^2 \lessgtr K$.

B.1 Intermediate weak audit ($K < f < \sqrt{K}$)

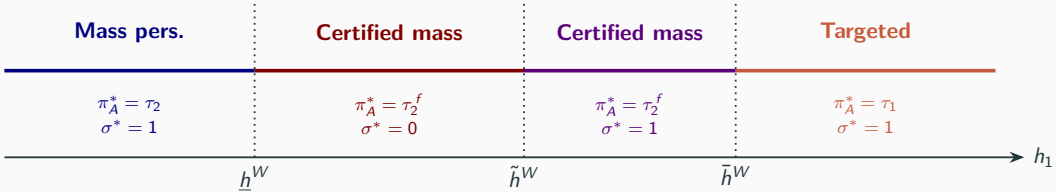
$\underline{h}^W = \dfrac{B(\tau_2)}{A(\tau_2^f)}$, $\tilde{h}^W = \rho(\tau_2^f)$, $\bar{h}^W = \dfrac{B(\tau_2^f)}{B(\tau_1)}$:

$$(\pi_A^*, \sigma^*) = \begin{cases} (\tau_2, 1), & h_1 \le \underline{h}^W, \\ (\tau_2^f, 0), & \underline{h}^W < h_1 \le \tilde{h}^W \quad \text{(certified mass, abstain)}, \\ (\tau_2^f, 1), & \tilde{h}^W < h_1 \le \bar{h}^W, \\ (\tau_1, 1), & h_1 > \bar{h}^W. \end{cases}$$

B.2 Very weak audit ($f \ge \sqrt{K}$)

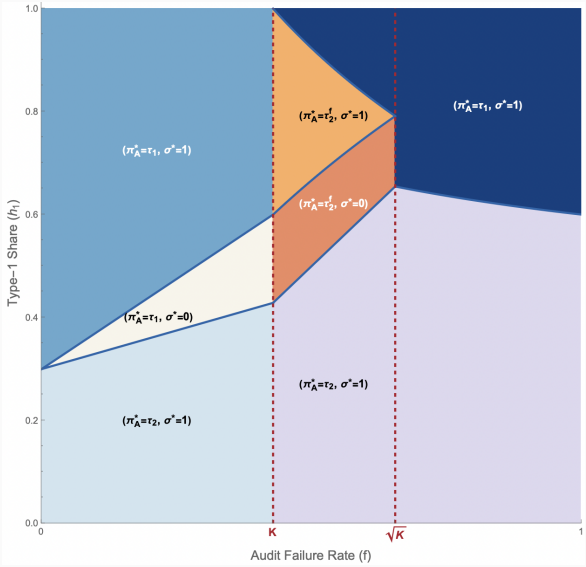
The calibrated cutoff drops out. With $\check{h}^W = B(\tau_2)/B(\tau_1)$: $(\tau_2, 1)$ if $h_1 \le \check{h}^W$, else $(\tau_1, 1)$.

Weak-audit regime — equilibrium in h_1



The Persuader injects calibrated noise τ_2^f to deter the audit. The interval $(\tilde{h}^W, \bar{h}^W]$ is empty when $\bar{h}^W \leq \tilde{h}^W$ (i.e. $f \geq \sqrt{K}$).

Equilibrium map



Equilibrium regimes in (f, h_1)-space. Dashed lines mark K and $\sqrt{K}$.
Axes: Horizontal: f audit failure rate ($f \uparrow =$ weaker auditing).
Vertical: h_1 mass of lenient voters.
Regions: Each region reports the equilibrium pair

$$(\pi_A^*, \sigma^*).$$

The cutoffs

$$K = \frac{\tau_2}{\tau_1}, \qquad \sqrt{K}$$

separate the strong, weak, and very-weak audit cases.

Partial Commitment

Welfare

KG benchmark

Kamenica–Gentzkow benchmark ($f = 1$)

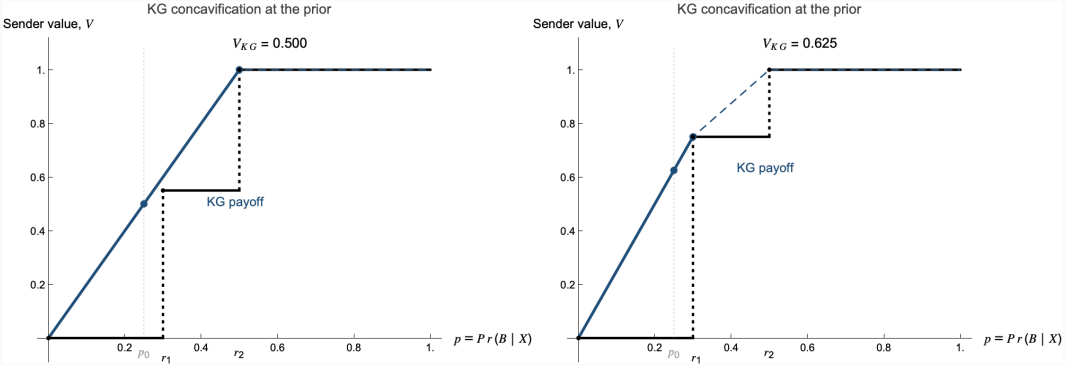
$$V^*_{KG} = \text{cav}\, v(1-\mu) = \max\{A(\tau_2),\ h_1 A(\tau_1)\}, \qquad h_{KG} = \frac{r_1}{r_2}.$$

Persuader chooses *mass persuasion* ($\pi_A = \tau_2$) if $h_1 \leq h_{KG}$, else *targeted* ($\pi_A = \tau_1$).

- Fact-checking adds an endogenous feasibility constraint: a Bayes-plausible signal must also be *compatible with the Checker's auditing incentives*.

Note: In Ederer and Min (2022), with exogenous lie-detection, the persuader can attain the same value.

KG concavification



The KG value is given by the upper-envelope segment evaluated at the prior.

Left panel: For $h_1 < r_1/r_2$, the relevant segment is the one connecting $(0, 0)$ and $(r_2, 1)$;

Right panel: For $h_1 \geq r_1/r_2$, it is the segment connecting $(0, 0)$ and (r_1, h_1).

Proposition 3 — persuader value vs. KG

Persuader value

If $(f, h_1) \neq (K, \frac{r_1}{r_2})$ then

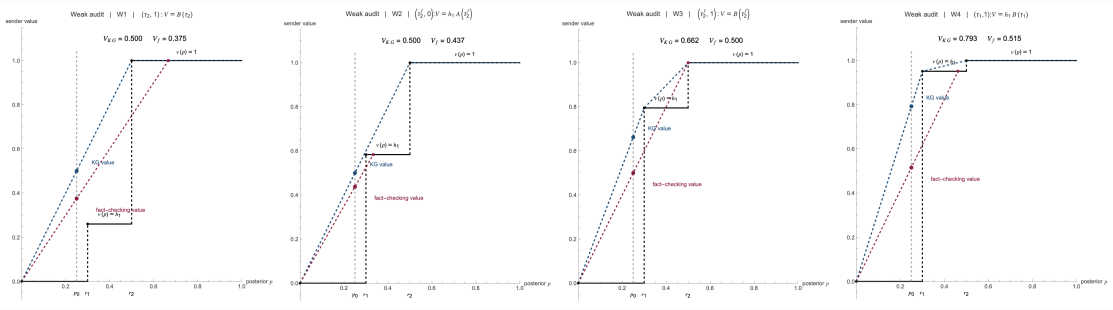
$$V^*(f) < V^*_{KG} \quad \text{for all } f \in (0, 1],$$

- The loss is not only from audits that occur on path.
- Even when the Checker *abstains*, the Persuader is constrained to keep abstention optimal
  $\Rightarrow$ strategic abstention does not restore the KG payoff.

Next step: Comparison with exogenous/mechanical fact checking.

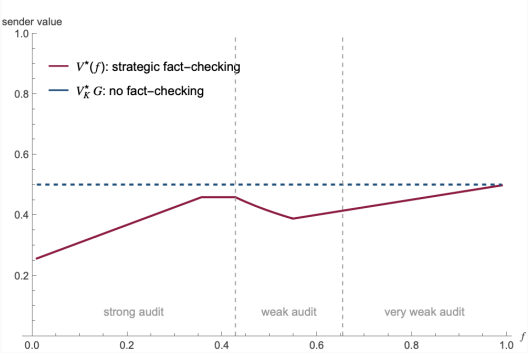
Strong audit | Weak audit | Very weak audit

Value comparison: Weak audit regime



Persuader value in the weak-audit regime. The figure compares the KG benchmark with the equilibrium value under strategic fact-checking for $K < f < \sqrt{K}$. Panel W1 shows mass persuasion at τ_2 with auditing, yielding $V = B(\tau_2)$. Panel W2 shows the calibrated cutoff τ_2^f with abstention, yielding $V = h_1 A(\tau_2^f)$. Panel W3 shows the same cutoff τ_2^f with auditing, yielding $V = B(\tau_2^f) = A(\tau_2)$; in this case, survival of the audit certifies the message and brings skeptical voters exactly to their cutoff r_2. Panel W4 shows targeted persuasion at τ_1 with auditing, yielding $V = h_1 B(\tau_1)$. Across all weak-audit regimes, strategic fact-checking yields a value weakly below, and generically strictly below, the KG benchmark.

The value is non-monotone in audit quality



Parameter values are $\mu = 0.75$, $r_1 = 0.3$, $r_2 = 0.5$, and $h_1 =$.

Corollary. For an open set of h_1, $V^*(f, h_1)$ is non-monotone in f:

- On the calibrated-abstention branch ($K < f < \sqrt{K}$), $V^* = h_1 A(\tau_2/f)$ is decreasing in f.
- After the branch switch it increases again, strictly throughout $\sqrt{K} \leq f < 1$.
- Converges to V_{KG}^* as $f \to 1$.

Blackwell informativeness

Lemma (fixed signal)

Fix $\pi = (\pi_A, 1)$ and condition on b being audited. If $f' \leq f$, the audit experiment with f' Blackwell-dominates that with f: auditing is a mean-preserving spread of the unaudited posterior, and lower f widens it.

Proposition (equilibrium)

The equilibrium public experiment $\mathcal{E}^*(f, h_1)$ is non-monotone in f: informativeness *increases* with f on an interval of $(K, \sqrt{K})$, then *decreases* throughout $\sqrt{K} \leq f < 1$.

- In equilibrium, f moves *both* the Persuader's cutoff and the Checker's audit decision $\Rightarrow$ it changes *which* experiment is induced.
- **Corollary:** no uniform Blackwell ranking between strategic fact-checking and KG.

Discussion and Conclusion

Discussion

1. Checker's commitment. The Checker conditions on *both* π *and* the message s. This contingent commitment drives $\pi_B = 1$; if she could not condition on s, we would expect Ederer and Min (2022)-style mixing with $\pi_B < 1$.

2. State verification vs. message-refutation. Our audit targets the *message*, not the state. "No refutation" is more likely under a true message $\Rightarrow$ the certification effect, which is absent under direct state verification.

3. Intermediate priors. $r_1 < 1 - \mu < r_2$: auditing after a can matter; defensive distortion $\pi_B < 1$ may arise.

Conclusion

- Imperfect, public fact-checking both disciplines and amplifies persuasion: "no refutation" is a persuasive event.
- This certification channel reshapes a strategic Checker's incentives, generating strategic abstention — auditing is *not* monotone in technological quality.
- Voters' heterogeneity is crucial
- The Persuader sometimes design more precise info structure to make the checker to abstainn.
- Welfare analysis suggests non monotonic results.

Thank you!

Microfoundation of voters' cutoffs

A microfoundation for the cutoff rule is as follows.

- Suppose a type-i voter chooses $v \in \{A, B\}$ and obtains payoff 0 if $v = \omega$, payoff $-\ell_i$ if $v = B$ when $\omega = A$, and payoff $-k_i$ if $v = A$ when $\omega = B$, where $k_i, \ell_i > 0$. Letting $p = \Pr(\omega = B \mid X)$, voting B yields $-(1-p)\ell_i$ and voting A yields $-pk_i$, so type i votes for B iff $p \geq \ell_i/(k_i + \ell_i)$.
- Thus $r_i = \ell_i/(k_i + \ell_i)$. Therefore the threshold rule can capture asymmetric losses. In the binary "judge" example used to illustrate Bayesian persuasion in Kamenica and Gentzkow (2011), the receiver gets payoff 1 from the correct decision and 0 from the incorrect one.
- Equivalently (up to an affine transformation), one can normalize payoffs so that the correct decision yields 0 and any mistake yields -1, i.e., $u_i(B, B) = u_i(A, A) = 0$ and $u_i(B, A) = u_i(A, B) = -1$. In the asymmetric-loss notation in the text, this corresponds to $k_i = \ell_i = 1$, so the induced cutoff is $r_i = \ell_i/(k_i + \ell_i) = 1/2$.

Back

Deception in Games (Sobel, 2020)

Sobel defines a message as deceptive in state ω if there exists an *available alternative message* that would induce a belief closer to the truth. Consequently, when the true state, for instance, is A, the realization $s = b$ is deceptive for two reasons:

1. Belief movement: It induces a posterior belief in the false state B that is strictly higher than the prior, moving the receiver 'further from the truth.[1]

2. Available truth: The Persuader could have chosen a structure that produced the message $s = a$ (which indicates A) but instead chose a structure that produced $s = b$. Under Sobel's taxonomy, because a more informative signal was feasible, the noisy signal constitutes deception.

Back

[1] Note that this concept of deception is fully consistent with voter rationality. Voters in our model are sophisticated Bayesians who perfectly understand the signal structure π and form accurate posterior beliefs. Deception in equilibrium does not imply that receivers are "fooled" or hold incorrect beliefs given their information. Rather, they are deceived in the instrumental sense: the signal structure is designed to induce a rational posterior belief that favors the false state.

Partial commitment-setup

Partial commitment — setup Commitment binds w.p. λ; otherwise the Persuader picks the message freely. The effective state-inconsistency rate is

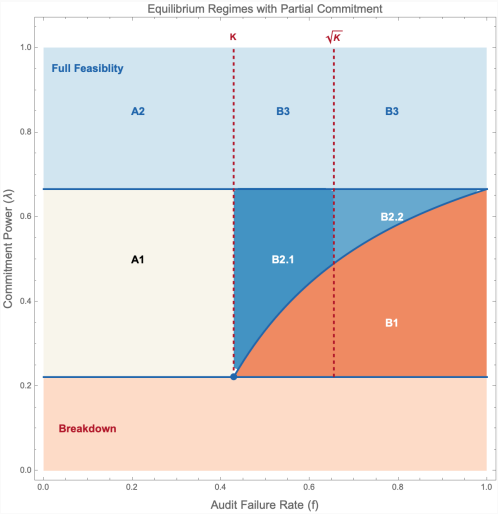
$$x := \bar{\pi}_A = 1 - \lambda + \lambda \pi_A \in [\underline{\pi}(\lambda), 1], \qquad \underline{\pi}(\lambda) := 1 - \lambda.$$

Reduction (Lemma)

There is a Persuader-optimal PBE with $\pi_B = 1$ and $\tilde{\pi}_A = \tilde{\pi}_B = 1$: the problem is the baseline with $\pi_B = 1$ and a feasibility floor $x \geq 1 - \lambda$.

- Weaker commitment (lower λ) raises the minimal noise $\Rightarrow$ rules out moderate policies.
- If $\lambda < \lambda_1 := 1 - \min\{\tau_1, 1\}$: persuasion collapses (babbling-like).
- **Not** cheap talk / costly lying: the certification mechanism is *endogenously shut down*, because auditing would only *help* the Persuader.

Partial commitment — equilibrium



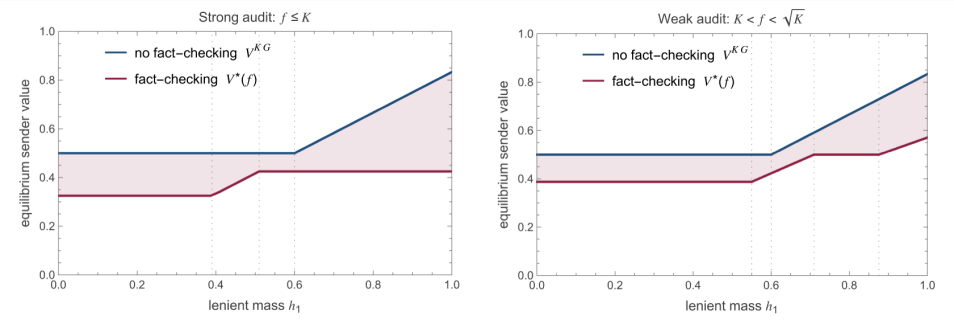
Proposition. With $\lambda_2^f = 1 - \min\{\tau_2^f, 1\}$, etc., the baseline regimes reappear once the relevant cutoff is feasible, and collapse sequentially as $\lambda \downarrow$.

- Calibrated deterrence at $\bar{\tau}_2^f$ needs $\lambda \geq \lambda_2^f$.
- Too little commitment $\Rightarrow$ Persuader *must* be noisy $\Rightarrow$ certification threat vanishes $\Rightarrow$ Checker audits.

Corollary

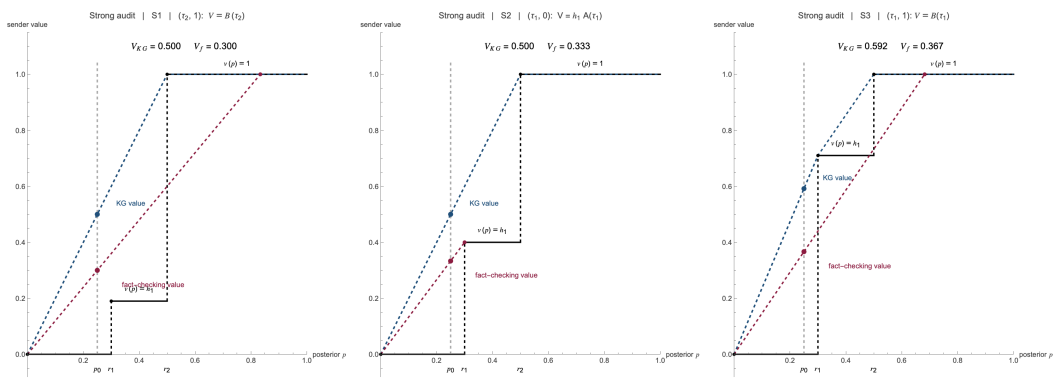
$V^*(f, \lambda)$ is weakly increasing in λ — a *commitment-to-moderation* effect.

Timing | Equilibrium



Persuader value with and without strategic fact-checking. The blue curve reports the KG benchmark V^*_{KG}, corresponding to uninformative auditing, while the red curve reports the equilibrium value $V^*(f)$ under strategic fact-checking. The shaded area is the Persuader's welfare loss. Parameter values are $\mu = 0.75$, $r_1 = 0.3$, and $r_2 = 0.5$, so $K = 3/7$ and $\sqrt{K} \simeq 0.655$. The left panel uses $f = 0.3 < K$ and the right panel uses $f = 0.5 \in (K, \sqrt{K})$.

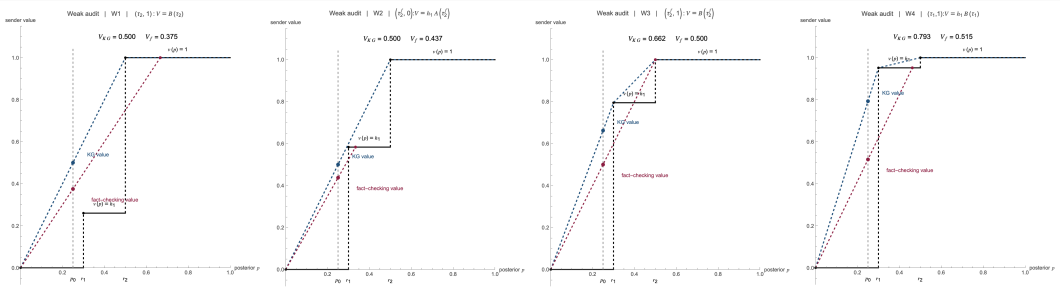
Value comparison: Strong audit regime



Persuader value in the strong-audit regime. The figure compares the KG benchmark with the equilibrium value under strategic fact-checking for $f \leq K$. The blue dashed line is the KG concavification, while the red dashed line is the value generated by the fact-checking technology. Panel S1 shows mass persuasion at τ_2 with auditing, yielding $V = B(\tau_2)$. Panel S2 shows targeted persuasion at τ_1 with abstention, yielding $V = h_1 A(\tau_1)$. Panel S3 shows targeted persuasion at τ_1 with auditing, yielding $V = B(\tau_1)$. In all three cases, the Persuader's value under strategic fact-checking is below the KG benchmark, except at the knife-edge equality case described in Proposition **??**.

Back

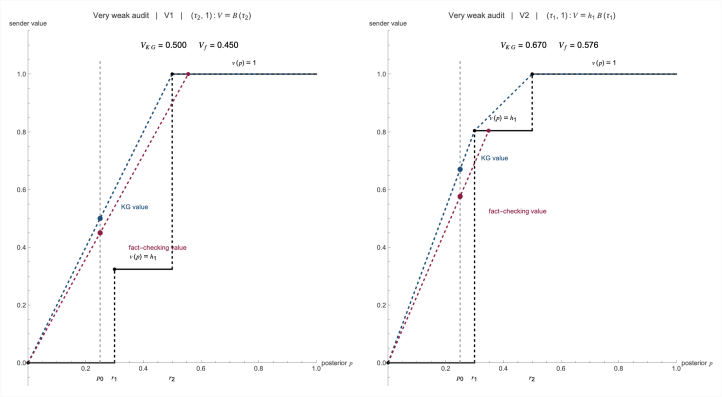
Value comparison: Weak audit regime



Persuader value in the weak-audit regime. The figure compares the KG benchmark with the equilibrium value under strategic fact-checking for $K < f < \sqrt{K}$. Panel W1 shows mass persuasion at τ_2 with auditing, yielding $V = B(\tau_2)$. Panel W2 shows the calibrated cutoff τ_2^f with abstention, yielding $V = h_1 A(\tau_2^f)$. Panel W3 shows the same cutoff τ_2^f with auditing, yielding $V = B(\tau_2^f) = A(\tau_2)$; in this case, survival of the audit certifies the message and brings skeptical voters exactly to their cutoff r_2. Panel W4 shows targeted persuasion at τ_1 with auditing, yielding $V = h_1 B(\tau_1)$. Across all weak-audit regimes, strategic fact-checking yields a value weakly below, and generically strictly below, the KG benchmark.

Back

Value comparison: Very weak audit regime



Persuader value in the very weak-audit regime. The figure compares the KG benchmark with the equilibrium value under strategic fact-checking for $f \geq \sqrt{K}$. Panel V1 shows mass persuasion at τ_2 with auditing, yielding $V = B(\tau_2)$. Panel V2 shows targeted persuasion at τ_1 with auditing, yielding $V = h_1 B(\tau_1)$. Unlike in the weak-audit regime, the calibrated abstention cutoff τ_2^f is no longer active. Thus, the Checker audits on the equilibrium path in both very weak-audit regions, and the Persuader's value remains below the KG benchmark except in the limit as $f \to 1$.

Back

References

Andreottola, G. and Louis-Sidois, C. (2025). Why us and not them? a theory of political fact-checking. *WP*.

Balbuzanov, I. (2019). Lies and consequences: The effect of lie detection on communication outcomes. *International Journal of Game Theory*, **48** (4), 1203–1240.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pp. 1431–1451.

Doval, L. and Skreta, V. (2024). Constrained information design. *Mathematics of Operations Research*, **49** (1), 78–106.

Dziuda, W. and Salas, C. (2018). Communication with detectable deceit. *Available at SSRN 3234695*.

Ederer, F. and Min, W. (2022). *Bayesian persuasion with lie detection*. Tech. rep., National Bureau of Economic Research.

Gratton, G., Teh, C. and Thornton, D. (2025). Targeted persuasion.

References ii

Guo, Y. and Shmaya, E. (2021). Costly miscalibration. *Theoretical Economics*, **16** (2), 477–506.

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

Kartik, N. (2009). Strategic communication with lying costs. *The Review of Economic Studies*, **76** (4), 1359–1395.

—, Ottaviani, M. and Squintani, F. (2007). Credulity, lies, and costly talk. *Journal of Economic theory*, **134** (1), 93–116.

Laclau, M. and Renou, L. (2017). Public persuasion. *Manuscript.[1079]*.

Levkun, A. (2022). *Communication with strategic fact-checking*. Tech. rep., mimeo. University of California, San Diego.

Lin, X. and Liu, C. (2024). Credible persuasion. *Journal of Political Economy*, **132** (7), 2228–2273.

Lipnowski, E., Ravid, D. and Shishkin, D. (2022). Persuasion via weak institutions. *Journal of Political Economy*, **130** (10), 2705–2730.

Louis-Sidois, C. (2025). Both judge and party? investigating the political unbiasedness of fact-checkers. *Journal of the European Economic Association*, **23** (6), 2137–2164.

References iii

Lukyanov, G. and Safaryan, S. (2025). Public persuasion with endogenous fact-checking. *arXiv preprint arXiv:2508.19682*.

Matyskova, L. and Montes, A. (2023). Bayesian persuasion with costly information acquisition. *Journal of Economic Theory*, **211**, 105678.

Min, D. (2021). Bayesian persuasion under partial commitment. *Economic Theory*, **72** (3), 743–764.

Pennycook, G., Bear, A., Collins, E. T. and Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, **66** (11), 4944–4957.

Perez-Richet, E. and Skreta, V. (2022). Test design under falsification. *Econometrica*, **90** (3), 1109–1142.

Rentschler, L. and Samad, Z. (2025). Bayesian persuasion with fact-checking: An experimental investigation. *European Economic Review*, p. 105114.

Sobel, J. (2020). Lying and deception in games. *Journal of Political Economy*, **128** (3), 907–947.

Venkatesh, R. S., Roy, S. and Pramanik, S. (2025). Bayesian persuasion and costly state verification. *Available at SSRN 5396241*.

Yang, L. L. (2024). Information design with costly state verification, cRC TR 224 Discussion Paper No. 502, University of Bonn and University of Mannheim.