# Supplemental Appendix to "A Model of Populism as a Conspiracy Theory"*

Adam Szeidl
Central European University and CEPR

Ferenc Szucs
Stockholm University

June 12, 2025

## Appendix

This material supplements our paper "A Model of Populism as a Conspiracy Theory" provides proofs for complementary theoretical results as well as additional evidence.

## A Definitions and proofs

### A.1 Formal model of audiences

Our baseline model has a unit mass of voters distributed uniformly on the unit square. We index voters by $i = (j, w)$, thus

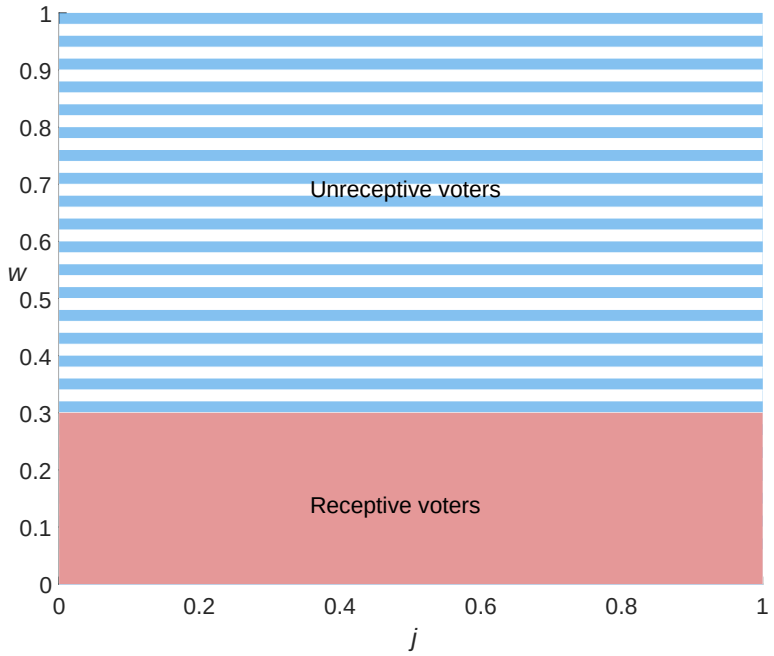
$$\int_0^1 \int_0^1 1 dj dw = 1$$

As it is shown in Figure 1, the first $\alpha$ share of voter along dimension $w$—$\alpha = 0.3$ in the figure—are receptive to propaganda, while the rest are unreceptive. We index media—both elite and new—by $j$ and its audience is given by

$$\text{Audience of media } j = \{(z, w) \in [0, 1]^2 : z = j\}$$

*Emails: szeidla@ceu.edu, ferenc.szucs@su.se.

Figure 1: Media audiences




## A.2 Microfoundation of messengers' objectives

In the main model we assume that the AR elite and the incumbent politician care about the average voter belief about the politician's type. Here we provide microfoundations for this assumption using a probabilitic voting model, in which after stage 2 of the game an election takes place between the incumbent and a challenger. The challenger is good with probability $q_c^c$. Voter $i$ chooses between the incumbent and a challenger to maximize utility

$$U_{v,i} = c\tilde{\theta}_c + \lambda \cdot 1_{\{\text{Incumbent}\}} + \epsilon + \eta_i, \tag{A1}$$

where $v \in \{rec, un\}$; $\tilde{\theta}_c$ is the competence of the elected politician; $\lambda$ is an additional preference component of the voter about the incumbent, which reflects ideological alignment; and $\epsilon$ and $\eta_i$ are mean-zero, independent, uniformly distributed common and individual preference shocks, which have supports $[-\bar{g}, \bar{g}]$ and $[-\bar{h}, \bar{h}]$, constant densities $g = 1/(2\bar{g})$ and $h = 1/(2\bar{h})$. We assume that $\bar{h} > c + \lambda + \bar{g}$ and $\bar{g} > c + \lambda$ to avoid corner outcomes.

Elite members' preferences are given by

$$\tilde{U}_{e,j} = c\tilde{\theta}_c - \lambda \cdot 1_{\{\text{Incumbent}\}} \tag{A2}$$

reflecting that their ideology is the opposite of the voters'. Thus, $\lambda > 0$ corresponds to the incumbent being ideologically pro-voter, while $\lambda < 0$ corresponds to the incumbent being ideologically pro-elite.

The incumbent politician's preferences are given by

$$\tilde{U}_p = E \cdot 1_{\{\text{In office}\}} - \tilde{f} \cdot p, \tag{A3}$$

where $E$ is an ego rent and $\tilde{f}$ is the cost of propaganda.

The following Lemma shows that the preferences in this microfounded model are equivalent to those in the model in the main text, implying that the two models have the same equilibria.

**Lemma 1.** *In this model, the expected utilities of the elite and the politician, conditional on the politician's type $\theta_c$ and the message profile $(\hat{s}, \hat{p})$, are positive affine transformations of the utility functions introduced in the main text*

$$U_{e,j}(\theta_c, \hat{p}, \hat{s}) = (\theta_c - \kappa)\bar{\mu}(\theta_c = 1|\hat{p}, \hat{s})$$
$$U_p(\theta_c, \hat{p}, \hat{s}) = \bar{\mu}(\theta_c = 1|\hat{p}, \hat{s}) - f \cdot p,$$

*where $\kappa \equiv q_c^c + \frac{\lambda}{c}$ is the cost of reelecting the incumbent for the elite, and $f \equiv \frac{\tilde{f}}{E \cdot g \cdot c}$ is the normalized cost of propaganda.*

**Proof of Lemma 1.** The probability, conditional on a fixed common shock $\epsilon$, that voter $i$ votes for the incumbent is

$$\Pr\left[c(q_c^c - \mu_{v,i}(\theta_c|\hat{p}, \hat{s}_{j(i)})) - \lambda - \epsilon < \eta_i|\epsilon\right] = 0.5 - h\left[c(q_c^c - \mu_{v,i}(\theta_c|\hat{p}, \hat{s}_{j(i)})) - \lambda - \epsilon\right]$$

because $\eta_i$ has a uniform distribution with a density $h$. The incumbent wins the election if he gets the majority of votes:

$$\int \left\{0.5 - h\left[c(q_c^c - \mu_i(\theta_c = 1|\hat{p}, \hat{s}_{j(i)})) - \lambda - \epsilon\right]\right\} di > 0.5$$
$$c(q_c^c - \bar{\mu}(\theta_c = 1|\hat{p}, \hat{s})) - \lambda < \epsilon,$$

where voters' average posterior belief of the politician's type is given by

$$\begin{aligned}\bar{\mu}(\theta_c = 1|\hat{p}, \hat{s}) &\equiv \int \mu_i(\theta_c = 1|\hat{p}, \hat{s}_{j(i)})di = \int\int \mu_{(j,w)}(\theta_c = 1|\hat{p}, \hat{s}_j)dwdj \\ &= \int \left[\alpha\mu_{rec,j}(\theta_c = 1|\hat{p}, \hat{s}_j)dj + (1-\alpha)\mu_{un,j}(\theta_c = 1|\hat{p}, \hat{s}_j)\right] dj.\end{aligned}$$

In the second line we use the notation that $\mu_{rec,j}$ and $\mu_{un,j}$ is the average belief of all receptive voters and all unreceptive voters, respectively, in the audience of elite member $j$. Because voters within the audience of elite member $j$ and voter type (receptive/unreceptive) access the same signals, their beliefs are the same, so both of these averages are averaging a constant. Moreover, since functions $\mu_{rec,j}(\theta_c = 1|\cdot)$ and $\mu_{un,j}(\theta_c = 1|\cdot)$ are the same for each elite member $j$, the integral is maximized by the same value of $\hat{s}_j$ for all $j$. Thus, the optimal behavior of the AR elite is to choose the same message $s$ for all members, and therefore below simply denote $\hat{s}_j$ by $\hat{s}$.

The incumbent's probability of winning is thus

$$\begin{aligned}P &\equiv \Pr\left[c(q_c^c - \bar{\mu}(\theta_c = 1|\hat{p}, \hat{s})) - \lambda < \epsilon\right] \\ &= g \cdot c \cdot \bar{\mu}(\theta_c = 1|\hat{p}, \hat{s}) + g(\lambda - c \cdot q_c^c) + 0.5.\end{aligned} \tag{A4}$$

Now consider the AR elite. Her conditional expected utility is

$$\begin{aligned}E[\tilde{U}_e|\theta_c, \hat{p}, \hat{s}] &= P(c\theta_c - \lambda) + (1-P)cq_c^c \\ &= g \cdot c^2(\theta_c - \kappa)\bar{\mu}(\theta_c = 1|\hat{p}, \hat{s}) \\ &\quad + [g(\lambda - cq_c^c) + 0.5][c(\theta_c - q_c^c) - \lambda] + cq_c^c \\ &= L_e[(\theta_c - \kappa)\bar{\mu}(\theta_c = 1|\hat{p}, \hat{s})]\end{aligned}$$

where $\kappa \equiv q_c^c + \frac{\lambda}{c}$ and $L_e$ is a positive affine transformation, as claimed. Note that $L_e$ depends on the state $\theta_c$, but this is not a problem because the state is exogenous from the perspective of all actors.

Next consider the politician. His expected utility is

$$\begin{aligned}E(\tilde{U}_p|\hat{p}, \hat{s}) &= E \cdot P - \tilde{f} \cdot p \\ &= E\left[g \cdot c \cdot \bar{\mu}(\theta_c = 1|\hat{p}, \hat{s}) + g(\lambda - c \cdot q_c^c) + 0.5\right] - \tilde{f} \cdot p \\ &= E \cdot g \cdot c\left[\bar{\mu}(\theta_c = 1|\hat{p}, \hat{s}) - f \cdot p\right] + E[g(\lambda - c \cdot q_c^c) + 0.5] \\ &= L_p[\bar{\mu}(\theta_c = 1|\hat{p}, \hat{s}) - f \cdot p],\end{aligned}$$

where $L_p$ is a positive affine transformation, as claimed.

### A.3 Definition of equilibrium

We start with introducing notation for players' types. We encode the reality state $\theta_r \in \Theta_r = \{R, AR\}$ in the types too. We define the politician's type to be $\theta_p = (\theta_c, \theta_r)$. We define the elite's type to be $\theta_e = (\hat{\theta}_c, \theta_r)$, which differs from the politician's type only because the elite does not observe $\theta_c$ directly, only a signal $\hat{\theta}_c$ on it. We define the receptive voter's type to be $\theta_{rec} = \theta_m$ because his priors depend on $\theta_m$. The unreceptive voter does not have a type. We denote the action of actor $k$ in stage $t \in \{1, 2\}$ by $a_k^t$. We let $\hat{a}_k^t$ stand for the realized action after Nature's tremble, and $\hat{a}^t$ for the realized action profile. The history at stage $t$ is denoted by $\hat{h}^t = (\hat{a}^1, ..., \hat{a}^t)$.

We define strategies as probability distributions over actions at the stages where an actor gets to move. Because the politician and the elite only move in stage 1, their strategies only depend on their type, and are denoted by $\sigma_p(a_p^1|\theta_p)$ respectively $\sigma_e(a_e^1|\theta_e)$. As the receptive voter moves in stage 2 after observing $\hat{a}^1 = (\hat{s}, \hat{p})$, his strategy depends on $\hat{a}^1$ and is denoted by $\sigma_{rec}(a_{rec}^2|\theta_{rec}, \hat{a}^1)$. The unreceptive also moves in stage 2 but only observes $\hat{s}$, not $\hat{p}$. Thus, his strategy only depends on $\hat{s}$, but for ease of notation we will denote it by $\sigma_{un}(a_{un}^2|\hat{a}^1)$. We let $\hat{\sigma}$ denote perturbed strategies that incorporate Nature's trembles. We denote the prior belief of actor $k$ of type $\theta_k$ by $\mu_k^0(\theta|\theta_k)$, and the posterior belief after history $\hat{h}^t$ by $\mu_k^t(\theta|\theta_k, \hat{h}^t)$. We allow beliefs to depend on types, both because the types of different actors are correlated so that the type of $k$ has information about the types of $-k$, and because different types can have different priors.

Our equilibrium concept is a version of perfect Bayesian equilibrium that recognizes our framework's departure from common priors and full rationality. As usual, equilibrium requires that actors best respond and form consistent beliefs. We begin with beliefs. We first note that because of the trembles beliefs will be always well defined. Belief consistency does not impose any condition on the politician or the elite, because they move only at stage 1 where they know only their priors. Belief consistency for the receptive voter requires that he follows Bayesian updating at the end of stage 1:

$$\mu_{rec}^1(\theta_{-rec}|\theta_{rec}, \hat{a}^1) = \frac{\mu_{rec}^0(\theta_{-rec}|\theta_{rec}) \cdot \hat{\sigma}_{-rec}^1(\hat{a}^1|\theta_{-rec})}{\sum_{\theta'_{-rec}} \mu_{rec}^0(\theta'_{-rec}|\theta_{rec}) \cdot \hat{\sigma}_{-rec}^1(\hat{a}^1|\theta'_{-rec})} \tag{A5}$$

where $\mu^0_{rec}(\theta_{-rec}|\theta_{rec})$ is the prior of the receptive voter of type $\theta_{rec}$ about the types of the other actors $\theta_{-rec} = (\theta_c, \hat{\theta}_c, \theta_r)$. This definition accounts for the model's deviation from rationality that the receptive voter's mind type and beliefs may change in stage 1, by computing the posterior for each mind type $\theta_m = N, P$ using the prior associated with that mind type. In particular, if the receptive voter is reached by propaganda and becomes persuaded, (A5) computes his posterior from the prior of the persuaded voter $\mu^0_{rec}(.|\theta_m = P)$. Intuitively, because the persuaded voter uses Bayes rule, he infers from the presence of propaganda about the politician's type; but because propaganda also influences his type, this inference is based on the prior modified by propaganda. Implicit in this is that when the receptive voter receives messages $\hat{a}^1 = (\hat{s}, \hat{p})$, first propaganda $\hat{p}$ changes his mind type and prior, and then he updates from his new prior based on the information content of $\hat{a}^1$. Finally, the unreceptive voter performs standard Bayesian updating based on observing $\hat{s}$.

We next formulate the best-response condition. To do so, we introduce subjective expected utility. In the model presented in the main text only the politician and the elite derive utility, while in the microfoundation presented above the voters also derive utility. In both cases, each actor who maximizes utility, at each stage where it moves, has a subjective probability distribution over final outcomes, where the final outcome is mean voter beliefs in the model presented in the main text. This distribution can differ from the objectively correct distribution because the persuaded voter has an incorrect prior about $\theta$. Actor $k$ at stage $t$ uses its subjective probability distribution over outcomes to compute its subjective expected utility, denoted $U_k(\sigma|\hat{h}^t, \theta_k, \mu_k(\theta|\theta_k, \hat{h}^t))$. For the unreceptive voter who does not observe the full history, we use the same notation to represent his expected utility conditional on only the part of history $\hat{h}^t$ that he does observe. Then the best-response property of equilibrium is that at each stage $t$ at which $k$ has a move, for all actions $\sigma'_k$ available to $k$,

$$U_k(\sigma|\hat{h}^t, \theta_k, \mu_k(.|\theta_k, \hat{h}^t)) \geq U_k((\sigma'_k, \sigma_{-k})|\hat{h}^t, \theta_k, \mu_k(.|\theta_k, \hat{h}^t)).$$

Finally, we need to define what we mean by a mixed equilibrium in this model with an infinitesimal lying cost. We say a mixed equilibrium respects the lying cost if (a) it is a mixed equilibrium; and (b) for any $\varepsilon > 0$ there exists $\delta > 0$ such that for a lying cost $\chi$ below $\delta$ there exists an equilibrium in which all mixing probabilities are within $\varepsilon$ of the original equilibrium. We only consider

equilibria that respect the lying cost.

### A.4 Proof of Proposition 1

Going beyond the result stated in the main text, we characterize the unique PPO equilibrium for all values of $\alpha$. We start with the definitions of two equilibrium profiles.

**Definition 1.** A strategy profile has the *simple propaganda form* if

1. In the reality (R):
    - The elite reports truthfully,
    - The politician sends propaganda if he can and he is bad.
2. In the alternative reality (AR):
    - The elite always reports that the politician is bad,
    - The politician sends propaganda if he can.

**Definition 2.** A strategy profile has the *complex propaganda form* if the elite in the AR, when the signal is good, randomizes between the good and the bad message, while the elite in the R and all politician types behave as in the simple propaganda profile.

We now prove the following generalization of Proposition 1.

**Proposition 1 (Appendix).** *Under Assumptions 1 and 2 there exists $\bar{\pi} < 1$ such that for $\pi > \bar{\pi}$ there exists $\alpha(\pi) > 0.5$ such that*

1. *For $\alpha < \alpha(\pi)$ the unique PPO equilibrium has the simple propaganda form.*
2. *For $\alpha > \alpha(\pi)$ the unique PPO equilibrium has the complex propaganda form.*

**Proof.** Because the proof is long, we have broken it into several numbered steps. As we have seen in the definition of the equilibrium, since messengers' observe the reality state $\theta_r$ their type vector contain the state of reality. Therefore, from now on, we refer to politician or elite observing R (AR) reality as R (AR) politician or elite.

1. Voter beliefs in the simple propaganda profile

We first derive voters' posterior beliefs assuming that play follows the simple propaganda profile. These formulas will be key for the analysis. The $1-\alpha$ share of unreceptive voters have the following posterior beliefs, irrespective of propaganda, as a function of the elite's message:

$$\mu_{un}(\theta_c = 1|\hat{s}) = \hat{s}\frac{\pi q_c}{\pi q_c + (1-\pi)(1-q_c)} + (1-\hat{s})\frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)}. \tag{A6}$$

This expression follows by straightforward Bayesian updating from the elite's message, under the assumption (made by these voters) that the elite's message equals her signal and hence is correct with probability $\pi$.

Consider next the share $\alpha$ of receptive voters. In the absence of propaganda, their beliefs are given by (6), which we repeat here for convenience

$$\mu_{rec}(\theta_c = 1|\hat{p} = 0, \hat{s}, \theta_m = N) = \hat{s}\frac{\pi q_c}{\pi q_c + (1-\pi)(1-q_c)\beta} + (1-\hat{s})\frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)\beta}. \tag{A7}$$

This equation is also the result of straightforward Bayesian updating. The difference relative to (A6) is that these voters, since they are capable of observing it, learn from the fact that there is no propaganda. This mechanism explains the terms involving $\beta$ in the denominators, since $\beta$ is the probability with which the bad politician is unable to send propaganda. Thus, a good message, absent propaganda, can reflect a bad politician, an incorrect elite signal, and the inability to send propaganda, captured in the denominator in the first term; and a bad message, absent propaganda, can reflect a bad politician, a correct elite signal, and the inability to send propaganda, captured in the denominator of the second term.

An implication is that because $\beta > 0$, the voter does not fully infer from the absence of propaganda that the politician is good, so that the elite's message is still informative for his updating. In fact, (A7) implies that for $\pi$ large (holding fixed $\beta$) beliefs are primarily determined by the elite's message $\hat{s}$, so that they are near one when $\hat{s} = 1$ and near zero when $\hat{s} = 0$. Intuitively, even though the absence of propaganda is informative, the elite's message is a more informative signal.

Finally, the beliefs of the receptive voter in the presence of propaganda are given by (7), which we repeat here for convenience

$$\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s}, \theta_m = P) = (1-\hat{s})\frac{q_{ar}q_c}{q_{ar} + q_r\pi(1-q_c)}. \tag{A8}$$

This formula too follows from Bayesian updating. Consider first a bad elite message. Since propaganda changed the voter's prior, he thinks that the politician may be good in the AR, explaining the numerator. However, propaganda and a bad signal can also emerge in the AR if the politician is bad, and in R if the politician is bad and the elite's message is correct, explaining the denominator. Consider next a good elite message. The profile of praise and propaganda is only possible in R and proves that the politician is bad.

2. Cutoff $\alpha$ value

We turn to characterize the condition on $\alpha$ under which the simple propaganda equilibrium exists. This will turn on whether, in the simple propaganda profile, the AR elite finds it optimal to criticize after propaganda. Since the goal of the AR elite is to minimize voter beliefs, it follows from the above expressions that she chooses to criticize if and only if

$$(1-\alpha)\left[\frac{\pi q_c}{\pi q_c + (1-\pi)(1-q_c)} - \frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)}\right] > \alpha \frac{q_{ar} q_c}{q_{ar} + q_r \pi (1-q_c)}. \tag{A9}$$

The left-hand side is the gain from worsening the beliefs of non-receptive voters, and is obtained by differencing (A6) between $\hat{s}=1$ and $\hat{s}=0$. The right-hand side is the loss from improving the beliefs of receptive voters, and is obtained by differencing (A8) between $\hat{s}=1$ and $\hat{s}=0$. It is straightforward to check that this inequality yields a threshold $\bar{\alpha}(\pi)$, such that for $\alpha < \bar{\alpha}(\pi)$ the AR elite strictly prefers to criticize the politician. Moreover, for $\pi$ approaching 1, the term multiplying $1-\alpha$ on the left hand side approaches 1, while the term multiplying $\alpha$ on the right hand side approaches $\hat{q}_c < 1$, implying that for $\pi$ large enough, $\bar{\alpha}(\pi) > 0.5$. As a result, when $\pi$ is large, for $\alpha < 0.5$ we are always in the range corresponding to the simple propaganda equilibrium.

We now turn to show that the proposed equilibrium exists, separately in the ranges below and above $\bar{\alpha}(\pi)$.

3. Equilibrium existence for $\alpha < \bar{\alpha}(\pi)$

We establish that the simple propaganda profile is an equilibrium using backward induction. The R elite always reports truthfully after any history to minimize the lying cost. The AR elite, absent propaganda, will (for $\pi$ large) always send a bad message, because that minimizes the posterior of both the non-receptive voter by (A6) and the receptive voter by (A7). The AR elite,

following propaganda, will send a bad message because $\alpha < \alpha(\pi)$ means that (A9) holds. Thus, all elite types find it optimal to follow the strategies in the proposed profile.

We next consider the politician types. Start with the good R politician. For $\pi$ high, he expects to be praised by the R elite, and is thus getting a payoff close to the highest possible in the game. Since the cost of propaganda $f$ is bounded away from zero, for $\pi$ high he does not send propaganda.

Consider the bad R politician. He will prefer to send propaganda if and only if

$$\alpha \left[ \pi \left( \frac{q_{ar} q_c}{q_{ar} + q_r \pi (1 - q_c)} - \frac{(1-\pi) q_c}{(1-\pi) q_c + \pi (1 - q_c) \beta} \right) + (1 - \pi) \cdot \frac{-\pi q_c}{\pi q_c + (1-\pi)(1-q_c)\beta} \right] > f. \tag{A10}$$

The left hand side measures the expected gain from propaganda. Propaganda only has an effect on the share of voters $\alpha$ who observe propaganda. For these voters, if the elite sends a bad message (with probability $\pi$), then propaganda changes beliefs to the value given in (A8) for $s = 0$, from the value given in (A7) for $s = 0$. This explains the first term. If the elite sends a good message (with probability $1 - \pi$), then propaganda changes beliefs to the value given in (A8) for $s = 1$, which is zero, from the value given in (A7) for $s = 1$. This explains the second term.

Observe that the limit of the left-hand side, as $\pi$ goes to one, is

$$\alpha \frac{q_{ar} q_c}{q_{ar} + q_r (1 - q_c)} = \alpha \hat{q}_c.$$

Thus, Assumption 2 implies that for $\pi$ sufficiently large the bad R politician will prefer to send propaganda.

Consider next the good and the bad AR politicians. They prefer to send propaganda if

$$\alpha \left[ \frac{q_{ar} q_c}{q_{ar} + q_r \pi (1 - q_c)} - \frac{(1-\pi) q_c}{(1-\pi) q_c + \pi (1 - q_c) \beta} \right] > f. \tag{A11}$$

This is slightly different from condition (A10), because while in the R the elite sends a bad message only with probability $\pi$, in the AR it sends a bad message with probability 1. However, it remains true that in the limit as $\pi$ goes to one, the left-hand side converges to $\alpha \hat{q}_c$, so that Assumption 2 implies that the AR politicians too will prefer to send propaganda.

4. Equilibrium existence for $\alpha > \bar{\alpha}(\pi)$

We prove that there exists an equilibrium that has the complex propaganda profile: following propaganda, the AR elite sends a bad message after a bad signal and plays a mixed action after a good signal, while all other players follow the simple propaganda profile. We proceed by backward induction. As before, the R elite reports truthfully to avoid the lying cost.

Now consider the condition for the AR elite's indifference after propaganda. Suppose that the mixing probability of sending a good report after a good signal is $r$. Voters' average belief after a good report is given by

$$\begin{aligned}
&\bar{\mu}(\theta_c = 1|\hat{p} = 1, \hat{s} = 1) \\
&= \alpha\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 1, \theta_m = P) + (1-\alpha)\mu_{un}(\theta_c = 1|\hat{s} = 1) \\
&= \alpha\frac{q_{ar}q_c\pi r}{q_{ar}q_c\pi r + q_{ar}(1-q_c)(1-\pi)r + q_r(1-q_c)(1-\pi)} + (1-\alpha)\frac{\pi q_c}{\pi q_c + (1-\pi)(1-q_c)}.
\end{aligned}$$

The first term follows from Bayesian updating by a receptive voter influenced by propaganda, who accounts for the fact that the AR elite randomizes after a good signal. This means that a good politician can be consistent with a good report and propaganda, if reality is AR, the elite's signal was good, and the elite randomized to follow that signal, explaining the numerator. However, the profile of propaganda and a good signal can also emerge in the AR if the politician is bad, the elite's signal was incorrect (good), and the elite randomized to follow it; and in the R if the politician is bad and the elite's signal was incorrect. This explains the denominator. The second term is the belief of the unreceptive voter and comes from (A6).

Voters' average belief after a bad report is given by

$$\begin{aligned}
&\bar{\mu}(\theta_c = 1|\hat{p} = 1, \hat{s} = 0) \\
&= \alpha\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 0, \theta_m = P) + (1-\alpha)\mu_{un}(\theta_c = 1|\hat{s} = 0) \\
&= \alpha\frac{q_{ar}\pi q_c(1-r) + q_{ar}(1-\pi)q_c}{q_{ar}\pi q_c(1-r) + q_{ar}(1-\pi)q_c + q_{ar}\pi(1-q_c) + q_{ar}(1-\pi)(1-q_c)(1-r) + q_r\pi(1-q_c)} \\
&+ (1-\alpha)\frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)}.
\end{aligned}$$

The first term is the update of the receptive voter after propaganda and a bad message. This profile can be consistent with a good politician if reality is AR, and either the elite's signal was good and he randomized not to follow it, or was bad in which case he always follows it, explaining

the numerator. However, this profile can also arise: in the AR if the politician is bad and the elite's signal was correct (bad); in the AR if the politician is bad, the elite's signal was incorrect (good) but she randomized to send a bad message; and in R if the politician is bad and the elite's signal was correct. This explains the denominator. The second term is the update of the unreceptive voter and comes from (A6).

It is tedious but straightforward to compute the partial derivatives of these beliefs with respect to $r$, and to sign them for $r \in [0,1]$:

$$\frac{\partial \bar{\mu}(\theta_c = 1 | \hat{p} = 1, \hat{s} = 1)}{\partial r} = \frac{\pi(1-\pi)q_c(1-q_c)q_{ar}q_r}{[q_{ar}q_c\pi r + q_{ar}(1-q_c)(1-\pi)r + q_r(1-q_c)(1-\pi)]^2} > 0$$

and

$$\frac{\partial \bar{\mu}(\theta_c = 1 | \hat{p} = 1, \hat{s} = 0)}{\partial r} =$$
$$- \frac{q_{ar}q_c(1-q_c)[\pi^2 q_r + q_{ar}(2\pi - 1)]}{[q_{ar}\pi q_c(1-r) + q_{ar}(1-\pi)q_c + q_{ar}\pi(1-q_c) + q_{ar}(1-\pi)(1-q_c)(1-r) + q_r\pi(1-q_c)]^2} < 0.$$

Thus, for $r \in [0,1]$ the mean belief after praise is strictly increasing, while the mean belief after criticism is strictly decreasing in $r$.[1] Direct substitution implies that for $r = 0$ the former mean belief is smaller than or equal than the latter mean belief if and only if

$$\alpha \frac{q_{ar}q_c}{q_{ar} + q_r\pi(1-q_c)} \geq (1-\alpha)\left[\frac{\pi q_c}{\pi q_c + (1-\pi)(1-q_c)} - \frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)}\right]$$

which is the opposite of (8), implying that it holds since we assume $\alpha > \bar{\alpha}(\pi)$. For $r = 1$ the former mean belief is larger than the latter mean belief if and only if

$$\alpha \frac{q_{ar}q_c\pi}{q_{ar}q_c\pi + q_{ar}(1-q_c)(1-\pi) + (1-q_{ar})(1-q_c)(1-\pi)} + (1-\alpha)\frac{q_c\pi}{q_c\pi + (1-q_c)(1-\pi)}$$
$$> \alpha \frac{q_{ar}q_c(1-\pi)}{q_{ar}q_c(1-\pi) + q_{ar}(1-q_c)\pi + (1-q_{ar})(1-q_c)\pi}$$
$$+ (1-\alpha)\frac{q_c(1-\pi)}{q_c(1-\pi) + (1-q_c)\pi}.$$

To evaluate this inequality, note that (i) the left-hand side is increasing in $\pi$ and (ii) we obtain the right-hand side from the left-hand side by replacing $\pi$ with $1-\pi$. Thus, the inequality follows from

[1] In fact, these expressions also show that as $\pi$ approaches 1, the first partial derivative approaches zero, while the second remains bounded away from zero, which is a property we will use later.

$\pi > 1 - \pi$ which holds since $\pi > 0.5$. It follows that there is a unique mixing probability $r$ that makes the AR elite indifferent after propaganda between praise and criticism. This establishes the optimality of the AR elite's behavior.

To establish optimality for the politician, we first need to characterize $r$ for $\pi$ approaching one. To do this, consider the indifference condition

$$\alpha\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 1, \theta_m = P) + (1-\alpha)\mu_{un}(\theta_c = 1|\hat{s} = 1)$$
$$= \alpha\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 0, \theta_m = P) + (1-\alpha)\mu_{un}(\theta_c = 0|\hat{s} = 0).$$

Combining this condition with the fact that $\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 0, \theta_m = P) \leq 1$ allows us to derive the inequality

$$\frac{q_{ar}q_c\pi r}{q_{ar}q_c\pi r + q_{ar}(1-q_c)(1-\pi)r + (1-q_{ar})(1-q_c)(1-\pi)}$$
$$\leq 1 - \frac{1-\alpha}{\alpha}\left(\frac{q_c\pi}{q_c\pi + (1-q_c)(1-\pi)} - \frac{q_c(1-\pi)}{q_c(1-\pi) + (1-q_c)\pi}\right)$$

which can be further rewritten as

$$\frac{q_{ar}(1-q_c)(1-\pi)r + q_r(1-q_c)(1-\pi)}{q_{ar}q_c\pi r + q_{ar}(1-q_c)(1-\pi)r + (1-q_{ar})(1-q_c)(1-\pi)}$$
$$\geq \frac{1-\alpha}{\alpha}\left(\frac{q_c\pi}{q_c\pi + (1-q_c)(1-\pi)} - \frac{q_c(1-\pi)}{q_c(1-\pi) + (1-q_c)\pi}\right)$$

and, increasing the left-hand-side, as

$$\frac{(1-q_c)(1-\pi)}{q_{ar}q_c\pi r} \geq \frac{1-\alpha}{\alpha}\left(\frac{q_c\pi}{q_c\pi + (1-q_c)(1-\pi)} - \frac{q_c(1-\pi)}{q_c(1-\pi) + (1-q_c)\pi}\right).$$

This implies

$$\frac{1}{r} \geq \frac{q_{ar}q_c\pi}{(1-q_c)(1-\pi)}\frac{1-\alpha}{\alpha}\left(\frac{q_c\pi}{q_c\pi + (1-q_c)(1-\pi)} - \frac{q_c(1-\pi)}{q_c(1-\pi) + (1-q_c)\pi}\right).$$

This expression implies that, uniformly in $\alpha \geq \bar{\alpha}(\pi)$, as $\pi$ goes to one $r$ goes to zero.

With this result in hand, we can establish the optimality of the politician's proposed behavior for $\pi$ large. Consider the limit, as $\pi$ approaches one, of the mean voter belief after a bad report and propaganda. Given that $r$ goes to zero uniformly in $\alpha$, the limit, uniformly in $\alpha \geq \bar{\alpha}(\pi)$, is

$$\alpha\frac{q_{ar}q_c}{q_{ar} + q_r(1-q_c)} = \alpha\hat{q}_c.$$

It follows that under Assumption 2, for $\pi$ sufficiently large (independently of $\alpha \geq \bar{\alpha}(\pi)$) the R politician and both AR politicians will find it optimal to send propaganda. And the good R politician still prefers not to, because absent propaganda his payoff approaches the possible maximum (as $\pi$ goes to one), while with propaganda he pays a non-negligible cost $f$. We conclude that the mixed equilibrium exists for $\pi$ sufficiently high and $\alpha \geq \bar{\alpha}(\pi)$.

We conclude this existence proof by showing that this mixed equilibrium respects the lying cost. For any small lying cost $\chi$, the indifference condition is distorted by a small additive constant. The argument for the mean beliefs after praise and propaganda are strictly increasing and decreasing, respectively, continues to be valid. Thus, it remains true that for any $\alpha > \bar{\alpha}(\pi)$, for a lying cost small enough there exists a mixing probability that ensures indifference. Moreover, as the lying cost approaches zero, the implied mixing probability approaches that corresponding to a zero lying cost. This follows from the observation made in footnote 1 that the slope in $r$ of the belief after criticism remains bounded away from zero (while that after praise approaches zero), which implies that a small wedge between the two beliefs can be compensated for by a small change in $r$. It follows that for any given $\pi$, when the lying cost is sufficiently small, the payoffs of all parties are going to be close to those in the original equilibrium. Now in the original equilibrium the AR elite after a good signal is indifferent and mixes, the AR elite after a bad signal is indifferent but sends a bad message, and all other parties strictly prefer their equilibrium action. In the new profile of the game with lying cost the AR elite after a good signal is indifferent by construction; therefore the AR elite after a bad signal—given the lying cost—strictly prefers to send a bad message and does so, generating the same action as in the original game. By continuity all other parties have a strict preference to take their prescribed action. Thus this new profile is indeed close to the original profile and is an equilibrium of the game with a small lying cost.

5. Equilibrium selection for $\alpha < \bar{\alpha}(\pi)$

We use the politician pure refinement, that is, we only consider equilibria in which all politician types play pure strategies. Our goal is to identify the politician pure equilibrium which is optimal for the R politician. Our proof strategy is to check for all possible pure strategy profiles of all politician types. We go through the politician types one-by-one.

[Good R politician.] In any equilibrium, for $\pi$ sufficiently high, the good R politician never sends propaganda. This is because with a high $\pi$ probability his good type is revealed, in which case his utility is maximized, and propaganda has cost $f$ bounded away from zero.

[Bad R politician.] Any equilibrium in which the bad R politician does not send propaganda, or is indifferent to not sending propaganda, is dominated by our preferred equilibrium. This is because payoffs absent propaganda would be the same for the bad R politician in all equilibria; and in our preferred equilibrium the bad R politician strictly prefers to send propaganda, implying that he earns a higher payoff from doing so. Thus, it suffices to consider equilibria in which the R politician strictly prefers to send propaganda if he is bad.

[Good AR politician.] Suppose that the good AR politician does not send propaganda. This means that propaganda reveals that the politician is bad. Hence, propaganda cannot be worthwhile for the bad R politician, a contradiction. Thus, the good AR politician must send propaganda.

[Bad AR politician.] This is the last step in the proof, but it is a complicated step. It will be useful to start this step by considering the behavior of the AR elite. It is immediate that after no propaganda, the AR elite always sends a bad message. After propaganda, we need to consider what the AR elite does as a function of the signal she receives.

- Propaganda and a good signal. Then, the AR elite must send a bad message with positive probability. Otherwise, a bad message after propaganda will prove that the elite received a bad signal (both in the R and the AR), implying that (for $\pi$ high) the bad R politician will not want to send propaganda.

- Propaganda and a bad signal. Then, the AR elite must send a bad message with probability one. This follows because of the infinitesimal lying cost: since she weakly prefers a bad message after a good signal, when it is a lie, she must strictly prefer it after a bad signal, when it is not a lie.

It follows that the AR elite always sends a bad message after a bad signal, but has two qualitatively different strategies after a good signal: she either randomizes or sends a good message.

We now return to the strategy of the bad AR politician. We have four subcases: whether the

bad AR politician does not or does send propaganda, and whether the AR elite after a good signal randomizes or always sends a bad message.

Subcase (1i): The bad AR politician does not send propaganda, and conditional on propaganda, after both signal realizations (good or bad) the AR elite criticizes. Since in this profile the bad AR politician is always criticized, he has even stronger incentives than the bad R politician to send propaganda. Indeed, the latter can sometimes get a good message, which reduces the payoff of propaganda and increases the payoff of no propaganda. Since the bad R politician prefers propaganda, so should the bad AR politician, a contradiction.

Subcase (1ii): The bad AR politician does not send propaganda, and conditional on propaganda, the AR elite mixes after a good signal and sends a bad message after a bad signal. In this subcase, ignoring the infinitesimal lying cost, the AR elite must be indifferent between the two messages, implying that the voters' mean beliefs after propaganda and a good message must be the same as after propaganda and a bad message. But then any politician type has the same payoff from propaganda: they may face a different distribution of elite messages, but mean beliefs after propaganda and any elite message at the same. Moreover, not sending propaganda is worse for the bad AR politician than for the bad R politician, since the latter sometimes gets a good message. Thus, propaganda should generate a strictly higher payoff gain for the bad AR politician than for the bad R politician, and since the latter prefers it, so should the former. This is a contradiction.

Subcase (2i): Both the good and the bad AR politician sends propaganda, and conditional on propaganda, after both signal realizations (good or bad) the AR elite criticizes. This is the structure of our preferred equilibrium, and the existence proof shows that given $\alpha < \alpha(\pi)$ this is an equilibrium.

Subcase (2ii): Both the good and the bad AR politician sends propaganda, and conditional on propaganda, the AR elite mixes after a good signal and sends a bad message after a bad signal. In this candidate equilibrium, relative to our preferred equilibrium, propaganda and a bad message are worse while propaganda and a good message are better for the politician. Indeed, in this candidate equilibrium propaganda and a bad message are stronger evidence that the politician is bad (because they arise with a lower probability when the AR politician is good) while propaganda

and a good message are weaker evidence that the politician is bad (because they arise with a higher probability when the AR politician is good). Since $\alpha < \alpha(\pi)$ ensures that the AR elite prefers to criticize in our preferred equilibrium, it follows that she will strictly prefer to criticize in this candidate equilibrium, a contradiction.

6. Equilibrium selection for $\alpha > \bar{\alpha}(\pi)$

Since in the proof for $\alpha < \alpha(\pi)$ we used that $\alpha < \alpha(\pi)$ only in subcases (2i) and (2ii), the previous steps continue to hold. It follows that in any equilibrium meeting our selection criteria, both the good and the bad AR politicians send propaganda, the elite after a bad signal sends a bad message, and the elite after a good signal either mixes or sends a bad message. Since $\alpha > \alpha(\pi)$, the elite after a good signal cannot be sending a bad message. Thus, she must be mixing. The existence proof characterizes the unique mixing probability that makes this profile an equilibrium.

### A.5 Alternative specification of voter types

Our assumption that the unreceptive voter does not even observe propaganda is stark. To relax this assumption, we introduce the following potential subtypes of the unreceptive voter.

**Definition 3.** We introduce two types of the unreceptive voter:

1. A voter is *naive unreceptive* if he does not observe propaganda,
2. A voter is *sophisticated unreceptive* if he does observe propaganda (and updates from it), but propaganda does not change his prior beliefs.

Thus, the unreceptive voter of our baseline model is naive unreceptive. Using these types, we introduce two modifications of our baseline model, both of which permit the mass of unreceptive voters to at least partially update from propaganda. As in the baseline model, we assume that a share $\alpha$ of voters are receptive to propaganda.

**Definition 4.** We introduce two alternative specifications of voter types.

- *Partially naive electorate.* A share $\alpha_s$ of voters are sophisticated unreceptive, and the rest of the unreceptive voters $(1 - \alpha - \alpha_s)$ are naive unreceptive.

- *Misspecified AR.* As in the baseline model, a share $1-\alpha$ of voters are sophisticated unreceptive. But the AR elite and the AR politician believe that all unreceptive voters are naive.

With a partially naive electorate, the share $1-\alpha$ of unreceptive voters consist of a mix of sophisticates and naifs and hence, while not being affected by propaganda in terms of their prior, on average they update partially from it. With a misspecified AR, in reality unreceptive voters update from propaganda, but the AR elite wrongly believes that they do not.

**Corollary 1.** *Under Assumptions 1 and 2, the statement in Proposition 1 applies if*

1. *We have a partially naive electorate and* $\alpha < (1-\alpha_s)/2$.

2. *We have a misspecified AR and* $\alpha < 0.5$.

**Proof of Corollary 1.**

The claim here is that in these settings with three rather than two voter types, the unique PPO equilibrium strategies of the politician and the elite are the same as in Proposition 1. We organize the proof the following way. First, we characterize the beliefs of all three voter types (receptive, sophisticated unreceptive, and naive unreceptive) at the end of stage 2 after observing the simple propaganda action profile by the politician and the elite. Second, we establish that in either model variant, there is an equilibrium which takes the simple propaganda form. Finally, we show that the simple propaganda equilibrium is the unique PPO equilibrium.

1. Voter beliefs

The beliefs of the naive unreceptive voter are still governed by (A6). However, the beliefs of the sophisticated unreceptive voter are given by

$$\mu_{un,so}(\theta_c = 1|\hat{p},\hat{s}) = (1-\hat{p})\left[\hat{s}\frac{\pi q_c}{\pi q_c + (1-\pi)(1-q_c)\beta} + (1-\hat{s})\frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)\beta}\right]. \quad \text{(A12)}$$

Observing propaganda, the sophisticated unreceptive voter learns immediately that the politician is bad, since in reality only bad politicians send propaganda. Thus, the expression is zero if $\hat{p} = 1$. In the absence of propaganda, the sophisticated unreceptive voter is similar to a receptive voter: both know that reality is R and both update from the absence of propaganda. Thus, after the history of no propaganda ($\hat{p} = 0$), the sophisticated unreceptive voter's beliefs are identical to those of

the receptive voter, (A7). Finally, the receptive voter's beliefs are formed the same way as in the baseline model, and are given by equations (A7) and (A8).

2. Existence

Similarly to the proof of Proposition 1, we use backward induction. We start with actors who have the same dominant strategies under the two model variants. The R elite always tells the truth, for the same reason as in the baseline model. Absent propaganda, for $\pi$ sufficiently large, the AR elite will always criticize: since there is no propaganda, the elite's message is taken at face value by all voter types. Finally, the good R politician never sends propaganda for the same reason as before.

The rest of the existence proof is different under the two model variants.

*Case 1—Partially naive electorate.* Consider the incentive compatibility constraint of the AR elite after propaganda. They send a bad message if and only if

$$(1-\alpha-\alpha_s)\left[\frac{\pi q_c}{\pi q_c+(1-\pi)(1-q_c)}-\frac{(1-\pi)q_c}{(1-\pi)q_c+\pi(1-q_c)}\right]>\alpha\frac{q_{ar}q_c}{q_{ar}+q_r\pi(1-q_c)}. \tag{A13}$$

To see the logic, note that after propaganda, the sophisticated voter knows the politician's type, so the elite's message only influences the naive unreceptive voters and the receptive voters. Then, the condition is very similar to the analogous condition in the baseline model, (A9). The left-hand-side measures the gain to the AR elite from worsening the beliefs of the naive unreceptive voters, and is different in that there are now $1-\alpha-\alpha_s$ naive unreceptive voters; the right-hand-side measures the loss to the AR elite from improving the beliefs of the receptive voter, and is identical to that in (A9). For $\pi$ large enough, the left-hand side converges to $1-\alpha-\alpha_s$, while the right-hand side converges to $\alpha\hat{q}_c$. Since $\hat{q}_c<1$, our assumption that $\alpha<(1-\alpha_s)/2$ ensures the condition.

Now, consider the incentive compatibility of the politician. The bad R politician sends propaganda if and only if

$$\begin{aligned}&\alpha\left[\pi\left(\frac{q_{ar}q_c}{q_{ar}+q_r\pi(1-q_c)}-\frac{(1-\pi)q_c}{(1-\pi)q_c+\pi(1-q_c)\beta}\right)+(1-\pi)\cdot\frac{-\pi q_c}{\pi q_c+(1-\pi)(1-q_c)\beta}\right]\\&+\alpha_s\left[\pi\cdot\frac{-(1-\pi)q_c}{(1-\pi)q_c+\pi(1-q_c)\beta}+(1-\pi)\cdot\frac{-\pi q_c}{\pi q_c+(1-\pi)(1-q_c)}\right]>f.\end{aligned} \tag{A14}$$

As before, the left hand side measures the gain from propaganda. However, now propaganda has an effect not only on the receptive voter but also on the sophisticated unreceptive voter. The effect

on the former is given by the first term and is the same as in (A10). The effect on the latter is given by the second term, and is new. In this term, $\alpha_s$ is the mass of sophisticated unreceptive voters. To interpret the expression in brackets, not that if the elite sends a bad message (with probability $\pi$), then propaganda changes beliefs from the value given in (A12) for $\hat{s} = 0$ to zero, since the sophisticated voter learns from propaganda that the politician is bad. If the elite sends a good message (with probability $1 - \pi$), then propaganda changes beliefs from the value given in (A12) for $\hat{s} = 1$ to zero for the same reason.

We have a similar incentive compatibility condition for both the good and the bad AR politician:

$$\alpha_s \left[ \frac{-(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)\beta} \right] + \alpha \left[ \frac{q_{ar}q_c}{q_{ar} + q_r\pi(1-q_c)} - \frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)\beta} \right] > f. \tag{A15}$$

This expression is simpler than (A14) because the AR politician expects criticism with certainty.

The left hand side of both (A14) and (A15), as $\pi$ goes to one, converges to $\alpha\hat{q}_c$, so under Assumption 2, for $\pi$ sufficiently large, the bad R and the good and the bad AR politician will prefer to send propaganda.

*Case 2—Misspecified AR.* In this model variant the AR elite believes that the true model is the baseline model. It follows that the AR elite sends a bad message if and only if the original (A9) condition holds. We know from the proof of Proposition 1 that this condition holds for $\alpha < 0.5$.

Consider next the incentive compatibility of the politician. The bad R politician sends propaganda if and only if

$$\begin{aligned}(1-\alpha) &\left[ \pi \cdot \frac{-(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)\beta} + (1-\pi) \cdot \frac{-\pi q_c}{\pi q_c + (1-\pi)(1-q_c)} \right] \\ + \alpha &\left[ \pi \left( \frac{q_{ar}q_c}{q_{ar} + q_r\pi(1-q_c)} - \frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)\beta} \right) + (1-\pi) \cdot \frac{-\pi q_c}{\pi q_c + (1-\pi)(1-q_c)\beta} \right] > f.\end{aligned} \tag{A16}$$

This condition is similar to (A14), since as in Case 1, propaganda affects both the receptive voter and the sophisticated unreceptive voter. However, now the mass of sophisticated unreceptive voters is $1 - \alpha$. The AR politician, whether good or bad, believes that unreceptive voters are naive, i.e., that the model is the same as our baseline model, and accordingly sends a bad message if (A11) holds.

The left hand side of both (A16) and (A11) converges, as $\pi$ goes to one, to $\alpha\hat{q}_c$. Thus, Assumption 2 implies that for $\pi$ sufficiently large, the bad R politician, and the good and the bad AR politician will prefer to send propaganda.

3. Equilibrium selection

Similarly to the proof of Proposition 1, we use the politician pure refinement and only consider candidate equilibria in which all politician types play pure strategies. As we established above, in both Case 1 and Case 2, in any equilibrium, for $\pi$ sufficiently high, the R elite is truthful and the good R politician never sends propaganda. It also follows that for $\pi$ sufficiently high the AR elite always criticizes. In Case 1 this follows because the benefit of criticism, coming from persuading naive unreceptive voters, converges to $1-\alpha-\alpha_s$, while the cost, coming from changing the beliefs of receptive voters, is at most $\alpha$, and we assume $\alpha < (1-\alpha_s)/2$. In Case 2, it follows because the AR elite imagines the world to be as in our baseline model, where we already established this point.

Our preferred equilibrium is better than any equilibrium in which the bad R politician refrains from propaganda, because in our equilibrium the bad R politician strictly prefers to send propaganda and thus benefits from doing so. It follows that in any PPO equilibrium the bad R politician sends propaganda. Parallel to our baseline model, there can be no PPO equilibrium in which the good AR politician does not send propaganda, because then the receptive voter would learn from propaganda that the politician is bad, destroying the gain from propaganda to the bad R politician. Finally, the bad AR politician must also send propaganda, since he faces a worse portfolio of elite messages (always criticism) than the bad R politician (often criticism).

## A.6 Proofs of Corollaries

**Proof of Corollary 1.** Under Assumptions 1 and 2, Proposition 1' implies that for $\pi > \bar{\pi}$ the bad R politician strictly prefers to send propaganda, which implies that

$$E[\bar{\mu}(\theta_c = 1|\hat{p} = 1, \hat{s})|\theta_c = 0] - E[\bar{\mu}(\theta_c = 1|\hat{p} = 0, \hat{s})|\theta_c = 0] > f$$

and hence that the left-hand-side is positive.

**Proof of Corollary 2.** Under Assumptions 1 and 2, Proposition 1' implies that depending on the value of $\alpha$ the unique PPO equilibrium takes either the simple or the complex propaganda form. In the simple propaganda equilibrium, after a history of propaganda, the beliefs of the receptive voter are given by equation (7), which immediately implies that a bad message improves the perception of the politician among receptive voters. In the complex propaganda equilibrium, the AR elite, after observing a good signal, is indifferent between reporting good and reporting bad. Since sending a bad message (relative to a good message) harms the politician's perceived competence among non-receptive voters, to make the elite indifferent, that bad message must improve the politician's perceived competence among receptive voters.

**Proof of Corollary 3.** Under Assumptions 1 and 2, by Proposition 1', the unique PPO equilibrium takes either the simple or the complex propaganda form. In either equilibrium, the receptive voter's posterior about the AR, after updating from propaganda, but before updating from the elite's message, is

$$\mu_{rec}(AR|\hat{p}=1,\theta_m=P)=\frac{q_{ar}}{1-q_rq_c}.$$

This follows because the receptive voter has a new prior $q_{ar}$, and from this prior and the observation of propaganda, he infers that the state $(\theta_r,\theta_c)$ is either (R,bad), or (AR, bad), or (AR, good). The unconditional joint probability of the AR states is $q_{ar}$, but the total probability of these three states is just $1-q_rq_c$. Observe that this expression is larger than $q_{ar}$.

Now consider the voter's posterior after observing the elite's message as well. In a normal Bayesian setting, the expected value of that posterior would equal the belief we just computed. That is not the case here, because the voter misunderstands the distribution of the elite's signals. Nevertheless, for $\pi$ approaching one the expected posterior will equal the above expression. This is because for $\pi$ approaching one, the voter expects that message to be almost always negative (since even in the complex propaganda equilibrium $r$ approaches zero) and hence his posterior after a bad message will be close to his posterior after propaganda. Moreover, it is also the case in the objective reality that the elite's message is almost always negative, implying that the objective expected posterior of the receptive voter will also be close to his post-propaganda posterior.

**Proof of Corollary 4.** The proof is organized in numbered steps.

1. Voter beliefs in the no propaganda profile

We say that a strategy profile has the *no propaganda form* if no politician type sends propaganda and all elite types report truthfully. We start by computing voter beliefs in this profile. First consider the history of no propaganda. Since no politician type sends propaganda in the equilibrium profile, then the receptive voter does not update form the absence of propaganda. Therefore, both voter types have the same posterior as the unreceptive voter in Proposition 1 given by equation (A6). Next, consider the off-equilibrium history of propaganda. The receptive voter attributes propaganda to a tremble, and since he knows that in both realities the elite is trustworthy he forms the same beliefs as the unreceptive voter in equation (A6).

2. Equilibrium existence

We establish that the no propaganda profile is an equilibrium using backward induction. The R elite reports truthfully after any history to minimize lying cost. The AR elite will report truthfully after any history too. This is because in the proposed equilibrium voters form beliefs according to equation (A6), which increases in $\hat{s}$ for large $\pi$, and the AR elite wants to maximize the average voter's posterior after a good signal and minimize it after a bad signal. In stage 1, no politician types chooses to send propaganda, since propaganda is costly but does not change any voter's posterior beliefs about the politician's type.

3. Equilibrium selection

Here we prove that the no propaganda equilibrium is the unique PPO equilibrium. The good R politician does not send propaganda in any equilibrium, since propaganda is costly and, as $\pi$ approaches one, his good type is almost completely revealed by the elite message.

The good AR politician does not send propaganda either. To see why, suppose he does, and consider a history without propaganda in the AR. Since there was no propaganda, the voter remains normal and follows the elite's signal.[2] Given this, the AR elite, who prefers to keep the good AR

[2] We can compute the receptive voter's belief explicitly. Denoting by $y$ whether in the candidate profile the bad R politician sends propaganda, it has the following form, which is increasing in $\hat{s}$ for $\pi$ large

$$\mu_{rec}(\theta_c = 1|\hat{p} = 0, \hat{s}, \theta_m = N) = \hat{s}\frac{\pi q_c}{\pi q_c + (1-\pi)(1-q_c)(1-y+y\beta)} + (1-\hat{s})\frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)(1-y+y\beta)}.$$

politician, will send a good message after a good signal. Thus, the AR politician will get a payoff near his first best for $\pi$ close to one. As a result, he does not engage in costly propaganda.

Finally, since no good politician type uses propaganda, no bad type uses it either to avoid revealing his type.

**Proof of Corollary 5.** We focus on the $\alpha < 0.5$ case in which we have the simple propaganda equilibrium. Begin with claim 1. In this equilibrium, the persuaded voter's posterior after observing propaganda and criticism is

$$\mu_{rec}(AR|\hat{p}=1,\hat{s}=1,\theta_m=P)=\frac{q_{ar}}{q_{ar}+q_r\pi(1-q_c)}.$$

This follows because in the AR the voter expects propaganda and criticism explaining the numerator; and in addition in the R he expects it if the politician is bad and the elite's signal is correct, explaining the denominator. This expression is clearly increasing in $q_c$.

Now consider claim 2. The difference between the unreceptive voter's beliefs with versus without propaganda is zero. To express the difference between the receptive voter's beliefs with versus without propaganda, note that his expected belief, *absent propaganda*, when the politician is bad, is

$$E[\mu_v(\theta_c=1|\hat{p}=0,\theta_m=N)|\theta_c=0]=\pi\frac{(1-\pi)q_c}{(1-\pi)q_c+\pi(1-q_c)}+(1-\pi)\frac{\pi q_c}{\pi q_c+(1-\pi)(1-q_c)}.$$

The receptive voter's expected belief, *with propaganda*, when the politician is bad, is

$$E[\mu_v(\theta_c=1|\hat{p}=1,\theta_m=P)|\theta_c=0]=\pi\frac{q_{ar}q_c}{q_{ar}+q_r\pi(1-q_c)}.$$

It is straightforward to verify that for $\pi$ large enough the difference is increasing in $q_c$. For a more direct argument, note that for $\pi$ converging to 1 the difference converges to $\hat{q}_c = q_c \cdot \hat{q}_{ar}$, which is strictly increasing in $q_c$ because both terms are increasing in $q_c$. Since all of the functions here are complex analytic, convergence of the function implies convergence of the derivative, implying that for $\pi$ large the difference is increasing in $q_c$.

## A.7 Microfundation of demand for alternative reality

*Model setup.* We present a simple model in which the receptive voter's propaganda-induced prior misbelief is endogenized. This model extends the probabilistic voting model presented in Appendix A.2. We assume that the receptive voter can only entertain the alternative reality proposed to him by the politician through propaganda, but—in the spirit of the idea of motivated beliefs—he can decide whether and to what extent to believe in it. More specifically, we assume that the receptive voter after observing propaganda, in stage 1, chooses how much prior belief $q_{ar} \in [0,1]$ to put on the AR presented by the politician. Receptive voter $i$'s objective function at this stage is

$$V_{rec,i} = \tilde{E}_{q_{ar}}[U_{rec,i}|\hat{p}=1] - E[C(\mu_{rec,i}(AR|\hat{p},\hat{s},q_{ar}))|\hat{p}=1]. \tag{A17}$$

We use the notation that $\tilde{E}_{q_{ar}}[.]$ computes the receptive voter's subjective expectation given his choice of prior belief $q_{ar}$. while $E[.]$ computes the objectively correct expectation. The first term is the receptive voter's subjective expectation of his utility defined by equation (A1). The second term represents the cost of holding incorrect posterior beliefs about the nature of reality in terms of subsequent outcomes. As common in the literature on motivated beliefs, this term is computed using the objectively correct expectations. We condition on $\hat{p}=1$ in both terms because we assume that the receptive voter chooses beliefs only when he receives propaganda, so that the expectations are taken over the realization of $\hat{s}$. As in Levy (2014), the cost is modeled in a reduced-form fashion; here it is a function of the voter's subjective posterior belief in the AR, $\mu_{rec,i}(AR|\hat{p},\hat{s},q_{ar})$, which depends on the voter's choice of prior $q_{ar}$. Assuming that the cost is a function of beliefs in the AR reflects that the cost is the result of taking bad personal decisions, such as not taking up vaccinations. The cost does not depend on beliefs about the politician's quality: voter $i$ understands that being infinitesimal he does not have an impact on the election outcome. This formulation is similar to Brunnermeier and Parker (2005) in that voters choose their optimal beliefs balancing between the benefit of optimism and the cost of worse decision making, but differs in that—for simplicity—we do not model the latter explicitly. We assume that $C'(\cdot)$ is convex, $C'(0)=0$, and $\lim_{x\to 1} C'(x) = \infty$.

*Analysis.* We assume that the conditions stated in Proposition 1 hold, and we will study the

equilibrium identified in that Proposition when $q_{ar}$ is endogenously chosen. We leave the question of whether other equilibria emerge for future work. Thus, in the derivations that follow we assume that strategies are as specified in our preferred equilibrium, and we will later confirm that those strategies continue to constitute an equilibrium.

Substituting in from (A1), the receptive voter's utility can be written as

$$U_{rec,i} = P(\hat{s},\hat{p}) \cdot c \cdot \mu_{rec,i}(\theta_c = 1|\hat{p},\hat{s},q_{ar}) + (1 - P(\hat{s},\hat{p}))c \cdot q_c^c + P(\hat{s},\hat{p}) \cdot \lambda. \tag{A18}$$

Here $P(\hat{s},\hat{p})$ is the probability that the incumbent wins the election, defined by equation (A4), except that here we made explicit its dependence on $\hat{p}$ and $\hat{s}$. Note that $P(\hat{s},\hat{p})$ is exogenous from the individual voter's perspective, because it is determined by other voters' beliefs about the AR. Since $q_c^c$ denotes the probability that the challenger is good, the first two terms measure the subjective expected value of the politician being good. The last term measures the subjective expected value of the politician being ideologically pro-voter.

We now turn to compute the subjective expected utility of voter $i$, that is, the subjective expected value of (A18). This requires some preliminaries. First, we note that although the maximization problem of voter $i$ is with respect to $q_{ar}$, we will find it convenient to treat it as a maximization problem with respect to $\hat{q}_{ar} = q_{ar}/(q_{ar} + q_r\pi(1 - q_c))$ which is the receptive voter's posterior belief in the AR after propaganda and criticism. This is an equivalent reformulation because $q_{ar}$ is a strictly monotone transformation of $\hat{q}_{ar}$. A consequence of this approach is that we will express terms of interest as functions of $\hat{q}_{ar}$.

Second, because the last two terms in (A18) will not contribute to the economics of the results, we introduce the notation

$$(1 - P(\hat{s},1))c \cdot q_c^c + P(\hat{s},1) \cdot \lambda = k(\hat{s}).$$

Third, to compute the subjective expected value of (A18), note that $\pi + q_{ar}(1 - \pi)$ is voter $i$'s subjective probability of observing elite criticism conditional on propaganda. We introduce the notation

$$\rho(q_{ar}) = \frac{\pi + q_{ar}(1 - \pi)}{\pi}$$

so that the subjective probability of observing criticism is $\pi\rho(q_{ar})$. We note for future reference that (i) with a slight abuse of notation we will often treat $\rho$ as a function of $\hat{q}_{ar}$, which is valid since $q_{ar}$ is a strictly monotonic transformation of $\hat{q}_{ar}$; (ii) as $\pi$ converges to one, $\rho(\hat{q}_{ar})$ converges to one uniformly in $\hat{q}_{ar}$; (iii) $\rho(\hat{q}_{ar})$ is a ratio of polynomials in $\hat{q}_{ar}$ and hence is (more precisely, can be extended into) a complex analytic function of $\hat{q}_{ar}$, which implies that as $\pi$ goes to one, all derivatives of $\rho$ with respect to $\hat{q}_{ar}$ converge to zero uniformly.

With these preliminaries, we can write the subjective expected value of (A18) as

$$\tilde{E}_{q_{ar}}[U_{rec,i}|\hat{p}=1] = \pi\rho(\hat{q}_{ar}) \cdot P(0,1) \cdot c \cdot \hat{q}_{ar} \cdot q_c + \pi\rho(\hat{q}_{ar}) \cdot [k(0) - k(1)] + k(1). \tag{A19}$$

Substituting back into the receptive voter's objective (A17) and noting that the cost is a function of the posterior AR belief $\hat{q}_{ar}$ yields

$$V_{rec,i} = \pi\rho(\hat{q}_{ar}) \cdot P(0,1) \cdot c \cdot \hat{q}_{ar} \cdot q_c + \pi\rho(\hat{q}_{ar}) \cdot [k(0) - k(1)] + k(1) - \pi C(\hat{q}_{ar}).$$

We maximize this with respect to $\hat{q}_{ar}$ by taking the first order condition, which yields

$$P(0,1)cq_c \cdot \rho(\hat{q}_{ar}) + \rho'(\hat{q}_{ar}) \cdot [P(0,1)cq_c\hat{q}_{ar} + k(0) - k(1)] = C'(\hat{q}_{ar}). \tag{A20}$$

This condition characterizes the equilibrium $\hat{q}_{ar}$. There are two important points to note. First, as mentioned above, when $\pi$ approaches one $\rho$ converges to one and $\rho'$ converges to zero, so that the first order condition converges to the much simpler form $P(0,1)cq_c = C'(\hat{q}_{ar})$. With that "approximate first-order condition" all the remaining analysis would follow easily. Much of the work below is showing that the results also obtain just before the limit. The second point to note is that the $P(0,1)$ on the left-hand-side also depends on $\hat{q}_{ar}$ in equilibrium (even if it was exogenous to voter $i$), because we know from equation (A4) that $P(0,1)$ is an increasing linear function of receptive voters' average posterior about the politician's type, that is $\mu_{rec}(\theta_c = 1|\hat{s}=0,\hat{p}=1) = \hat{q}_{ar}q_c$.

We now analyze this first-order condition, first under the (false) assumption that $\rho \equiv 1$ (which implies $\rho' \equiv 0$), and then properly. If $\rho \equiv 1$ were true, then we could directly trace the two sides of the approximate first-order-condition $P(0,1)cq_c = C'(\hat{q}_{ar})$ as a function of $\hat{q}_{ar}$. For $\hat{q}_{ar} = 0$, the left-hand-side is positive given the definition of $P(0,1)$, while the right-hand-size is zero by assumption. As $\hat{q}_{ar}$ increases, the left-hand-side traces out an increasing linear function, while the

right-hand-side an increasing convex function which asymptotes to infinity. Thus, there is a unique point of equilibrium.

Relaxing the false assumption, but taking $\pi$ large so that the deviations from the approximate first-order condition are small, it is still the case that the left-hand-side starts from a positive value while the right-hand-side starts from zero. Moreover, given the properties of $\rho$ highlighted above, the left-hand side remains arbitrarily close to a increasing linear function, and its derivative remains arbitrarily close to the positive constant slope of that function. The right-hand-side is still a smooth convex function, thus there is at least one point of intersection. Since the intersection requires that the right-hand-side "catches up" to the left-hand-side, in its neighborhood the slope of the right-hand-side must be strictly higher than the constant slope of $P(0,1)cq_c$. Thus, for $\pi$ sufficiently large, the slope of the right-hand-side will be strictly higher than the slope of the left-hand-side (which is arbitrarily close to the aforementioned constant). It follows that there cannot be a second intersection. We conclude that for $\pi$ large there is a unique $q_{ar}$. Moreover, the arguments also imply that as $\pi$ converges to 1, that $q_{ar}$ converges to the solution of the approximate first-order condition $P(0,1)cq_c = C'(q_{ar})$.

**Assumption 1.** Assumption 2 holds with the unique solution $q^*_{ar}$ of $P(0,1)cq_c = C'(\hat{q}_{ar})$.

**Proposition 1.** *Suppose that Assumptions 1 and 1 hold and $\alpha < 0.5$. For $\pi$ sufficiently large, the equilibrium of Proposition 1 remains an equilibrium with a unique endogenously chosen $q_{ar}$. Moreover, $q_{ar}$ is increasing in the voter's preference for an incumbent government $\lambda$ and in the voter's prior probability of a good politician $q_c$.*

**Proof of Proposition 1.** Consider the proposed equilibrium profile. In that profile, for $\pi$ large, the unique optimal $q_{ar}$ will satisfy Assumption 2. As a result, Proposition 1 shows that the profile is an equilibrium.

To establish the comparative statics, we need two preliminary steps. First, (A4) implies that the probability the incumbent politician remains in power, conditional on propaganda and criticism, is

$$P(0,1) = q \cdot c \cdot \left[\alpha \hat{q}_{ar} q_c + (1-\alpha)\frac{(1-\pi)q_c}{(1-\pi)q_c + \pi(1-q_c)}\right] + g(\lambda - c \cdot q_c^c) + 0.5,$$

where $\hat{q}_{ar}q_c$ is the receptive and $(1-\pi)q_c/[(1-\pi)q_c+\pi(1-q_c)]$ is the non-receptive voter's posterior belief. Second, if we rearrange equation (A20) and define

$$F \equiv \rho(\hat{q}_{ar}) \cdot P(0,1)cq_c + \rho'(\hat{q}_{ar}) \cdot [P(0,1)cq_c\hat{q}_{ar} + k(0) - k(1)] - C'(\hat{q}_{ar})$$

then

$$\frac{\partial F}{\partial \hat{q}_{ar}} = \rho'(\hat{q}_{ar}) \cdot P(0,1)c(1+q_c) + \rho''(\hat{q}_{ar}) \cdot [P(0,1)cq_c\hat{q}_{ar} + k(0) - k(1)] - C''(\hat{q}_a r),$$

and because when $\pi$ approaches one both $\rho'(\hat{q}_{ar})$ and $\rho''(\hat{q}_{ar})$ converge uniformly to zero, while $C''(\hat{q}_{ar})$ is by definition positive, we have that for $\pi$ large $\partial F/\partial \hat{q}_{ar} < 0$.

Given these preliminaries, we can apply the Implicit Function Theorem to obtain

$$\frac{\partial \hat{q}_{ar}}{\partial \lambda} = -\frac{\partial F/\partial \lambda}{\partial F/\partial \hat{q}_{ar}} = \frac{\rho(\hat{q}_{ar})\frac{\partial P(0,1)}{\partial \lambda} \cdot cq_c + \rho'(\hat{q}_{ar}) \cdot \frac{\partial}{\partial \lambda}[P(0,1) \cdot cq_c\hat{q}_{ar} + k(0) - k(1)]}{-\partial F/\partial \hat{q}_{ar}}$$
$$\xrightarrow[\pi \to 1]{\text{unif.}} \frac{\frac{\partial P(0,1)}{\partial \lambda} \cdot cq_c}{C''(\hat{q}_{ar})} = \frac{g \cdot cq_c}{C''(\hat{q}_{ar})} > 0$$

which implies that for $\pi$ large enough $\hat{q}_{ar}$ is increasing in $\lambda$. And then $q_{ar}$ is also increasing in $\lambda$ because $q_{ar}$ is an increasing transformation of $\hat{q}_{ar}$.

The intuition for the result is that $\lambda$ increases the probability $P(0,1)$ that the incumbent remains in power. Intuitively, the voter, who enjoys being optimistic, want to protect his positive belief about the politician who is likely to win the election.

The second comparative static also follows from the implicit function theorem:

$$\frac{\partial \hat{q}_{ar}}{\partial q_c} = -\frac{\partial F/\partial q_c}{\partial F/\partial \hat{q}_{ar}} = \frac{\rho(\hat{q}_{ar})\frac{\partial}{\partial q_c}[P(0,1)cq_c] + \rho'(\hat{q}_{ar})\frac{\partial}{\partial q_c}[P(0,1)cq_c\hat{q}_{ar} + k(0) - k(1)]}{-\partial F/\partial \hat{q}_{ar}}$$
$$\xrightarrow[\pi \to 1]{\text{unif.}} \frac{\frac{\partial}{\partial q_c}[P(0,1)cq_c]}{C''(\hat{q}_{ar})} = \frac{c[g \cdot c \cdot \alpha \hat{q}_{ar}q_c + P(0,1)]}{C''(\hat{q}_{ar})} > 0$$

which proves the result. The intuition here operates through two channels. First, $q_c$ directly increases the benefit of believing in the alternative reality, since the AR allows the voter to maintain the pleasurable prior belief $q_c$ that the politician is good. Second, $q_c$ increases the incumbent's probability of reelection; and the voter prefer to maintain a favorable opinion about the likely winner of the election.

*Implications.* We show that even with the endogenous demand for misbeliefs the predictions of Corollaries 1-5 continue to hold for the equilibrium identified in Proposition 1. For Corollaries 1-3 this is immediate, since they characterize properties of the equilibrium, and the equilibrium has the same form as in the basic model. Corollary 4 follows because for any endogenously chosen value of $q_{ar} < 1$ the current proof applies. Finally, the result of Corollary 5 is strengthened, because a higher $q_c$ also increases $q_{ar}$ through the demand side, acting to further amplify the belief in the AR.

## A.8 Proof of Proposition 2

Denote the lying cost AR by AR1 and the conspiracy AR by AR2.

Case 1: $\chi_f < (1-2\alpha)/N$.

*Behavior in any equilibrium.* We begin by characterizing the behavior of some actors in any large-$\pi$ equilibrium. Since the R elite's reputation costs are prohibitively large, the R elite is truthful in any profile. Given this, for $\pi$ large enough, the good R politician does not send propaganda.

Fix an equilibrium and consider a member $j$ of the AR1 elite after some history of propaganda $\hat{p}$. The impact on $\hat{\mu}$ of reporting good rather than bad after a good signal is

$$\frac{(1-\alpha)}{N} \cdot [\mu_{un,i(j)}(\hat{s}_j = 1) - \mu_{un,i(j)}(\hat{s}_j = 0)] + \frac{\alpha}{N} \cdot [\mu_{rec,i(j)}(\hat{p}, \hat{s}_j = 1) - \mu_{un,i(j)}(\hat{p}, \hat{s}_j = 0)].$$

In the limit as $\pi$ approaches one the elite signal becomes perfectly informative and the first term approaches $(1-\alpha)/N$. The second term, since beliefs are always between zero and one, is always bounded from below by $-\alpha/N$. Thus, as long as

$$\frac{1-\alpha}{N} - \frac{\alpha}{N} > \chi_f$$

holds, for $\pi$ large enough elite member $j$—who cares about reducing $\bar{\mu}$ but has a cost $\chi_f$ from lying—will report bad after a good signal. Since we are in Case 1, this condition holds. Thus, the AR1 elite always criticizes after a good signal. Since after a bad signal the gain from criticism is the same and the cost of criticism (relative to praise) becomes $-\chi_f$, the AR1 elite always criticizes after a bad signal as well.

Consider the AR2 elite. Since $N > 1$, we have $1 - 2\alpha > \chi_f$, and an analogous argument shows that the AR2 elite (as $\pi$ approaches 1), when reporting bad rather than good after a good signal, gains $1 - \alpha$ from unreceptive voters but loses at most $\alpha$ from receptive voters. Thus, the AR2 elite always criticizes after a good signal; and then it always criticizes after a bad signal as well.

*Existence of candidate equilibrium.* We now show that the following strategy profile is an equilibrium: the R elite is truthful; the good R politician does not send any propaganda; the bad R politician sends AR1; both AR politicians send AR1; and the elite in both ARs always criticizes. We have already established that the R elite is truthful, that the good R politician does not send propaganda, and that the elite in both ARs criticizes. It remains to characterize the behavior of the bad R politician and the AR politicians.

To do this, note that in the proposed equilibrium the belief of the voter who observed no propaganda continues to be given by (A7), while the belief of the voter who observed AR1 is

$$\mu_v(\theta_c|\hat{s}, \hat{p} = AR1) = (1 - \hat{s})\frac{q_{ar}q_c}{q_{ar} + q_r(1 - q_c)\pi} \tag{A21}$$

This expression is derived analogously to our basic model. Propaganda and praise ($\hat{s} = 1$) conclusively prove that the politician is bad. For propaganda and criticism, the numerator reflects that in the AR a good politician always sends propaganda and gets criticism, while the denominator reflects that propaganda and criticism can also arise in the R if the politician is bad.

The belief of the voter who observed AR2 is

$$\mu_v(\theta_c|\hat{s}, \hat{p} = AR2) = \hat{s}\frac{q_c\pi}{q_c\pi + (1 - q_c)(1 - \pi)} + (1 - \hat{s})\frac{q_{ar}q_c + q_r(1 - \pi)q_c}{q_{ar} + q_r[(1 - \pi)q_c + (1 - q_c)\pi]}. \tag{A22}$$

The first term represents beliefs after observing AR2 propaganda and praise by the elite. This term is no longer zero because the outcome is attributed to a tremble. More precisely, propaganda shifts the prior to put a positive weight on AR2, but, because AR2-propaganda is not observed on the equilibrium path, it is attributed to a tremble and does not generate updating. Since praise never occurs in AR2, given praise the voter updates that reality is R, thinks that the AR2 propaganda was a tremble, and forms beliefs based on the signal only. The second term represents beliefs after observing AR2 propaganda and criticism from the elite. As in the first term, the voter puts a positive weight on the AR2, but since AR2 propaganda never happens on the equilibrium path, it

is attributed to a tremble and does not generate updating. Therefore, the numerator reflects that in the AR2 a $q_c$ share of politicians are good and in the R a good politician is only criticized if the elite receives a bad signal (which happens with probability $1-\pi$). The denominator reflects that the elite always sends a bad message in AR2, while in reality she criticizes the incumbent if the politician is good but she received an incorrect signal or if the politician is bad and she received a correct signal.

Similarly to the basic model, (A21) implies that on the proposed equilibrium path, as $\pi$ converges to one, the R politician's return to successful AR1 propaganda is governed by $\hat{q}_c$. Hence, by Assumption 2, for the bad R politician AR1 propaganda is better than no propaganda. Moreover, AR1 propaganda is better than AR2 propaganda because in the limit as $\pi$ goes to one, (A21) and (A22) imply that the return to AR1 is the same as that to AR2, but AR1 has a lower cost. The same logic implies that the AR politicians—in both AR1 and AR2—choose to send AR1 propaganda. This confirms that the proposed profile is an equilibrium.

*Equilibrium selection.* We show that for $\pi$ large the proposed equilibrium is the unique PPO equilibrium. Recall that PPO implies that the politician uses pure strategies. We already characterized the behavior in any equilibrium of the R and AR elites and the good R politician. Our preferred equilibrium is better than any equilibrium in which the bad R politician refrains from propaganda, because here he strictly prefers to send AR1 propaganda and thus doing so improves his payoff. Thus, in any PPO equilibrium, the bad R politician must send either AR1 or AR2 propaganda. We consider these cases in turn.

[Bad R politician sends AR1 propaganda.] Then the AR1 politician must also send AR1 propaganda, because otherwise observing AR1 would lead the persuaded voter, who now has a positive prior on R and AR1 (but not on AR2) to conclude that reality is R, which cannot be profitable for the bad R politician. This already shows that the equilibrium path is the same as in our preferred equilibrium. We now show that for $\pi$ large enough the equilibrium is also the same. It is not optimal for the AR2 politician to send no propaganda, since the R politician, who gets criticized less often, sends propaganda. Suppose that the AR2 politician sends AR2 propaganda. Then the persuaded voter's beliefs after AR2 propaganda are that reality is AR2 and the politician is good

with probability $q_c$. Deviating to AR1 propaganda would instead generate beliefs that are identical to those that emerge after the AR1 politician sends AR1 propaganda, as given by (A21). Thus, except for the knife-edge case of indifference, which can only happen for one value of $\pi < 1$ given the strict monotonicity of (A21) in $\pi$, if the AR2 politician prefers to send AR2 propaganda, then so does the AR1 politician, a contradiction. It follows that for $\pi$ large enough in any PPO equilibrium the AR2 politician sends AR1 propaganda. This is our preferred equilibrium.

[Bad R politician sends AR2 propaganda.] Then the AR2 politician must also send AR2 propaganda. Consider the AR1 politician. No propaganda cannot be optimal for her, since the R politician, who gets criticized less often, sends AR2 propaganda. If he sends AR1 propaganda, the voter will conclude that reality is AR1 and he is good with probability $q_c$. This is better than AR2 propaganda, which is more expensive and leads to worse beliefs, so he sends AR1. Given this, the AR2 politician also prefers to send AR1, a contradiction.

Case 2: $1/N < \chi_f < (1 - 2\alpha)$.

*Behavior in any equilibrium.* We begin by characterizing the behavior of some actors in any equilibrium. As in Case 1, the assumption that $\chi_h$ is prohibitively large implies that the R elite is truthful. Therefore, for $\pi$ large the good R politician does not send propaganda. For $\pi$ large the AR1 elite is also truthful. This is because, in the limit as $\pi$ goes to one, the maximal gain from changing the perception of her audience is $1/N$, which, since we are in Case 2, is smaller than her lying cost of $\chi_f$. However, for $\pi$ large the AR2 elite always sends a bad message after a good signal, because doing so generates a gain of $1 - \alpha$ in the limit from unreceptive voters, and a loss of at most $\alpha$ from persuaded voters, and in Case 2 we have that $1 - 2\alpha > \chi_f$.

*Existence of candidate equilibrium.* We now show that the following strategy profile is an equilibrium. The R and the AR1 elite are truthful; the AR2 elite always criticizes; the good R politician does not send any propaganda; the bad R politician sends AR2; both AR politicians send AR2. Given the results above, we only need to verify the optimality of the behavior of the bad R and the AR politicians.

Observe that no politician sends AR1 propaganda. This follows from the fact that the AR1 elite is truthful, which implies that AR1 propaganda has no effect on the voter's interpretation of

the elite's message, while having a positive cost. However, sending AR2 propaganda is optimal for the bad R politician, for the same reason that propaganda is optimal in the basic model. Indeed, since AR1 is off the table, the setup is identical to that of the basic model, and by Assumption 2, for $\pi$ sufficiently high the benefit of propaganda exceeds the cost. The same logic implies that sending AR2 propaganda is optimal for the AR1 and the AR2 politician.

*Equilibrium selection.* In any equilibrium weakly better for the bad R politician that the one proposed here, he has to send AR2 propaganda. This is because the proposed equilibrium yields a higher payoff than that of not sending propaganda, and sending AR1 propaganda—as established in the previous paragraph—is not useful given that the AR1 elite is truthful. Since the bad R politician is sending AR2, the AR2 politician must also be sending AR2, otherwise the voter learns from observing AR2 (and having a positive prior on R and AR2) that reality must be R. Finally, the AR1 politician must also prefer to send AR2 propaganda, since doing so is more attractive than sending no propaganda, and sending AR1, as established above, is even worse than sending no propaganda.

Case 3: $1 < \chi_f$.

We prove that in the unique equilibrium the elites in all realities are always truthful and the politicians never send propaganda. As before, the R elite is truthful. The assumption that $1 < \chi_f$ implies that the gain to any AR elite from fully influencing the entire electorate is smaller than the fabrication cost. It follows that telling the truth is optimal for them as well. Since neither propaganda changes the interpretation of the elite's message, no politician chooses propaganda.

## A.9 Proof of Proposition 3

Key to the proof is that for $\pi$ large, both when $e = 0$ and when $e = 1$, the elite's signal is almost perfectly informative. As a result, the large-$\pi$ arguments used in the proof of the main result also apply here.

*Behavior in any equilibrium.* We begin by characterizing the behavior of some actors in any large-$\pi$ equilibrium. Begin with the elite. As in the basic model, since its members have no impact on the outcome, the R elite is always truthful. Consider the AR elite. In the absence of propaganda

they always send a bad message. In the presence of propaganda, the gain from sending a bad rather than a good message, as $\pi$ approaches one, approaches $1-\alpha$, because the $1-\alpha$ share of unreceptive types believe (for $\pi$ large) that the elite's message is almost perfectly informative. The loss from sending a bad rather than a good message is at most $\alpha$ because in the worst case the share $\alpha$ of receptive voters react in the exact opposite way to her message. Since $\alpha < 0.5$, for $\pi$ large enough the AR elite always sends a bad message.

Now consider the good R politician. For $\pi$ large enough, he earns close to the maximal payoff absent propaganda, and hence refrains from costly propaganda.

*Existence.* We turn to establish that the proposed profile constitutes an equilibrium. Given the above results, to prove existence, we only need to focus on the bad R politician and the good and bad AR politicians. First consider their decisions about propaganda. For $\pi$ large, the bad R politician, and the good and bad AR politician all prefer to send propaganda by Assumption 2. This is for the same logic as in the main result. Absent propaganda the elite (i) almost certainly sends a bad message (both when $e = 0$ and when $e = 1$), and (ii) is perceived by all voters to be almost fully informative. Hence expected average beliefs about competence become approximately zero. In the presence of propaganda, because the elite almost always sends a bad message, the expected weighted average belief $\mu'$ approximates $\alpha'\hat{q}_c$, since receptive voters' belief approximates $\hat{q}_c$ (for the same reason as in our main setting) while unreceptive voters' beliefs approximate zero. Since $\alpha' > \alpha$, the result follows from Assumption 2.

Now consider the bad politician' decision about $e$. The bad AR politician, since he expects to be criticized no matter what he does, is indifferent between more or less precise elite signals and chooses $e = 0$. The bad R politician who can not send propaganda (which happens with a probability $\beta$) expects, for $\pi$ large, that voter beliefs will be close to zero after a bad elite message and close to one after a good elite message. Thus, he would like to minimize the probability of a bad elite message and chooses $e = 0$.

Finally, consider the bad R politician who can send propaganda. At this step we need to explicitly calculate voters' beliefs after propaganda. In the proposed equilibrium the politician chooses $e = 1$, making the elite's signal correct with probability $\pi'$. Therefore the belief of the

receptive voter after propaganda, as a function of the elite's message, is

$$\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s}) = (1 - \hat{s})\frac{q_{ar}q_c}{q_{ar} + q_r\pi'(1 - q_c)}.$$

As in the basic model, when the elite praises the politician ($\hat{s} = 1$), posterior beliefs are that the politician is bad. When the elite criticizes, posterior beliefs are a function of the probability of criticism when the politician is good, which can only happen in the AR ($q_{ar}q_c$), relative to the probability of criticism, which always happens in the AR ($q_{ar}$) and happens in R for the bad politician if the elite's message is bad ($q_r(1 - q_c)\pi'$). Note that the last term accounts for the fact that the bad R politician chooses a more precise signal.

To compute the belief of the unreceptive voter about the politician, we introduce $\hat{\pi} = \beta\pi + (1-\beta)\pi'$, which is the unreceptive voter's belief about the precision of the elite's signal. This holds because in the proposed path the bad R politician sets $e = 1$, implying precision $\pi'$, precisely in the $\beta$ probability event in which he can send propaganda. The beliefs of the unreceptive voter after propaganda are given by

$$\mu_{un}(\theta_c = 1|\hat{p} = 1, \hat{s}) = \hat{s}\frac{\pi q_c}{\pi q_c + (1 - \hat{\pi})(1 - q_c)} + (1 - \hat{s})\frac{(1 - \pi)q_c}{(1 - \pi)q_c + \hat{\pi}(1 - q_c)}.$$

The first term says that when observing a good signal, posterior beliefs are governed by the probability of that good signal under a good politician, $\pi q_c$, relative to the probability of a good signal under a good or a bad politician $\pi q_c + (1 - \hat{\pi})(1 - q_c)$, where the $\hat{\pi}$ reflects the probability of a correct signal under a bad politician. The intuition for the second term, which expresses posterior beliefs after a bad signal, is similar.

The condition that the bad R politician prefers $e = 1$ is

$$\begin{aligned} &\alpha' \cdot [\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 0) - \mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 1)] \\ +(1 - \alpha') &\cdot [\mu_{un}(\theta_c = 1|\hat{p} = 1, \hat{s} = 0) - \mu_{un}(\theta_c = 1|\hat{p} = 1, \hat{s} = 1)] > 0. \end{aligned} \tag{A23}$$

Indeed, the left-hand-side is a weighted average of the belief changes of the unreceptive and receptive voter in response to improving the precision of the signal, which here means that a positive signal is turned into a negative signal. The weights are those that the politician assigns to the two classes

of voters. Substituting in the above expressions for the beliefs, we obtain

$$\alpha' \frac{q_{ar} q_c}{q_{ar} + q_r \pi' (1 - q_c)} + (1 - \alpha') \left[ \frac{(1 - \pi) q_c}{(1 - \pi) q_c + \hat{\pi} (1 - q_c)} - \frac{\pi q_c}{\pi q_c + (1 - \hat{\pi})(1 - q_c)} \right] > 0.$$

It is straightforward to check that as $\pi$ and $\pi'$ approach one, the condition collapses to $\alpha' > 1/(1 + \hat{q}_c)$. Thus, for any such $\alpha'$, we can find $\pi$ large enough that the result holds.

*Equilibrium selection.* We show that for $\pi$ large the proposed equilibrium is the unique PPO equilibrium. We already characterized in any equilibrium the behavior of the R and AR elites and the good R politician. Our preferred equilibrium is better than any equilibrium in which the bad R politician refrains from propaganda, because here he strictly prefers to send propaganda and thus doing so improves his payoff. Thus, in any PPO equilibrium, the bad R politician must send propaganda. But then the good AR politician must also send propaganda, since otherwise propaganda would reveal that the politician is bad, in which case it would not be worth it for the bad R politician. At this step we used the fact that we are looking for a PPO equilibrium, so that the good AR politician is not mixing. And then the bad AR politician must also send propaganda, since he faces a worse portfolio of elite messages (always criticism) than the bad R politician (often criticism). Thus, the propaganda decisions are uniquely pinned down.

We now turn to the policy decision. The AR politician, since he is always criticized anyway, chooses $e = 0$. The bad R politician who cannot send propaganda, since he would like to minimize the probability that the elite sends a bad message, chooses $e = 0$. Finally, consider the bad R politician who can send propaganda. Consider a candidate equilibrium in which this politician sets $e = 0$. Then the equilibrium path, including actions and beliefs, is exactly identical to the simple propaganda equilibrium of the basic model. Thus, we can evaluate the condition that the bad R politician prefers $e = 1$ by substituting in the beliefs from (A6) and (7) into (A23). It is straightforward to check that as $\pi$ approaches one, the condition approaches $\alpha' > 1/(1 + \hat{q}_c)$, which holds by assumption. Thus, setting $e = 1$ is optimal, a contradiction. The only remaining case is our preferred equilibrium.

Table A1: Impact of redistricting on contributions from Trump-supporter and other donors

| | Trump donors Share | Trump donors Amount (1,000 dollars) | Other donors Amount (1,000 dollars) |
|---|---|---|---|
| $\Delta$ predicted Dem margin | 0.001 | -1.07 | 1.43 |
| | (0.001) | (1.60) | (3.57) |
| Old predicted Dem margin | 0.001 | 0.402 | 5.36*** |
| | (0.0006) | (0.454) | (1.05) |
| Observations | 266 | 296 | 296 |

Note: Observations are house representative by quarter cells. The sample consists of all Republican Representatives in the last two quarters of 2022. The Democratic vote margin of a Representative's old district and its change between the old and the new district are taken from FiveThirtyEight (2022). In column 1, the dependent variable is the share of donations from Trump supporters; in columns 2 and 3 it is the volume of donations from Trump supporters and from other Republicans, respectively. Standard errors are clustered by state.

# B Evidence

A possible alternative explanation for the scandal effects documented by Table 3 is that scandals increase donations because they intensify electoral competition. We provide evidence gainst this explanation by exploiting the redistricting of congressional districts before the 2022 midterm elections. We combine data on predicted Democratic vote margins for both the old and the new districts of Republican representatives from FiveThirtyEight (2022) with donations data from the Federal Elections Commission. We estimate

$$y_{iq} = \text{const} + \beta \Delta DVM_i + \gamma DVM_i^{old} + \delta_q + \varepsilon_{iq}, \tag{A24}$$

where $y_{iq}$ measures donations received by candidate $i$ in quarter $q$ of the 2022 midterm elections campaign; $DVM_i^{old}$ is the predicted Democratic vote margin of candidate $i$ in their electoral district in the period 2011-2020; $\Delta DVM_i = DVM_i^{new} - DVM_i^{old}$ is the change in predicted Democratic vote margin between the new and the old district; and $\delta_q$ denotes quarter fixed effects.

Table A1 reports the results. Column 1 shows that a reduction in the chance of winning—

induced by an unfavorable change in the electoral map—has a small and insignificant effect on the Trump-supporter share, while columns 2 and 3 document small impacts on the volume of donations. Thus, a decline in the electoral prospects of Republican house candidates changes neither the volume nor the composition of donations.

## References


**Brunnermeier, Markus K and Jonathan A Parker**, "Optimal expectations," *American Economic Review*, 2005, *95* (4), 1092–1118.

**FiveThirtyEight**, "What Redistricting Looks Like In Every State," `https://web.archive.org/web/20220603143216/https://projects.fivethirtyeight.com/redistricting-2022-maps/` 2022.

**Levy, Raphaël**, "Soothing politics," *Journal of Public Economics*, 2014, *120*, 126–133.