A Model of Populism as a Conspiracy Theory[*]

Adam Szeidl
Central European University and CEPR

Ferenc Szucs
Stockholm University

May 14, 2025

Abstract

We model populism as the dissemination of a false “alternative reality”, according to which the intellectual elite conspires against the populist for purely ideological reasons. If enough voters are receptive to it, this alternative reality—by discrediting the elite’s truthful message—reduces political accountability. Elite criticism, because it is more consistent with the alternative reality, strengthens receptive voters’ support for the populist. Alternative realities are endogenously conspiratorial to resist evidence better. Populists, to leverage or strengthen beliefs in the alternative reality, enact harmful policies that may disproportionately harm the non-elite. These results explain previously unexplained facts about populism.

Keywords: populism, conspiracy theory, alternative reality, propaganda, misbeliefs, distrust

JEL codes: D03, D72, D82, D83

[*] Emails: szeidla@ceu.edu, ferenc.szucs@su.se. An earlier draft of this paper was titled “The Political Economy of Alternative Realities.” We thank Nageeb Ali, Ruben Durente, Ruben Enikolopov, Matthew Gentzkow, Sergei Guriev, Helios Herrera, Botond Koszegi, Kristof Madarasz, Maria Petrova, Giacomo Ponzetto, Jesse Shapiro, David Stromberg, Ekaterina Zhuravskaya, and conference and seminar audiences for comments and discussions. We thank the European Research Council (ERC), under the European Union’s Horizon 2020 research and innovation programme grant agreement number 724501, and under the Horizon Europe grant agreement number 101163526; and the Austrian Science Fund (FWF), under grant F100600, for funding.

1 Introduction

Populist leaders paint a grim picture of the world. The populist ideology is often centered around a false narrative, an *alternative reality*, in which a conspiracy of the elite shapes major events to the detriment of the people. In this narrative, the elite is not only “corrupt”, as described in leading accounts of populism (Guriev and Papaioannou 2022); it is conspiratorial and all-powerful. For example, a central part of Donald Trump’s political narrative is that the 2020 US election was stolen by a conspiracy of the deep state. The populist narrative matters because it shapes supporters’ beliefs: the majority of Republicans believe that Trump did not lose the 2020 election legitimately.[1] These sorts of misbeliefs are potentially highly consequential.

Populism is also associated with a profound decline in political accountability. Funke, Schularick and Trebesch (2023) show that populist leaders, despite substantially reducing GDP per capita, stay in power for twice as long as non-populists. Moreover, populists seem to achieve electoral success despite widely-publicized acts that would normally be extremely damaging: for example, Donald Trump, a convicted felon, won the 2024 election.

Existing economic models do not explain the use of conspiratorial narratives in politics and their association with reduced accountability. Models of political narratives do not study conspiratorial narratives (Eliaz and Spiegler 2020, Eliaz, Galperti and Spiegler 2022). Models of reduced accountability work through the silencing of the media, or through repression (Guriev and Treisman 2020, Egorov and Sonin 2024), hence do not explain reduced accountability in populist democracies that have an independent media. And in models of populism, populist policies are a positive signal, so that populism is associated with *increased* accountability (Acemoglu, Egorov and Sonin 2013, Bellodi, Morelli, Nicolò and Roberti 2023).

We present a new theory of populism that helps us understand both conspiratorial narratives and reduced accountability. In our theory, the goal of populism is to provide a false *alternative reality* that discredits the intellectual elite’s message about the politician. Specifically, we assume that populist propaganda can (partially) persuade voters that the elite conspires to criticize the

[1] See Figure 1 below. People not only claim to hold these beliefs, they act on them, as illustrated by the 2021 January 6 attack on the United States Capitol.

politician's competence purely because they disagree with his ideology. This alternative reality discredits elite criticism that would normally reveal the politician's type. "Bad" politicians, expecting elite criticism, propagate the alternative reality to remain in power. Thus, our theory predicts both the use of conspiratorial propaganda and its association with reduced accountability.

We formalize these ideas in a model which explicitly incorporates the false alternative reality. Beyond explaining our motivating facts, this model makes several new predictions. It predicts that truthful elite criticism can backfire and strengthen some voters' support for the politician; that alternative realities are endogenously conspiratorial to better resist evidence; and that populists, despite their "pro-people" rhetoric, may set policies that disproportionately harm the non-elite. These results offer a new understanding of populism.

In our model, presented in Section 2, an incumbent politician is characterized by a type dimension, e.g., competence, along which he can be good or bad. Voters do not directly observe the politician's type but form beliefs over it, and political accountability is measured with the accuracy of their beliefs. The politician and the intellectual elite send messages which affect these beliefs. First, the politician chooses whether to send conspiratorial propaganda. Then the elite (including the news media), having received an informative signal about the politician's type, sends a message that reports on that signal. We assume that the elite consists of a continuum of small members, who individually cannot influence voters and thus report about the signal truthfully.

A share α of voters are receptive to propaganda, so that propaganda exogenously and counterfactually increases their prior belief in the alternative reality (AR). The AR is a state of the world with zero objective probability, which differs from the objective reality in precisely one way: In the AR the continuum of small elite members can coordinate—effectively conspire—and thus can collectively choose their message to influence voters. It follows that if elite members sufficiently dislike the politician, perhaps because they disagree with his ideology, then in the AR they will always report that politician bad. Intuitively, in the AR the "fake news media" criticize Trump's competence not because he is incompetent, but because he is "anti-woke." In turn, a voter persuaded by propaganda partially believes this alternative reality and distrusts elite criticism.

We analyze the model in Section 3. We show that an equilibrium of the following form emerges.

(i) In the objective reality only the bad politician sends propaganda, and the elite always reports truthfully. (ii) In the alternative reality both the good and the bad politician sends propaganda, and the elite always criticizes the politician. Intuitively, in reality the good politician has no reason to send propaganda as he expects praise from the elite. The bad politician, who expects criticism, has an incentive to send propaganda if doing so discredits elite criticism. Discrediting only works if the narrative of the alternative reality is plausible: if it is incentive compatible for the conspiring elite to criticize even a good politician. This holds provided that elite members sufficiently dislike the politician (sufficiently disagree with his ideology). Under this assumption, the equilibrium admits the above form; otherwise it does not feature propaganda. These results immediately predict the equilibrium use of conspiratorial propaganda, and its association with reduced accountability. Thus, our model helps explain our motivating facts.

The model also yields new theoretical implications. First, it predicts that propaganda *inverts* the effect of the elite's message on receptive voters, so that elite criticism increases their beliefs that the politician is good. The key intuition is that for a receptive voter who experienced propaganda, the elite's message is primarily informative about the nature of reality. In particular, he knows that in the alternative reality the (conspiring) elite always criticizes, while in the objective reality the (honest) elite only sometimes criticizes. Thus, observing elite criticism is more consistent with the alternative reality, increases his posterior of the alternative reality, and with it, his posterior that the politician is good. It follows that truthful critical information can increase receptive voters' support for the politician, a result that overturns standard intuitions about the impact of information in political economics.

Second, the politician's choice of when to send propaganda *amplifies* receptive voters' misbelief about the alternative reality. Intuitively, the politician supplies the alternative reality precisely when events, such as elite criticism, are expected to be consistent with it. The receptive voter neglects this correlation, implying that his misbelief is (on average) strengthened by events. As a result, even propaganda that plants a small initial misbelief can have large societal effects.

These implications help explain previously unexplained facts. The inversion result explains a key fact in contemporary US politics: that the four criminal indictments against Trump in 2023 were accompanied by an *increase* in his support among Republican voters (Swan, Igielnik, Goldmacher and Haberman 2023). This reaction by supporters of the presumptive party of law and order is puzzling, especially when compared to the case of Nixon, who lost Republican support after Watergate. Our inversion result explains the increased support for Trump by predicting that it was the *causal effect* of the indictments. This prediction is in line with survey evidence that Republicans claimed to increase support for Trump due to the indictments. It is also in line with new evidence we present that scandals of Republican politicians caused an increase in the donations they received from Trump supporters. Moreover, the mechanism for inversion, increased beliefs in the alternative reality, is consistent with the fact that following the indictments, Republicans sharply increased their beliefs in the conspiracy theory that the 2020 election was stolen. Finally, our model explains the contrast between Trump and Nixon through the logic that only Trump had a sufficiently large ideological cleavage with the elite to make the alternative reality plausible. Nixon, representing the more educated party (Republicans around 1970), could not credibly argue that the intellectual elite conspired to remove him.

The amplification result may help explain why beliefs in the alternative reality are so prevalent (e.g., held by most Republicans), a fact that seems difficult to attribute purely to the persuasive effect of propaganda. Consistent with the logic of amplification, we argue that populists in multiple countries supplied the alternative reality precisely in situations where it matched headline facts.

In Section 4 we develop two applications of the model. First, we investigate the reason that alternative realities are often conspiracy theories. In our basic model, the conspiracy was purely by assumption. We now allow the politician to choose between two types of alternative realities: one in which elite members have a low lying cost but cannot conspire, and another in which they can also conspire. Sending propaganda about the latter is more expensive. We show that the conspiracy theory often dominates, because it solves a collective action problem of the elite. Intuitively, each elite member's lie about the politician benefits all other elite members, resulting in a within-elite externality which the conspiracy internalizes. Thus, the ability to conspire makes the elite more powerful. As a result, the conspiracy alternative reality is more attractive to the politician, because the more powerful elite can explain away even more credible critical evidence. We believe that this is the first formal explanation for the prevalence of political conspiracy theories.

This analysis predicts that alternate realities are often resistent to evidence, because in response to more credible criticism the politician can "upgrade" the narrative from a lying elite to a conspiring elite. Upgrading is socially harmful, because it increases distrust in the elite beyond politics. Once the voter contemplates an elite conspiracy, he fears that the conspiracy's interests may be driving elite messages even in other domains. This logic helps explain why misbeliefs under populism extend beyond politics, including Republicans' general distrust in science.

In our second application, we investigate the effect of conspiratorial populism on government policy. This is an important topic since populism is associated with large economic and non-economic costs (Guriev and Papaioannou 2022). We find that populists introduce harmful policies for two distinct reasons. First, there is a direct effect of reduced accountability: populism enables "bad" politicians to maintain power, who then enact "bad" policies. Second, our model predicts that populists will choose harmful polcies *purely to trigger the elite.* The intuition follows from the inversion result: since elite criticism increases the support of receptive voters, the politician chooses harmful policies to invite elite criticism.

Harmful policies also emerge in the Acemoglu et al. (2013) model, where populists signal their independence from the elite using policies that disproportionately harm the elite. The key difference is that our model can also account for harmful policies that *do not* disproportionately harm the elite. As a result, our model helps explain the previously unexplained fact that populists, despite their pro-people rhetoric, are not actually siding with the "people": their policies seem to hurt the non-elite as much as they hurt the elite. Indeed, Funke et al. (2023) show that populists reduce GDP per capita without meaningfully reducing inequality, i.e., that they seem to cause equal economic harm to the elite and the non-elite. Populists also favor specific policies that especially harm the "people." They tend to be massively corrupt (Zhang 2024), thereby reducing the quality of government services; they implement tariffs that harm their core supporters (Fajgelbaum, Goldberg, Kennedy and Khandelwal 2019); and they oppose environmental policies that would help the non-elite (Friedman, Plumer and Stevens 2025). We conclude that the current wave of populism will likely create substantial harm to both the elite and the non-elite.

Our paper builds on overlapping literatures in political, behavioral, and information economics. We build on theories of populism (Acemoglu et al. 2013, Bellodi et al. 2023, Agranov, Eilat and Sonin 2023) and identity politics (Besley and Persson 2021, Bonomi, Gennaioli and Tabellini 2021). Our main contribution to this work is a conspiracy-theory-based model of populism. Our model makes a number of new predictions about populism, including reduced accountability, the inverted effect of elite criticism, the emergence of conspiracy theories, and broadly harmful policies.

We also build on work studying the supply of misinformation in politics, including the supply of hatred (Glaeser 2005), media capture (Besley and Prat 2006), censorship and positive propaganda (Guriev and Treisman 2020, Egorov and Sonin 2024), and worldview politics (Ash, Mukand and Rodrik 2021). Much of this work assumes that voters update in a Bayesian fashion, as in models of Bayesian persuasion (Kamenica and Gentzkow 2011). We contribute to this work by modeling misinformation as a strategic alternative reality, an approach potentially portable to other settings; and with the aforementioned new implications.

Our modelling approach builds on theories of distorted belief formation in political economics. Theories of motivated beliefs in an alternative state of the world have been used to study beliefs in a just world (Bénabou and Tirole 2006), groupthink (Bénabou 2013), inefficient policy-making (Levy 2014) and partisan disagreement (Le Yaouanq 2023). Theories of model misspecification have been used to study persuasion (Bénabou, Falk and Tirole 2018, Galperti 2019, Schwartzstein and Sunderam 2021, Aina 2023) and political narratives (Eliaz and Spiegler 2020, Eliaz et al. 2022). We contribute to this work with a model-misspecification-based approach to model conspiratorial populism, and with its new implications.

Finally, we build on a multidisciplinary literature studying misinformation and conspiracy theories, especially in political science and psychology, reviewed for example by Nyhan (2020) and Douglas, Uscinski, Sutton, Cichocka, Nefes, Ang and Deravi (2019). Our main contribution to this work is a formal model of conspiracy theories.

2 Model

2.1 Motivation

Our model is motivated by two observations. First, in a number of democracies, the populist ideology is centered around a conspiratorial alternative reality, according to which the elite conspires to attack the competence of the populist purely because they disagree with his ideology. This alternative reality goes beyond the Mudde (2004) and Guriev and Papaioannou (2022) descriptions of populist ideology, which emphasize the antagonism between the "pure people" and the "corrupt elite", in that here the elite is conspiratorial and all-powerful. The following examples illustrate.

- In the United States, Trump claims that the deep state and the media conspire to criticize him (e.g., claim him a criminal) because they find his cultural values too conservative. Trump talks about the conspiracy explicitly, "Either the deep state destroys America, or we destroy the deep state" (Allen 2023); ties the incentives of the conspiracy to cultural values, "they won't hesitate to ramp up their persecution of Christians, pro-life activists"; and suggests that the goal of the conspiracy is to limit conservative values "they want to silence me because I will never let them silence you" (Corasaniti and Gabriel 2023).
- In Hungary, Orban claims that the members of the "Soros network"—including Brussels and the media—conspire to attack him for dismantling checks and balances because they find him too anti-immigration. Orban is explicit about this narrative. "And we understand what is happening. George Soros has bought people, he has bought organisations, he is feeding them out of the palm of his hand, Brussels is under his influence, and it is his plan that the Brussels machine is implementing in the case of immigration. They want to remove the fence, they want to let in millions of immigrants and they want to divide them up on a compulsory basis. And they want to punish those who do not obey." (Kocsis 2017).
- In Israel, Netanyahu claims that the judiciary and the media conspire to attack him on charges of corruption because they find him too anti-Palestinian. As Horovitz (2020) explains, Netanyahu's thesis is that "a strong, pro-annexation, right-wing prime minister is facing an illicit attempt — perpetrated by a vast, leftist alliance of politicians, media, cops and state

prosecutors — to oust him because of his ideology and policies".

Our second observation is that in the same democracies, supporters of the populist leader tend to hold misbeliefs consistent with these alternative realities.

- In the US, the majority of Republicans believe that Biden did not win the 2020 election legitimately (see Figure 1 below). Since large-scale election fraud requires a conspiracy, these beliefs reflect beliefs in the deep state conspiracy.
- In Hungary, the majority of Orban's supporters believe in the existence of a Soros-plan (hvg.hu 2017), i.e., the conspiracy that the Soros network is bringing migrants into Europe.
- In Israel, a large fraction of Netanyahu's supporters doubt the corruption charges against him (Navot 2022), suggesting beliefs in a conspiracy of the justice system.

The fact voters believe in the specific and often elaborate alternative reality supplied by the politician (e.g., the "Soros-plan") suggests that these beliefs are at least partly driven by the supply of populist propaganda.[2] This motivates our model in which a politician can supply a conspiratorial alternative reality to change voter beliefs.

2.2 Setup

Players, types, and actions. Our model is an information game in which a politician and the intellectual elite send messages to influence voters' beliefs about the politician's type. Both the intellectual elite, which represents the news media, and the voters consist of a unit mass of members. Each elite member has limited influence: each sends its message to an audience of voters who have measure zero. In turn, each voter has access to the message of exactly one elite member, i.e., consumes exactly one news media.[3]

At the beginning of the game the politician's type $\theta_c \in \{0, 1\}$ is realized, where $\theta_c = 1$ with probability q_c. Here $\theta_c = 1$ means that the politician is "good" and $\theta_c = 0$ means that the

[2] There is much evidence that the political supply of misinformation changes beliefs (Yanagizawa-Drott 2014, Adena, Enikolopov, Petrova, Santarosa and Zhuravskaya 2015, Blouin and Mukand 2019, Barrera, Guriev, Henry and Zhuravskaya 2020, Ajzenman, Cavalcanti and Da Mata 2023) and that populists value the supply sufficiently to capture media (Mcmillan and Zoido 2004, Szeidl and Szucs 2021).

[3] We present a formal construction of the elite and the voters in Appendix A.1.

politician is "bad." Good politicians are valued by both elite members and voters. We refer to θ_c as competence, but it could represent some other broadly valued attribute such as being honest (as opposed to corrupt), lawful (as opposed to criminal), or democratic (as opposed to authoritarian). We assume that θ_c is observed only by the politician.

After observing his type, the politician has an opportunity with probability $1 - \beta$ (where $0 < \beta < 1$) to send propaganda. Only the politician knows whether he has the opportunity to send propaganda. We let $p = 1$ denote that the politician sends propaganda, and $p = 0$ that he does not, either because he did not have the opportunity or because he chose not to. The role of β is to ensure that the absence of propaganda does not fully reveal the politician's type.

Members of the elite observe the propaganda realization and receive a signal $\hat{\theta}_c \in \{0, 1\}$ about the politician's type θ_c. This signal is correct ($\hat{\theta}_c = \theta_c$) with probability $\pi \in (0.5, 1]$. All elite members receive the same signal; voters do not receive a signal. We think of π as relatively high. Then, each elite member j sends a message $s_j \in \{0, 1\}$ about the signal to its zero measure of voters, where $s_j = 1$ means that the signal is good. We sometimes refer to $s_j = 1$ as praise and $s_j = 0$ as criticism.

There are two kinds of voters. A share α are receptive to propaganda, and observe both the elite's message and propaganda. The remaining share $1 - \alpha$ are unreceptive, and only observe the elite's message, but not propaganda. This is a stylized representation of the idea that unreceptive voters primarily follow the news media and are less exposed to propaganda. Each elite member j has an audience consisting of a share α of receptive and a share $1 - \alpha$ of unreceptive voters.

We assume that the elite's message s_j and propaganda p are subject to vanishing noise. This ensures that beliefs are well-defined off the equilibrium path. With probability ε_e, perfectly correlated across elite members, every elite member's realized message $\hat{s}_j$ is the opposite of the message s_j sent; and with independent probability ε_p, realized propaganda $\hat{p}$ is the opposite of the propaganda p sent. We let ε_e and ε_p go to zero and characterize the equilibrium in the limit.

Alternative reality. To model the alternative reality, we assume that there is a state of the world $\theta_r \in \Theta_r = \{R, AR\}$, where R represents the objective reality and AR the alternative reality. We

Stage:	0	1	2
Politician	$\theta_c,\ \theta_r$	$\hat{p}$	$\hat{s}$
Elite		$\hat{p},\ \hat{\theta}_c,\ \theta_r$	$\hat{s}$
Receptive voter		$\hat{p}$	$\hat{s}$
Unreceptive voter			$\hat{s}$

Table 1: Timing and allocation of information

assume that the true prior probability of $\theta_r = AR$ is zero.[4] The difference between the two realities is that in R the elite cannot, but in AR the elite can coordinate. Thus, if $\theta_r = R$, then each elite member j chooses her message s_j individually to maximize her own utility, but if $\theta_r = AR$, then the elite collectively chooses an identical message $s_j = s$ for all of its members to maximize the sum of their utilities.[5]

At the beginning of the game all voters hold the correct prior about θ_r, but propaganda exogenously increases receptive voters' prior that $\theta_r = AR$ to $q_{ar} > 0$. We let $q_r = 1 - q_{ar}$. We encode the change in the prior by assuming that each receptive voter i has a mind type $\theta_{mi} \in \{N, P\}$ (for normal and persuaded), that i becomes persuaded if and only if he encounters propaganda, and that the prior of receptive voter i as a function of his mind type is $\mu^0_{rec,i}(\theta_r = AR|\theta_{mi}) = 1_{\{\theta_{mi}=P\}} \cdot q_{ar}$. Each receptive voter then updates from his prior in a Bayesian fashion. Since either all or none of the receptive voters encounter propaganda, their mind types are identical and denoted by θ_m.

Motives. We assume that the payoffs of the politician and the elite are determined by voters' average beliefs about θ_c. Assuming that payoffs depend on beliefs simplifies presentation because it allows us to abstract from voters' preferences and actions. In Appendix A.2 we show that a probabilistic voting model with a common preference shock provides microfoundations for this assumption, through the logic that voters' beliefs govern the probability that the politician is reelected, which in turn determines payoffs.

[4] In principle we could allow this prior to be positive, but to make the results stark we set it to zero.

[5] In our microfoundation in Appendix A.2 we show that sending an identical message is the optimal strategy for a coordinating elite.

We define voters' average posterior belief about θ_c as

$$\bar{\mu}(\theta_c = 1|\hat{p}, \hat{\mathbf{s}}) = \alpha \cdot \overline{\mu_{rec,i}}(\theta_c = 1|\hat{p}, \hat{s}_{j(i)}, \theta_m) + (1 - \alpha) \cdot \overline{\mu_{un,i}}(\theta_c|\hat{s}_{j(i)}).$$

On the left-hand-side, the conditioning shows that the average posterior depends both on realized propaganda $\hat{p}$ and the full collection of realized elite messages $\hat{\mathbf{s}} = (\hat{s}_j)_{j \in \text{elite}}$. In the first term on the right-hand-side, $\mu_{rec,i}(\theta_c = 1|\hat{p}, \hat{s}_{j(i)}, \theta_m)$ stands for the belief of receptive voter i who observes realized propaganda $\hat{p}$, belongs to the audience of elite member $j(i)$ and thus observes elite message $\hat{s}_{j(i)}$, and has mind type θ_m (which in turn is pinned down by $\hat{p}$). The bar means that this belief is averaged across all receptive voters i. In the second term, $\overline{\mu_{un,i}}(\theta_c|\hat{s}_{j(i)})$ is the average belief over unreceptive voters i, who only observe the elite's message $\hat{s}_{j(i)}$, not propaganda.[6]

Using these voter beliefs, we define the preferences of elite member j as

$$U_{ej} = (\theta_c - \kappa) \cdot \bar{\mu}(\theta_c = 1|\hat{p}, \hat{\mathbf{s}}). \tag{1}$$

Here $\theta_c - \kappa$ reflects that the elite likes competence θ_c but dislikes the incumbent politician by κ, where κ measures the ideological disagreement between the incumbent and the elite. These terms are multiplied by voters' average posterior belief, which, intuitively, governs the probability that the incumbent stays in power. We further assume that each elite member j has a small preference for sending a truthful message s_j, thus if otherwise indifferent tells the truth.

Our assumptions imply that in state R, because each elite member acts independently and influences a zero measure of voters, each sends her message truthfully. In contrast, in state AR, because elite members act as a single decision maker, they choose their message to maximize (1). In both states, all elite members send an identical message, denoted s. Thus, for the purposes of characterizing behavior, we can represent the elite as a single player which maximizes

$$U_e = 1_{\{\theta_r = AR\}} \cdot (\theta_c - \kappa)\bar{\mu}(\theta_c = 1|\hat{p}, \hat{s}) + 1_{\{\theta_r = R\}} 1_{\{s = \hat{\theta}_c\}}. \tag{2}$$

The first term, active in the AR, represents the collective interests of the elite. The second term, active in R, represents that in isolation, each elite member chooses to tell the truth.

[6] We use i to denote both receptive and unreceptive individual voters.

Since all elite members send the same message, all receptive voters, and all unreceptive voters, form the same beliefs. Thus, we can represent them with a representative receptive voter and a representative unreceptive voter, respectively.

The preferences of the politician are given by

$$U_p = \bar{\mu}(\theta_c = 1|\hat{p}, \hat{s}) - f \cdot p. \tag{3}$$

The first term captures the politician's preference to get reelected, which is governed by voters' belief about his type. The second term captures the cost $f > 0$ of sending propaganda.

Timing. In summary, the model consists of the following stages.

0. The politician's type θ_c and the reality state θ_r are realized and observed by the politician.

1. With probability β the politician cannot send propaganda; with probability $1 - \beta$ he can, and he decides on propaganda $p \in \{0, 1\}$. Propaganda is subject to trembles. The elite observes the reality state θ_r, the realized propaganda $\hat{p}$, and receives a signal $\hat{\theta}_c$ on the politician's type (correct with probability π).

2. The elite sends message $s \in \{0, 1\}$, which is subject to trembles. Voters observe the realized message $\hat{s}$. Receptive voters (share α) also observe the realized propaganda message $\hat{p}$. If $\hat{p} = 1$ then receptive voters' become persuaded ($\theta_m = P$), implying that their prior that $\theta_r = AR$ changes to $q_{ar} > 0$. Voters form posterior beliefs.

2.3 Equilibrium

Our equilibrium concept is a version of perfect Bayesian equilibrium that recognizes our framework's departure from common priors and rationality. We assume that actors correctly anticipate each others' strategies, compute expected utilities using their subjective beliefs, and choose strategies to maximize these expected utilities. We also assume that actors update in a Bayesian fashion. The trembles ensure that these updates are well defined.

The key novelty in this equilibrium is the Bayesian updating of the receptive voter. We assume that in stage 2, the posterior of the receptive voter is computed from the prior associated with his mind type θ_m. This definition allows the persuaded voter to make Bayesian inference from the elite's message and propaganda; but the order of updating is that first propaganda changes his prior, and then he makes the inference. Because aside from this novelty our equilibrium concept is standard, we relegate the formal definition to the Appendix.

Equilibrium selection. Given the complexity of our game we expect multiple equilibria, and we introduce the following criteria for selection. First, we focus on equilibria which are *politician-pure*: in which all politician types in all states use pure strategies. Second, among these equilibria, we focus on *politician-optimal* equilibria, which maximize the ex ante expected utility of the incumbent politician in state R. We refer to equilibria satisfying these conditions as *PPO* equilibria.

2.4 Discussion of model assumptions

Modeling alternative realities. Central to our model is to explicitly incorporate the false alternative reality that voters may believe in. Importantly, this alternative reality contains optimizing agents, who impose constraints on real-world outcomes that parallel the out-of equilibrium constraints of perfect Bayesian equilibrium. Indeed, perfection requires that agents, even at information sets never reached, behave optimally; whereas we require that agents, even in imaginary states, behave optimally. This approach of explicitly modeling a strategic alternative reality—also used by Bénabou (2013) in a different setting—may be portable to other systems of misbeliefs in economics.

Elite conspiracy. In our model, the elite conspiracy emerges by assumption. Moreover, as we show in Section 4.1, our qualitative results would also obtain in a framework without a conspiracy, in which the key difference between the R and the AR is that in the latter the elite has a lower lying cost. We chose to incorporate the conspiracy into our basic model both because it is realistic (Douglas et al. 2019) and because it highlights that allowing for coordination fundamentally alters the equilibrium. We endogenize the conspiracy in Section 4.1 by showing that when the politician can choose between a lying cost and a conspiracy narrative, he will often prefer the latter because it makes the elite appear more powerful.

Propaganda is only observed by part of the electorate. In our model unreceptive voters do not observe propaganda, implying that they are neither manipulated by it nor learn from it about the politician's type. We think of unreceptive voters as those who primarily consume the traditional media, while receptive voters as those who primarily consume social media and propaganda news. Unreceptive voters focus on the elite's message and do not internalize the politician's narrative. In contrast, receptive voters consume propaganda and fully internalize its narrative, but still encounter headline news, consistent with evidence that even strong partisans are aware of the headlines (Angelucci and Prat 2024).

The assumption that unreceptive voters do not observe propaganda at all is for tractability. The key part of this assumption is that unreceptive voters do not fully update from propaganda, so that they may still find the elite's message informative. To demonstrate this, in Appendix A.5 we develop two realistic ways of allowing unreceptive voters to learn from propaganda. In the first, a bounded share of unreceptive voters update from propaganda, while the rest do not. In the second, all unreceptive voters update from propaganda, but the alternative reality falsely claims that they do not, i.e., that they remain malleable to the elite's lies. In both cases, our main results continue to hold.

Receptive voters are a minority. In the main text below we assume that $\alpha < 0.5$, i.e., receptive voters are a minority. But in Appendix A.4 we show that our main results hold for $\alpha > 0.5$ as well, albeit may require mixed strategies. We further note that for the pure strategy equilibrium we only need that receptive voters *believe* $\alpha < 0.5$, i.e., that only a minority see through the conspiracy, even if the true α is larger. Real-world conspiracy theories often assume that believers are a minority (Douglas et al. 2019).

Belief changes. We assume that propaganda can exogenously change the prior beliefs of receptive voters. This assumption is consistent with the descriptive evidence in Section 2.1 that supporters tend believe in the specific alternative reality disseminated by the politician. However, it is reasonable to assume that misbeliefs are also shaped by voters' demand. To address this issue, in Appendix A.7 we develop a simple model of the demand for misbeliefs based on motivated beliefs. This model provides microfoundations for the reduced-form framework presented here.

3 Results

3.1 Equilibrium

We will characterize the equilibrium for $\pi < 1$ large. Empirically, this is the right parameter range, as the signal of the elite is plausibly fairly informative but imperfect. From the perspective of the analysis, assuming that π is large means that we can simplify some derivations by working them out for $\pi = 1$ and then using arguments based on continuity.

Assumption 1. The elite wants to remove the politician irrespective of his type:

$$\kappa > 1.$$

This assumption captures that the ideological disagreement between the politician and the elite is large. Recalling from (1) that the utility of the elite is $(\theta_c - \kappa) \cdot \bar{\mu}(\theta_c = 1 | \hat{p}, \hat{s})$, since the assumption implies that $\theta_c - \kappa < 0$ for any value of θ_c, it implies that the elite—if it can influence voters—wants to minimize voters' average belief that the politician is good. This assumption ensures the incentive compatibility of elite criticism in the AR.

Assumption 2. For π approaching 1, the cost of propaganda is smaller then its gain:

$$f < \alpha \hat{q}_c.$$

Here, as we will explain in detail below, $\hat{q}_c = q_{ar} q_c / (q_{ar} + q_r(1 - q_c))$ is the persuaded voter's limiting posterior belief (as $\pi \to 1$), after observing propaganda and criticism, that the politician is good. Assumption 2 captures that propaganda has the potential to improve outcomes for the politician. The left-hand side is the cost of propaganda, while the right-hand side is the limit of the gain from propaganda as π approaches one. This gain derives from increasing the beliefs about the politician for the share α of receptive voters, from (approximately) zero to (approximately) $\hat{q}_c$. As we explain after stating the result, this assumption ensures the incentive compatibility of the politician's equilibrium strategy.[7]

[7] Assumption 2 implies that when the share α of voters persuadable by propaganda is higher, even a lower q_{ar}, i.e., less persuasive propaganda, is sufficient to incentivize the politician.

Proposition 1. *If Assumptions 1 and 2 hold, and $\alpha < 0.5$, then there exists $\bar{\pi} < 1$ such that for $\pi > \bar{\pi}$ in the unique PPO equilibrium*

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

All proofs are in the Appendix. At a high level, the intuition for the result is as follows. In reality (part 1 of the result), the good politician has no reason to send propaganda as he will most likely be praised by the elite. The bad politician, who will likely be criticized by the elite, does have an incentive, and will do so by Assumption 2 if propaganda succeeds in discrediting criticism. But discrediting elite criticism requires a persuasive alternative explanation for that criticism: here an elite conspiracy (part 2 of the result). For this conspiracy theory to be persuasive, it is necessary that members of the elite, if they could, would in fact conspire to act against the politician. This is ensured by Assumption 1 which states that members of the elite sufficiently dislike the politician. The narrative then is that conspiring elite members always criticize, leaving the politician no choice but to spread propaganda to counter the elite's lies.

To see more precisely how discrediting works, note that in equilibrium, the receptive voter's posterior about θ_c after propaganda and elite criticism equals

$$\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 0, \theta_m = P) = \frac{q_c q_{ar}}{q_{ar} + q_r \pi (1 - q_c)}. \tag{4}$$

To understand this expression, recall that since $\hat{p} = 1$, the receptive voter is persuaded ($\theta_m = P$) and assigns prior q_{ar} that reality is AR. His posterior belief that the politician is good conditional on propaganda and criticism is then formed by Bayes' rule. The numerator measures the joint probability that (i) the politician is good (q_c), (ii) propaganda, which as the politician is good only happens in the AR (q_{ar}), and (iii) criticism, which always happens in the AR.[8] The denominator measures the probability of propaganda and criticism. In the AR (q_{ar}) the politician always sends propaganda (if he can) and the elite always criticizes, explaining the first term. In the R (q_r), the bad politician sends propaganda ($1 - q_c$) and the elite criticizes when it receives a correct signal (π), explaining the second term.

The key is that as $\pi \to 1$, these beliefs converge to

$$\hat{q}_c = \frac{q_c q_{ar}}{q_{ar} + q_r(1 - q_c)} > 0 \tag{5}$$

so that even as the elite's message becomes arbitrarily precise, the persuaded voter's beliefs after elite criticism remain bounded away from zero. This is because the persuaded voter entertains the possibility of the AR, and in the AR the elite sends criticism even when the politician is good. This limits the perceived informativeness of the elite's message for the persuaded voter. It follows that $\hat{q}_c$ measures (for π large) the effectiveness of discrediting, explaining why it appears in Assumption 2.

3.2 Detailed logic of equilibrium

To fully flesh out the equilibrium logic, we derive voter beliefs and explain how these beliefs ensure incentive compatibility for the elite and the politician.

Voter beliefs. In the proposed equilibrium, the receptive voter's posterior, *absent propaganda* ($\hat{p} = 0$), as a function of the elite's message $\hat{s}$ is

$$\mu_{rec}(\theta_c = 1 | \hat{p} = 0, \hat{s}, \theta_m = N) = \hat{s} \frac{\pi q_c}{\pi q_c + (1 - \pi)(1 - q_c)\beta} + (1 - \hat{s}) \frac{(1 - \pi) q_c}{(1 - \pi) q_c + \pi(1 - q_c)\beta}. \tag{6}$$

On the left-hand side, note the receptive voter's mind type θ_m: because $\hat{p} = 0$, he is normal (N) and retains his prior that reality is R. On the right-hand side, the first term is active when the voter receives a good message from the elite ($\hat{s} = 1$). Such a message typically comes when the politician is good, but may also come when the politician is bad if the elite's signal is incorrect. However, in the latter case, the politician must not be able to send propaganda, otherwise in the proposed

[8] These terms should also be multiplied by $1 - \beta$ to reflect that the politician can send propaganda, but all terms in the denominator should also be multiplied by $1 - \beta$ so we divided through with it.

profile we would observe $\hat{p} = 1$. The formula then follows via Bayes' rule. The numerator is the probability that the politician is good (q_c) and the signal is correct (π); while the denominator also includes the probability that the politician is bad ($1 - q_c$), the signal is incorrect ($1 - \pi$) and the politician cannot send propaganda (β). The second term, active when the elite sends a bad message ($\hat{s} = 0$), follows analogous logic. Observe that as π approaches one, these beliefs converge to $\hat{s}$: for large π the elite's report almost fully reveals the politician's type.[9]

The receptive voter's posterior, *after propaganda*, as a function of the elite's message $\hat{s}$ is

$$\mu_{rec}(\theta_c = 1 | \hat{p} = 1, \hat{s}, \theta_m = P) = (1 - \hat{s}) \frac{q_{ar} q_c}{q_{ar} + q_r \pi (1 - q_c)}. \tag{7}$$

Equation (4) already derived these beliefs after elite criticism ($\hat{s} = 0$). In the case of elite praise ($\hat{s} = 1$) the expression is zero: the AR elite never sends praise, and in R only the bad politician sends propaganda.

Finally, the beliefs of the unreceptive voter are similar to those of the receptive voter absent propaganda (6), except that the unreceptive voter does not observe propaganda and hence does not infer from its absence, so that we do not have the β factors in the denominator.

Incentive compatibility of the elite. Having characterized beliefs, we explain why the elite follows the proposed equilibrium. In R, the behavior of the elite is straightforward: because its members are atomistic and cannot influence voter beliefs, they report truthfully. In the AR, since the elite acts as a single actor and wants to lower voter beliefs, sending criticism is incentive compatible if

$$(1 - \alpha) \left[\frac{\pi q_c}{\pi q_c + (1 - \pi)(1 - q_c)} - \frac{(1 - \pi) q_c}{(1 - \pi) q_c + \pi (1 - q_c)} \right] > \alpha \frac{q_{ar} q_c}{q_{ar} + q_r \pi (1 - q_c)}. \tag{8}$$

The left-hand side is the elite's gain from criticism: the reduced beliefs of the $1 - \alpha$ unreceptive voters. Inside the brackets, we have the difference between unreceptive voters' belief after praise versus criticism. As noted above, these expressions are similar to those of the receptive voter absent propaganda (6), with the difference that the denominators do not have the β factors. The right-hand side is the elite's loss from criticism: the increased beliefs of the α receptive voters who "see through" the conspiracy. This loss is computed by differencing (7) between $\hat{s} = 1$ and $\hat{s} = 0$.

[9] When taking the limit in the second term, we used that $\beta < 1$.

If π is large, then the left hand side of (8) is close to $1-\alpha$ while the right-hand side is close to $\alpha\hat{q}_c$. Therefore, if $\alpha < 0.5$, as assumed in Proposition 1, then for π large the inequality holds. Intuitively, unreceptive voters—who do not entertain the alternative reality—are manipulable by the elite; and if there are enough of them, then their impact incentivizes the AR elite to criticize.

Incentive compatibility of the politician. Finally, we turn to the politician. In R, as noted above, the good politician who expects praise from the elite has no reason to send propaganda. For the bad politician, sending propaganda is incentive compatible if

$$\alpha\left[\pi\left(\frac{q_{ar}q_c}{q_{ar}+q_r\pi(1-q_c)}-\frac{(1-\pi)q_c}{(1-\pi)q_c+\pi(1-q_c)\beta}\right)+(1-\pi)\cdot\frac{-\pi q_c}{\pi q_c+(1-\pi)(1-q_c)\beta}\right]>f. \quad (9)$$

The left hand side measures the expected gain from propaganda. Propaganda only has an effect on receptive voters (α). For them, propaganda changes expected beliefs about competence θ_c, from the expected value of beliefs absent propaganda (6) to the expected value of beliefs in the presence of propaganda (7). These expectations are computed over the distribution of the elite's message that $\hat{s}=0$ with probability π and $\hat{s}=1$ with probability $1-\pi$. The formula then follows by direct substitution. For the politician to prefer propaganda, this expected gain has to exceed the cost of propaganda f. As π approaches one, the second and third fractions on the left-hand side vanish and the remaining terms converge to $\alpha\hat{q}_c$. By Assumption 2, $\alpha\hat{q}_c > f$, thus for π sufficiently large the bad R politician's incentive compatibility constraint holds.

Next consider the politician's incentive compatibility constraint in the AR. Both the good and the bad types observe the state and know that the elite always sends criticism. This means that their incentive compatibility follows from (9) because they expect more criticism. Formally, on the left-hand side the weight on the first two terms increases from π to 1 while the weight on the third term decreases to zero. It follows that for large π, the AR politicians also send propaganda.

Note that the good politician's choice of propaganda in the AR is key for the updating of the voter, who, if propaganda is to be effective, should not be able to infer from it that the politician is bad. This logic underlies equation (4) and prevents the revelation of the bad R politician's type.

Relaxing the constraint on receptive voters. Proposition 1 focuses on the case when only a share $\alpha < 0.5$ of voters are receptive. However, we show in Appendix A.4 that the model has a unique PPO equilibrium which features propaganda even when $\alpha > 0.5$. This equilibrium is identical to that of Proposition 1 in the behavior of the politician. But for α high, the behavior of the elite in the AR is more complex: they now mix between criticizing and praising the politician. The core intuition is that for α high, the conspiracy theory has to address an internal consistency problem: why should elites lie once they know that most people (receptive voters) see through their lies? In our model, the alternative reality evolves to address this problem by making the elites more cunning. Elites now sometimes tell the truth, to confuse voters and ensure that voters no longer see through their lies. Importantly, our main qualitative predictions continue to hold in this more complex equilibrium.[10]

3.3 Theoretical implications of equilibrium

The equilibrium has a number of new theoretical implications which lead to testable predictions.

Deflection. A core implication is that propaganda can deflect elite criticism.[11]

Corollary 1. *Suppose that Assumptions 1 and 2 hold, $\pi > \bar{\pi}$, and $\alpha < 0.5$. In the PPO equilibrium, propaganda by the bad politician increases voters' beliefs that the politician is good:*

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

Note that $E[.]$ represents objectively correct expectations. The result follows essentially from equation (4), which showed (for π large) that under propaganda, elite criticism does not persuade receptive voters. Thus, propaganda enables bad politicians to remain in power, and by doing so, reduces political accountability. Importantly, this result applies in the presence of independent media that provides reliable information on the politician's type. It follows that in our model, populist propaganda reduces accountability.

Inversion. A second implication of the model is that among receptive voters, propaganda *inverts* the effect of the elite's message: elite criticism increases, while elite praise decreases receptive voters' support for the politician.

[10] We note that conspiracy theories could resolve the internal consistency problem in other ways too, such as by falsely claiming that α is low, i.e., that only a minority are aware of the conspiracy.

[11] For consistency with the Proposition we focus on the $\alpha < 0.5$ case in stating our corollaries, but we show in the Appendix that Corollaries 1, 2 and 3 hold for all values of α.

Corollary 2. *Suppose that Assumptions 1 and 2 hold, $1 > \pi > \bar{\pi}$, and $\alpha < 0.5$. In the PPO equilibrium, in the presence of propaganda, elite criticism strictly increases the receptive voter's support for the politician: $\mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 1) < \mu_{rec}(\theta_c = 1|\hat{p} = 1, \hat{s} = 0)$.*

Formally, this result follows from (7), which shows that the persuaded voter's belief after elite criticism remains bounded away from zero, but after elite praise becomes equal to zero. The former statement holds because propaganda discredits elite criticism; the latter holds because elite praise is not possible in the alternative reality, so that observing it punctures the alternative reality.

The underlying intuition is that for the persuaded voter, the elite's message is primarily informative about the nature of reality, not the politician's type. This force emerges because in our model the voter questions the motives of the elite (i.e., the nature of reality). In particular, because in the AR the elite always criticizes, while in the R (for $\pi < 1$) the elite only sometimes criticizes, elite criticism is more consistent with the AR and increases beliefs in the AR. Then, because in the AR propaganda may come from a good politician, while in the R it always comes from a bad politician, the increased belief in the AR increases beliefs that the politician is good. As this logic makes it clear, $\pi < 1$ is necessary for inversion, because it ensures that elite criticism is relatively more consistent with the AR. We conclude that populist propaganda inverts the standard effect of elite information on receptive voters' beliefs.[12]

Although in our model inversion is driven by beliefs about the politician's competence, in practice there can be an alternative mechanism driven by beliefs about the politician's anti-elite nature. That is, persuaded voters may infer from elite criticism not that the politician is (relatively) competent, but that he is anti-elite. We find this alternative mechanism plausible, but for simplicity we did not incorporate the additional type dimension necessary to model it.

Amplification. The previous result explored how beliefs in the presence of propaganda vary with the elite's message. We next characterize beliefs in the presence of propaganda *on average.* The key insight is that propaganda-induced AR beliefs are amplified by Bayesian updating.

[12] We note that inversion can be broken down into two predictions: that the persuaded voter's beliefs (i) increase after elite criticism, and (ii) decrease after elite praise. Note that (i) cannot hold without (ii): since the persuaded voter updates as a Bayesian, his prior must be a convex combination of his posteriors.

Corollary 3. *Suppose that Assumptions 1 and 2 hold, $\pi > \bar{\pi}$, and $\alpha < 0.5$. In the PPO equilibrium, even though signals are generated by R, the receptive voter's expected posterior, relative to his (propaganda-induced) prior, moves towards the AR: $E[\mu_{rec}(AR|\hat{p}, \hat{s})|\hat{p} = 1] > q_{ar}$.*

This result seems counterintuitive from a Bayesian perspective. A standard Bayesian with the wrong prior, as long as his prior assigns positive probability to the truth, should form posteriors that on average drift towards the truth. Corollary 3 says that here, even though R is always included in the receptive voter's prior, he forms beliefs that on average drift away from the truth.

To understand the result, recall that propaganda increases the receptive voter's prior about the AR to q_{ar}. He then uses this new prior to interpret the sequence of messages he receives. For π large, this sequence is very likely to be propaganda and criticism, which leads him to increase his posterior belief about the AR to (approximately)

$$\hat{q}_{ar} = \frac{q_{ar}}{1 - q_r q_c} > q_{ar}. \tag{10}$$

This increase is non-standard: the sequence of propaganda and criticism contains (for π large) essentially no new information beyond the fact that the voter's prior changed. It emerges because the voter with the new prior still entertains the state of the world that reality is R and the politician is good (which has probability $q_r q_c$) even though the fact that his prior changed already ruled it out. Updating eliminates that state from the voter's mind, increasing his beliefs in the AR (for π large) by a factor $1/(1 - q_r q_c)$. Intuitively, the voter neglects the fact that his prior is large precisely when the likely outcomes are propaganda and criticism. His correlation neglect amplifies AR beliefs because these likely outcomes are more consistent with the AR than with the R.

These arguments imply that the success of propaganda is determined by two factors: (i) planting initial misbeliefs; (ii) a plausible narrative that amplifies misbeliefs by explaining the observed reality better than the true narrative. We note that these two factors are already reflected in Assumption 2. The assumption requires that the posterior belief about the politician's quality $\hat{q}_c$ is large enough, and we can express that posterior, normalized by the prior q_c, as

$$\frac{\hat{q}_c}{q_c} = \frac{q_{ar}}{1 - q_r q_c}. \tag{11}$$

Here q_{ar} is the initial false belief, while $1/(1 - q_r q_c)$, as discussed above, is the amplification.

The mechanisms identified here are related to two lines of research on misbeliefs. First, amplification depends on our timing assumption that first propaganda changes the prior and then the voter updates from the new prior, which is related to the sequential updating assumption of Cheng and Hsiaw (2022) and Koçak (2018) that a person first updates about the credibility of a source and then about the information provided by that source. A key difference is that in our setting the change in the prior is driven by the supply side, leading to predictions about when misbeliefs arise. Second, the logic of amplification is related to the research on persuasion with models, in which models more consistent with the data are found to be more persuasive (Schwartzstein and Sunderam 2021, Aina 2023). Although we also differ in our formal approach, our key contribution to this work is the applied finding that beliefs in politically-supplied conspiracy theories are amplified by observed outcomes.

Comparative statics of presence of propaganda. Proposition 1 shows that bad politicians always choose propaganda, but focuses on the case when the elite strongly dislikes the incumbent politician: $\kappa > 1$ by Assumption 1. We now relax that assumption.

Corollary 4. *Suppose that Assumption 2 holds, $\pi > \bar{\pi}$, and $\alpha < 0.5$. If $1 - \pi < \kappa < \pi$, then there is a unique PPO equilibrium, and in that equilibrium no politician sends propaganda.*

The point here is that when κ is (somewhat) lower than 1, propaganda is no longer used. This is because when κ is in the specified range, the elite wants to keep the politician if and only if he is good. But then the narrative in which the elite conspires to remove a good politician is not plausible and will not be believed by voters. The politician then has no reason to send propaganda. It follows that successful propaganda requires the ideological disagreement κ between the politician and the elite to be sufficiently large.

Comparative statics of effectiveness of propaganda. We turn to explore the conditions under which propaganda is more or less effective. We focus on the effect of q_c, the prior probability that the politician is good, which may be reflected in the politician's baseline popularity.

Corollary 5. *Suppose that Assumptions 1 and 2 hold, $\pi > \bar{\pi}$, and $\alpha < 0.5$. In the PPO equilibrium,*

1. *The receptive voter's belief in the AR after propaganda and criticism, $\mu_{rec}(AR|\hat{p} = 1, \hat{s} = 1)$, is increasing in q_c.*

2. The bad R politician's expected gain from propaganda

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

is increasing in q_c.

The first result is a comparative static of the amplification in Corollary 3 and follows essentially from equation (10). When the politician is more likely to be good (q_c higher), propaganda and criticism are less likely in the R but more likely in the AR, and thus increase more the posterior of the AR. Intuitively, when the politician is more likely to be good, a conspiracy is a more plausible explanation for elite criticism.

The second result, that propaganda is more valuable for the politician when q_c is higher, follows essentially from equation (11). There are two forces. First, as just noted, with q_c higher, propaganda induces a larger belief change towards the AR. Second, belief in the AR is better for the politician, because in that state he is perceived to be good with a higher q_c probability.

An interesting implication of the second result is that as a politician looses popularity, propaganda becomes less effective in shoring up support. When perceived competence falls, the conspiracy becomes a relatively less plausible explanation for elite criticism; and even if it does discredit elite criticism, it does not undo voters' direct perception of the politician's incompetence.

Demand for misbeliefs. A weakness of our theoretical model is that we take the demand for misbeliefs as given. To address this, in Appendix A.7 we develop a microfounded model of the demand side, which is based on the idea of motivated beliefs (Brunnermeier and Parker 2005, Bénabou and Tirole 2006). In this model, a voter who experiences propaganda becomes aware of the elite conspiracy alternative reality, and chooses his prior belief q_{ar} in that alternative reality. This choice is made before the voter updates from propaganda and the elite's message. The voter chooses q_{ar} by trading off his subjective expected utility from the election outcome against a cost of changing his prior. The utility from the election outcome comes from our microfoundation of voter behavior (Appendix A.2). The cost of changing the prior is a function of the expected posterior belief in the AR, capturing the idea that beliefs in the AR may impair decisions in other domains.

We show in Proposition 4 in the Appendix that our equilibrium is robust to incorporating the demand for misbeliefs, but features an endogenously chosen $q_{ar} > 0$.[13] We also show that in this equilibrium the predictions of Corollaries 1-5 continue to hold.

3.4 Evidence

We turn to discuss evidence on the model's implications, focusing on Corollaries 1-4.

Propaganda lowers accountability in democracies. A growing body of evidence documents that populism is associated with reduced accountability. Funke et al. (2023) show that populism reduces GDP per capita and consumption by 10% relative to a plausible non-populist benchmark; yet populists stay in power for twice as long as non-populists. More anecdotally, Donald Trump won the 2024 US election despite being a convicted felon, while Hungary's Orban and Turkey's Erdogan stayed in power for extended periods despite evidence that they eroded democratic institutions (Guriev and Treisman 2022).

To our knowledge, existing theories do not explain populism's association with reduced accountability. Models of populism, including Acemoglu et al. (2013) and Bellodi et al. (2023), work through the logic that populism is a positive signal about the politician, hence cannot easily explain reduced accountability. Models that do predict reduced accountability do so in non-democracies, through the mechanisms of repressing voters or silencing the media (Egorov and Sonin 2024, Guriev and Treisman 2020). For example, in Guriev and Treisman (2020), propaganda paints the politician in a false positive light, and independent media cannot correct this false view because it is silenced. These mechanisms cannot easily explain reduced accountability in democracies that have independent media.

Our model explains reduced accountability with Corollary 1, which predicts that populist propaganda reduces accountability. The reason propaganda works despite an independent media is that it discredits that media. The mechanism of discrediting—a conspiracy theory—is also consistent with evidence: as we discussed in Section 2.1, the populist narrative is often centered around an elite conspiracy. We note that although in this paper we focus on democracies, our mechanism may also play a role in autocracies and hybrid regimes, where it can complement other anti-media

[13] We no longer show that the equilibrium is unique.

	All	Moderate	Conservative
More likely to vote for him	41%	24%	44%
Less likely to vote for him	4%	13%	3%
Not affect whether you vote for him	55%	63%	53%
Observations	488	80	408

Table 2: Impact of indictment on Trump's support by Republicans intending to vote in primary

strategies such as censorship and media capture.

Propaganda inverts the elite's effect on receptive voters. A key fact in contemporary US politics is that during 2023, the growing body of critical evidence against Donald Trump, including four criminal indictments, was accompanied by an *increase* in his support among Republican voters (Swan et al. 2023). Since the indictments were produced by the US legal system, this reaction by the presumptive party of law and order is puzzling. It is even more puzzling when compared to two other salient legal cases against leading politicians. President Richard Nixon, following the Watergate scandal, and New York mayor Eric Adams, following his 2024 criminal indictment, both experienced large reductions in popular support even among supporters of their own party (Franklin 2018, McFadden and Mays 2024). We not aware of other formal models that explain why, following their legal challenges, support increased for Trump but declined for Nixon and Adams.

Our model can explain these facts through its inversion and comparative statics predictions. We first describe how inversion explains the increase in support for Trump and present supporting evidence; and then describe how the comparative statics explain the opposite pattern for Nixon and Adams. The explanation for Trump follows directly from Corollary 2, under the assumptions that Trump spreads propaganda and that Republicans correspond to the model's receptive voters. Then, the Corollary predicts that the indictments—elite criticism—*causally* increase Republicans' support for Trump.

We present two pieces of evidence that support this causality. First, in Table 2 we show results from a 2023 poll investigating the impact of the indictments on Trump's political support (YouGov 2023). Among registered Republicans intending to vote in the primaries, 41% claimed that they would be more likely, and only 4% claimed that they would be less likely, to vote for Trump if he was indicted in the matter of handling classified documents. The effects were large even among moderate Republicans. Thus, Republicans anticipated that their own support would increase in response to critical evidence.

But this evidence is about voters' hypothetical behavior. For evidence on voters' actual behavior, we turn to the impact of Republican politicians' scandals on campaign contributions. Scandals, diffused by the news media, are an example of elite criticism, thus the logic of our model predicts that—given the presence of propaganda—they should increase campaign contributions from receptive voters. To explore this effect, we take Wikipedia's list of political scandals of Republican House candidates during 2017-2022, and select the 11 scandals that are related to sexual misconduct, financial misconduct, election fraud, or violence. These are issues on which probably most voters and elite members agree, thus they correspond to θ_c. We combine these data with donation data from the Federal Election Commission.[14]

We estimate difference-in-differences regressions of the effect of a scandal on donations that come from Trump supporters and from other donors. We define Trump supporters as individuals who donated to the Make America Great Again PAC in the 2020 election campaign. Our control group includes donations to other Republican House candidates in the same period. Table 3 reports the results. Column 1 shows that relative to a control mean of 6.5 percent, the share of donations coming from Trump-supporter donors increased after the scandal by a significant 7.5 percentage points. Columns 2 and 3 show that this increase was largely driven by a significant increase in Trump-supporters' donations of about $20,000 per quarter, with no significant change in other donors' donations. We conclude that the evidence supports the causal link between elite criticism and political views predicted by our model.[15]

Beyond predicting that elite criticism should increase Republicans' support for Trump, the model also predicts the underlying mechanism: that elite criticism should increase Republicans'

[14] We use quarterly data on contributions made by private individuals to the election committees of congressional candidates.

[15] A possible alternative explanation is that the scandal increased election competitiveness, and competitiveness affected donations. Two pieces of evidence speak against this. First, as Table 3 shows, the effect is concentrated among Trump-supporters, and it is not clear why they should care more about the election. Second, in Appendix A.10 we show that when competitiveness increases because of redistricting, there is no analogous impact on donations.

	Trump donors	Trump donors	Other donors
	Share	Amount (1,000 dollars)	
Scandal	0.075***	20.35**	-9.86
	(0.009)	(9.88)	(16.59)
Representative and quarter f.e.	yes	yes	yes
Control mean	0.065	16.06	119.0
Observations	3,393	4,382	4,382

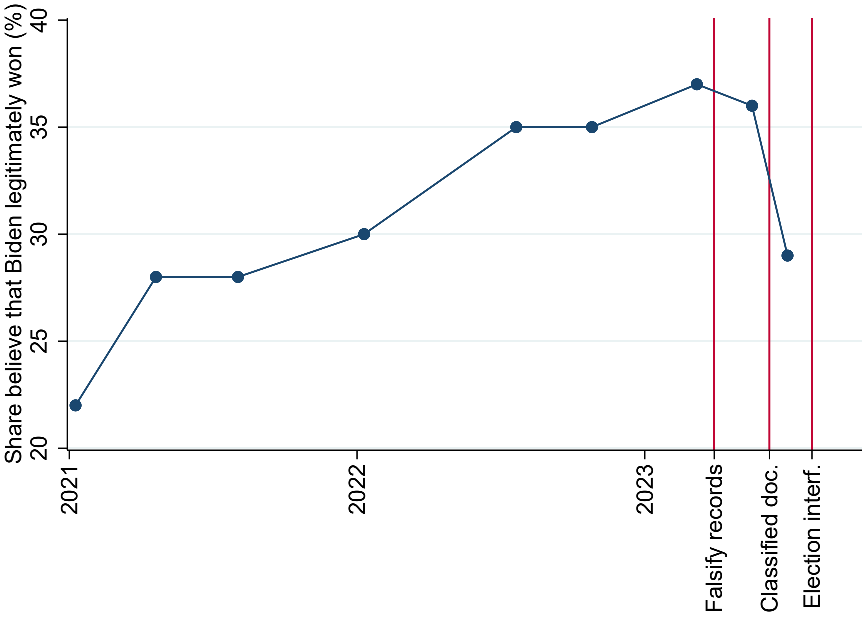
Note: Observations are house representative by quarter cells. Representatives with a scandal during 2017-22 are in the sample in a one-year window around their scandal. Representatives without a scandal are in the sample in all quarters in which they were in office during 2017-22. Scandal is one for representatives that have a scandal in the quarters after the scandal. Column 1 is restricted to observations with non-zero total donations. 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.

Table 3: Impact of scandals on contributions from Trump-supporter and other donors

beliefs in the alternative reality. This prediction is consistent with survey evidence we present in Figure 1, which plots, over time, the share of Republican-leaning voters who believe that Biden legitimately won the 2020 presidential election. The share steadily increases during 2021 and 2022, but sharply drops in the summer of 2023, exactly around the time of the first three Trump indictments. That is, as predicted by the model, beliefs in the conspiracy theory that the 2020 election was stolen (AR beliefs) increased precisely at the time of the indictments (elite criticism). Although this evidence does not conclusively prove that the indictments caused the increase in misbeliefs, the sharp change in the absence of other salient events is suggestive of causality.

Why did receptive voters respond differently to the elite criticism of Trump than they did to the elite criticism of Nixon or Adams? Although both Nixon and Adams attempted to use conspiracy theories to deflect criticism (Shabecoff 1974, Mays 2024), they were unsuccessful. Corollary 4 says that propaganda is only successful when the cleavage between the politician and the intellectual elite is sufficiently large. But both Nixon and Adams represented the more educated party—Republicans in the early 1970s, Democrats today (Kuziemko, Marx and Naidu 2022)—suggesting

Figure 1: Share of Republican-leaning voters who believe that Biden legitimately won in 2020



Source: Nine CNN opinion polls conducted by SSRS between January 2021 and August 2023 (SSRS 2023). Vertical lines indicate dates of indictments against Donald Trump in 2023: (1) in March 30 for falsifying business records; (2) in June 8 for mishandling of classified documents; (3) in August 1 for attempting to overturn the 2020 US presidential election.

that the cleave between them and the elite was small, making it implausible that the elite would conspire to remove them from power. We conclude that the model is consistent with the presence of inversion for Trump and its absence for Nixon and Adams.

Finally, we relate our results on inversion to the research on the "backfire effect" that corrective information can sometimes lead individuals to more strongly endorse a misbelief. Early work showed evidence for this backfire effect, but more recent research suggests that corrective information is usually somewhat effective in correcting beliefs (Nyhan 2021). Our inversion prediction is different from the backfire effect: it is not about the impact of corrective information but about the impact of elite criticism of the politician. Corrective information is often not elite criticism of the politician. In fact, our model suggests the comparative static that the backfire effect should be stronger when the salient interpretation of corrective information is elite criticism of the politician.

Politically supplied misbeliefs are amplified by outcomes. Beliefs in the alternative reality are surprisingly widespread: e.g., as shown in Figure 1, the overwhelming majority of Republicans believe in the conspiracy theory that Biden did not legitimately win the 2020 elections. Such widespread misbeliefs seem difficult to explain purely with propaganda's ability to move priors. But they may be easier to explain with Corollary 3, which predicts that realized outcomes amplify those prior beliefs. We do not have direct evidence on amplification, but we note that in line with its logic, alternative realities tend to be supplied precisely in conditions in which they are more consistent with realized outcomes. Returning to the contexts of Section 2.1, in the US, the deep state conspiracy was supplied around the time of the Trump indictments; in Hungary, the Soros-Brussels conspiracy to import immigrants was supplied around the time of the European migrant crisis; in Israel, the judiciary-media conspiracy was supplied around the time of the legal cases against Netanyahu. In each of these cases, observed outcomes—indictments, immigration, court cases—were almost inevitable in the alternative reality and hence should have strengthened beliefs in that alternative reality.

Propaganda is only used in divided societies by anti-elite politicians. Corollary 4, in combination with Proposition 1, show that propaganda is only used if disagreement between the politician and the elite is large, that is, if (i) there is large division in society, and (ii) the politician is on the other side of the elite in that division. Part (i) highlights a new mechanism that links societal cleavages to populism. A novelty relative to prior work is that the cleavage need not be about differences in income, with the elite being the rich; but may be about differences in cultural values, with the elite being intellectuals. Thus, in the US context, our model formalizes the narrative that Democrats' move to the left enabled right-wing populism (Norris and Inglehart 2019), through the logic that the resulting increase in cultural disagreement κ made the elite conspiracy narrative plausible. Part (ii), as we noted above, helps explain the presence of inversion under Trump and its absence under Nixon and Adams, through the logic that the latter politicians represented pro-elite parties.

4 Applications

We turn to develop two applications of our model: endogenizing the conspiratorial nature of the alternative reality, and studying the impact of populism on government policy. In each application, we introduce additional assumptions to capture new features of the environment, but do not change our fundamental assumptions concerning the alternative realities.

4.1 Endogenizing the conspiracy theory

In our basic model, the AR features a conspiracy only by assumption. Moreover, for our qualitative results, this assumption is not strictly necessary: we could obtain our qualitative results in a model without a conspiracy, in which in the AR the elite has a lower cost of lying. Thus, incorporating a conspiracy into the model may seem superfluous. Here, we argue that conspiracy theories are in fact a natural implication of our framework, justifying our approach and helping to explain prevalence of political conspiracy theories.

The basic insight we develop here is that the elite conspiracy solves a collective action problem. This problem arises because a lie about the politician's competence by any given elite member benefits every other elite member, since they all benefit from lower support for the politician. The ability to coordinate allows these externalities to be internalized, strengthening the incentives to lie. As a result, the conspiracy-based alternative reality can explain away a wider range of criticism: even credible evidence like an indictment that individual elite members would not, but collectively the "deep state" might have the incentive to manufacture.

To explore these issues formally, we extend the model to allow for two different types of alternative realities. In the first, elite members have a lower lying cost but do not have the ability to coordinate; in the second, they also have the ability to coordinate. In addition, we introduce a variable that measures the credibility of the evidence the elite provides in support of its message: a publicly known fabrication cost that each elite member has to pay in order to send a false message.

Model. Modeling the lying-cost alternative reality requires that each elite member has some individual-level incentives to manipulate. We thus assume that the elite consists of a finite number of members N, and each of them accesses a mass $1/N$ of voters. We further assume that there is a non-infinitesimal lying cost χ which can be written as the sum of a fabrication cost χ_f and an integrity or honor cost χ_h. The fabrication cost χ_f is the cost of manufacturing the evidence presented in the elite's message—such as videos of intensive care units during Covid—which we assume is known by the voter and cannot be changed by the alternative reality. The honor cost χ_h is the private cost to an elite member for telling a lie. In addition, there is an organizing cost χ_o which each elite member has to pay if they conspire. In the objective reality both χ_o and χ_h are prohibitively high, so that elite members do not conspire and tell the truth.

The politician chooses to send propaganda about one of the following two alternative realities.

1. Lying cost AR. In this AR, χ_o continues to be prohibitively high but $\chi_h = 0$. The cost of sending propaganda to make the voter believe in this AR is $f' < f$.
2. Conspiracy AR. In this AR both $\chi_o = 0$ and $\chi_h = 0$. The cost of sending propaganda to make the voter believe in this AR is f. Since $\chi_o = 0$, we assume that in this AR the elite always coordinates if it is in their joint interest, i.e., there are no coordination problems.

We will denote the lying-cost AR by AR1 and the conspiracy AR by AR2. There are three reality states: $\theta_r \in \{R, AR1, AR2\}$ such that the objective probability of $\theta_r = R$ is 1. If the receptive voter receives propaganda on AR1, his prior of AR1 increases to q_{ar}; if he receives propaganda on AR2, his prior of AR2 increases to q_{ar}. His prior of the other possible AR remains zero. The politician in any reality state can send either AR1 or AR2 propaganda. This is the natural generalization of our model to the setting with multiple alternative realities.

Since the elite has a finite number of members, average voter beliefs become

$$\bar{\mu} = \frac{\sum_{j=1}^{N} \bar{\mu}_j(\hat{p}, \hat{s}_j)}{N}$$

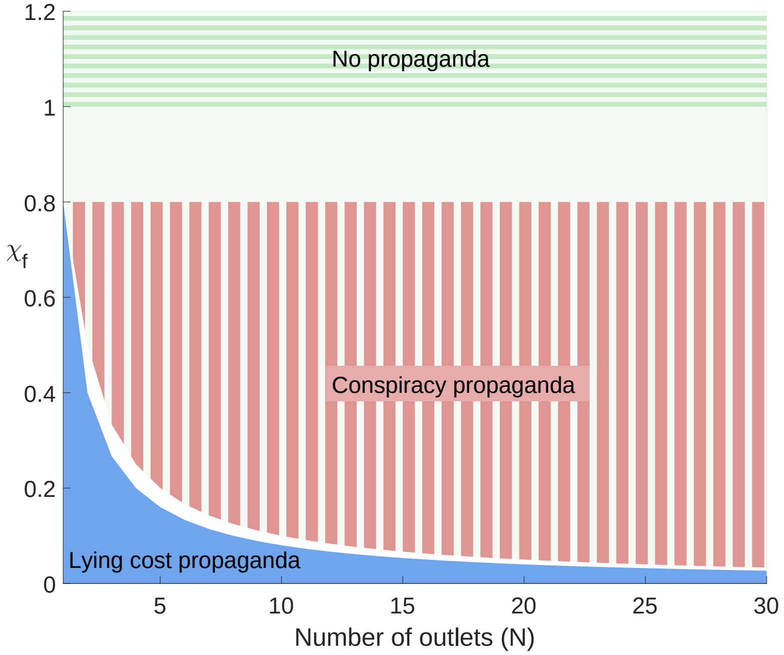
where $\bar{\mu}_j(\hat{p}, \hat{s}_j)$ is the average belief among voters influenced by elite member j. We assume $N > 1$.

Given the non-infinitesimal lying costs, elite member j's utility becomes

$$U_{ej} = (\theta_c - \kappa) \cdot \bar{\mu} - \chi_f \cdot 1_{\{s_j \neq \hat{\theta}_c\}} - \chi_h \cdot 1_{\{\theta_r = R\}} 1_{\{s_j \neq \hat{\theta}_c\}} - \chi_o \cdot 1_{\{\theta_r \neq AR2\}} 1_{organize_j}. \tag{12}$$

The first term captures the impact of the average voter belief $\bar{\mu}$. The remaining terms reflect that the elite must pay a fabrication cost χ_f for fabricating a lie in all realities, an honor cost χ_h for not telling the truth in R, and an organizing cost χ_o for attempting to organize in R and AR1. Since χ_o and χ_h are prohibitively high, in equilibrium the last two costs are never paid. The politician's utility is still given by (3) with $\bar{\mu}$ governing voter beliefs.

Figure 2: Endogenous AR



Proposition 2. *Under Assumptions 1-2, if $\alpha < 0.5$ then, in the following ranges for χ_f, for π large enough, the unique PPO equilibrium is such that*

1. *If $\chi_f < (1-2\alpha)/N$, then in reality the bad politician sends lying cost propaganda;*

2. *If $1/N < \chi_f < (1-2\alpha)$, then in reality the bad politician sends conspiracy propaganda;*

3. *If $1 < \chi_f$, then no politician sends propaganda.*

Figure 2 is helpful for understanding the result. The horizontal axis is N, the number of elite members, and the vertical axis is χ_f, the publicly known fabrication cost. The first part of the Proposition says that for χ_f low, i.e., when the evidence supporting the elite's message is easy to fabricate, lying cost propaganda is sufficient. In this case, the narrative that elite members have "no honor" is sufficient to explain away the weak evidence. More precisely, because each elite member influences a share $1/N$ of voters, each has a non-negligible gain from manipulating these voters.

Hence, absent an honor cost and given the low fabrication cost, each is willing to fabricate a fake message. This range corresponds to the blue (solid) region in the Figure.

The second part of the Proposition says that for χ_f in the middle range, the equilibrium uses conspiracy propaganda. In this range lying cost propaganda no longer works: the individual-level gain to each elite member no longer covers the fabrication cost. But conspiracy propaganda works, because if elite members act collectively, then the individual-level gains increase by a factor of N. Intuitively, each elite member now internalizes that her action benefits all other elite members, and thus has higher-powered incentives to fabricate her message. This is the equilibrium in the red (vertical stripes) region in the Figure. Observe that the higher N, i.e., the more fragmented the elite, the wider the range of the conspiracy equilibrium. Since in practice N is likely to be large, the Proposition suggests that the conspiracy AR is a likely outcome.

The third part of the Proposition says that for χ_f high, corresponding to the green (horizontal stripes) region, propaganda is not used. At such a high cost, even a collectively acting elite does not have sufficient incentives to fabricate lies.[16]

Implications and evidence. The Proposition has three main implications. First, it predicts that misbeliefs should often feature conspiracy theories. Conspiracy theories are indeed prevalent (Douglas et al. 2019) and we are not aware of other formal theories that explain their emergence.

Second, the Proposition predicts that increasing the credibility of evidence need not correct beliefs. This is because the politician can respond to an increase in the fabrication cost χ_f by escalating the alternative reality from a lying cost AR to a conspiracy AR, and explain away the more credible evidence with a more powerful elite.[17] Through this logic, alternative realities can resist evidence. Thus, we formalize the argument of Sunstein and Vermeule (2009) that maintaining a conspiracy theory in the face of contradictory evidence requires an ever-widening conspiracy.

Third, the proposition implies that propaganda—because it often uses a conspiracy theory—can lead to distrust in science and the non-adoption of best practices. This is because the conspiracy

[16] As illustrated by the white areas between the three regions in the Figure, the Proposition does not cover the full range of χ_f values. In the intermediate ranges mixed equilibria are possible, which did not seem central to our message. We also note that the π large condition in the Proposition is required by χ_f, rather than uniformly.

[17] More generally, and outside our current model, the politician could escalate the scale of the conspiracy theory by claiming that it involves more actors.

narrative makes the elite more powerful in other domains too, which affects the behavior of the voter in those domains. Once the voter believes that the elite can conspire, he will suspect that even seemingly credible elite messages in the health or climate domains may be driven by the elite's private interest. For example, reports about climate change by scientists, which seem prohibitively expensive to fabricate individually, may be driven by their collective desire to control the population (Uscinski, Douglas and Lewandowsky 2017).

This last point helps explain the evidence that misbeliefs under populism go beyond politics, including Republicans' attitudes in the health and climate domains. For example, Allcott, Boxell, Conway, Gentzkow, Thaler and Yang (2020) show that under Covid Republicans were less likely to engage in social distancing; Wallace, Goldsmith-Pinkham and Schwartz (2022) show that they had higher excess death rates attributable to Covid; and Hotez (2023) shows the persistence of Covid-denialism in the face of credible evidence. Our model explains these facts through the logic that populism causes distrust in the elites. This is in contrast to prior work that emphasized the causality from distrust to populism (Bellodi et al. 2023, Guiso, Helios, Morelli and Sonno 2023). Our chain of causality suggests that eliminating propaganda should improve trust in science.

4.2 Government policy under conspiratorial populism

In our second application we explore how conspiratorial populism shapes government policy. This is a fundamental question since much evidence shows that populism is associated with large economic and non-economic costs (Guriev and Papaioannou 2022) which presumably derive from harmful policies. Our model identifies two mechanisms through which populism leads to harmful policies. First, there is a direct effect of reduced accountability: populism enables incompetent, corrupt, or authoritarian politicians to maintain political power, and implement incompetent, corrupt, or authoritarian policies. Second, on top of this basic effect, our model predicts that populist politicians will often choose harmful policies *purely to trigger elite criticism* and thereby strengthen beliefs in the alternative reality. We turn to formally develop this second insight.

To incorporate government policy into the model in a simple way, we assume that the bad politician can set policies that make his bad type more visible to the elite. Formally, in stage 1, the bad politician, simultaneously with his propaganda decision, can take an action $e \in \{0, 1\}$, where $e = 1$ represents his intent to set a bad policy. We assume that $e = 1$ has vanishingly small cost to the politician. Although e is not directly observable, it reduces the quality of the policy mix, and thereby increases the probability that the elite's signal $\hat{\theta}_c$ is correct to $\pi' > \pi$.[18] The key here is that setting $e = 1$ invites elite criticism.

We make one other substantive departure from the basic model: we assume that the politician cares more than the elite about the beliefs of receptive voters. Formally, the politician maximizes

$$U_p = \tilde{\mu}(\theta_c = 1|\hat{p}, \hat{s}) - f \cdot p, \tag{13}$$

where $\tilde{\mu}$ is the weighted average of receptive and unreceptive voters' beliefs with a new weight α'

$$\tilde{\mu}(\theta_c = 1|\hat{p}, \hat{s}) = \alpha' \cdot \mu_{rec}(\theta_c = 1|\hat{p}, \hat{s}) + (1 - \alpha') \cdot \mu_{un}(\theta_c|\hat{s})$$

and we will assume that $\alpha' > \alpha$ by a sufficiently large margin. This assumption creates a wedge between the incentives of the politician (who weight receptive voters with α') and those of the elite (who weight receptive voters with α). One interpretation of this wedge, already suggested before, is that receptive voters incorrectly perceive that $\alpha < 0.5$, even though the true α is larger. Since the perceived α governs the incentives of the elite in the AR, this misperception creates the desired wedge. Another interpretation is that the politician also cares about winning a primary election, and hence overweights the beliefs of his core (receptive) supporters.[19] And a third interpretation is that the elite cares disproportionately about some part of the audience, such as foreigners or donors, who are plausibly less receptive to propaganda. Under all three interpretations, the wedge allows both the politician and the conspiring elite to gain from elite criticism.

Proposition 3. *Under Assumptions 1 and 2, if $\alpha < 0.5$ and $\alpha' > 1/(1 + \hat{q}_c)$, then for π large enough and $\pi' > \pi$, in the unique PPO equilibrium*

1. *All propaganda and message choices are as in Proposition 1.*

[18] Formally, define the realized policy as $\theta_c - \phi e + \xi$ where ϕ is the effectiveness of the politician's action and ξ is a random policy shock. The elite's signal is $\tilde{\theta}_c = 1\{\theta_c - \phi e + \xi > \tau\}$, where the τ threshold is such that $\Pr[\xi > \tau - 1] = \pi$, $\Pr[\xi > \tau] = 1 - \pi$, and $\Pr[\xi > \tau + \phi] = 1 - \pi'$.

[19] The assumption that the politician's core supporters are the receptive voters seems realistic in that Republicans are the voters who believe in the elite conspiracy.

2. *The bad politician chooses to set a bad policy ($e = 1$) if and only if reality is R and he sends propaganda.*

That is, populist propaganda drives bad policies. To see the intuition, first note that because $\alpha < 0.5$ we are in the parameter range of Proposition 1. In this range, by equation (8), it is a dominant strategy for the elite in the AR to always criticize the pro-voter politician, because among the majority $1 - \alpha$ of unreceptive voters elite criticism reduces voter beliefs. In contrast, as Corollary 2 demonstrated, among the minority α of receptive voters, elite criticism increases voter beliefs. Thus, for a politician who puts a sufficiently large weight on receptive voters, elite criticism is beneficial. This force induces the politician to take the bad policy action and trigger criticism.

Implications and evidence. The key empirical prediction is that populist politicians, both because their type is bad and in order to invite elite criticism, choose harmful policies. Harmful policies are also a prediction in the Acemoglu et al. (2013) model of populism, where populists signal their independence from the elite using policies that disproportionately harm the elite. The key difference is that our model does not make any assumption about the nature of the bad policy, and is thus consistent with harmful policies that *do not* disproportionately harm the elite.

Because of this difference, our model can help explain the puzzling fact that populists, despite their pro-people rhetoric, do not appear to be siding with "the people:" their policies seem to hurt the non-elite at least as much as they hurt the elite. Macro evidence on this comes from Funke et al. (2023), who show not only that populists reduce GDP per capita, but also that they fail to reduce inequality. Thus, populists seem to cause equal harm to the non-elite and the elite.

Populists also favor specific policies that seem to disproportionately harm the non-elite. Perhaps the most direct example is corruption. Populism's erosion of democratic institutions (Funke et al. 2023) can enable large-scale corruption, and indeed, populism is associated with a large increase in executive corruption (Zhang 2024). In turn, stealing government funds plausibly harms those the most who rely on government services the most, i.e., the non-elite. A second example is tariffs, which are commonly used by populists (Funke et al. 2023) including Donald Trump. Although tariffs do not necessarily hurt the non-elite, there are good reasons why they might: they raise import prices, and induce retaliatory tariffs targeted at the populist's non-elite supporters. Consistent with this logic, Fajgelbaum et al. (2019) find in a model-based evaluation that the tariffs of the first Trump administration, due to tariff retaliations, affected tradeable sector workers in heavily Republican countries the most negatively.[20] A third example is climate policy. The Biden Administration's climate bill, the Inflation Reduction Act, was opposed by Republican representatives and suspended by President Trump, despite its widely acknowledged benefits for blue-collar workers in Republican-leaning states (Friedman et al. 2025).

To our knowledge, this evidence is not explained by prior models. Although our model does not predict which bad policies get implemented, it is consistent with bad policies that harm the non-elite as much as they harm the elite, and in this sense can explain the evidence.[21] We conclude that the current wave of populism may generate substantial harmto both the elite and the non-elite.

5 Conclusion

In this paper we built a new model of populism as a conspiracy theory. In our model, a politician can supply a false alternative reality claiming that members of the elite conspire to attack him because they disagree with his ideology. We show that, among voters receptive to it, this alternative reality can discredit the elite's truthful message about the politician. In turn, discrediting is beneficial for a "bad" politician because it enables him to remain in power. Through this logic, our model explains two previously unexplained facts about populism: the political use of conspiratorial narratives, and their association with reduced accountability.

A key prediction of the model is that conspiratorial propaganda inverts the effect of elite criticism among receptive voters, so that elite criticism increases these voters' support for the populist. The underlying intuition is that for receptive voters, the elite's message is primarily informative about the nature of reality, and elite criticism is more consistent with the alternative reality. This result explains Republicans' increased support for Trump after the indictments. The model also explains

[20] Grossman and Helpman (2021) explain populism-induced tariffs in an identity-based model of trade, but in their model tariffs do not materially harm the non-elite.

[21] Neither does our model predict which of its two mechanisms—lack of accountability or the desire to trigger the elite—is responsible for specific policies. But we note that one channel for the second mechanism may be the populist's personnel policy. Selecting experts for key policy roles can invite elite praise and should therefore be avoided; and the selection of non-experts leads to harmful policies.

the absence of such an increase for Nixon after Watergate, through the logic that Nixon did not have sufficient ideological cleavage with the elite to make an ideology-motivated attack plausible. Another prediction of the model is that beliefs in the alternative reality tend to strengthen in response to realized outcomes, because the alternative reality was chosen in anticipation of those outcomes. This result helps explain why beliefs in political conspiracy theories are widespread.

We then developed two applications of the model. In the first, we showed that alternative realities often endogenously feature conspiracies in order to better resist evidence. Thus, our model provides a formal explanation for the emergence of political conspiracy theories. An implication is that increasing the credibility of evidence can make the alternative reality conspiratorial, which in turn can create distrust in the elite beyond politics. This result helps explain Republicans' general distrust in science.

In our second application, we studied government policy under conspiratorial populism. We found that populists may set harmful policies that disproportionately harm the non-elite, both because—given reduced accountability—they can, and because doing so triggers elite criticism. This result helps explain why populism is associated with economic underperformance without meaningful reductions in inequality, as well as policy choices such as corruption, tariffs, and anti-environmentalism which may disproportionately harm the non-elite. We conclude that our theoretical results shed light on a number of key facts about populism.

Our model studied the supply side of populism. But much empirical research shows that populism is more likely to emerge following economic crises (Guriev and Papaioannou 2022), suggesting that the demand side also plays a role. Building a psychologically realistic model of the demand side that informs this evidence is an important topic for future work.

In this paper, we used the framework of a false alternative reality to study populist ideology. The same framework may be used to study other ideologies as well. One possible example is nationalism. Aiming to deflect criticism or initiate collective action, political leaders may demonize the citizens of the other country, an alternative reality that captures some elements of nationalism. Modeling this alternative reality may lead to predictions about the emergence and persistence of conflict, based on the idea that nationalistic ideology leads to a misinterpretation of other countries' actions. More generally, formalizing other ideologies as strategic alternative realities is a potentially important avenue for future research.

References

Acemoglu, Daron, Georgy Egorov, and Konstantin Sonin, " A Political Theory of Populism ," *The Quarterly Journal of Economics*, 02 2013, *128* (2), 771–805.

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya, " Radio and the Rise of The Nazis in Prewar Germany," *The Quarterly Journal of Economics*, 07 2015, *130* (4), 1885–1939.

Agranov, Marina, Ran Eilat, and Konstantin Sonin, "Information Aggregation in Stratified Societies," Working Paper 31510, National Bureau of Economic Research July 2023.

Aina, Chiara, "Tailored Stories," Working Paper, Harvard University 2023.

Ajzenman, Nicolás, Tiago Cavalcanti, and Daniel Da Mata, "More than words: Leaders' speech and risky behavior during a pandemic," *American Economic Journal: Economic Policy*, 2023, *15* (3), 351–371.

Allcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang, "Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic," *Journal of public economics*, 2020, *191*, 104254.

Allen, Jonathan, "Awaiting possible indictment, Trump rallies in Waco and vows to 'destroy the deep state'," NBC News, `https://www.nbcnews.com/politics/awaiting-possible-indictment-trump-rallies-waco-rcna75684`, 2023.

Angelucci, Charles and Andrea Prat, "Is journalistic truth dead? Measuring how informed voters are about political news," *American Economic Review*, 2024, *114* (4), 887–925.

Ash, Elliott, Sharun Mukand, and Dani Rodrik, "Economic Interests, Worldviews, and Identities: Theory and Evidence on Ideational Politics," Working Paper 29474, National Bureau of Economic Research November 2021.

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya, "Facts, alternative facts, and fact checking in times of post-truth politics," *Journal of Public Economics*, 2020, *182*, 104123.

Bellodi, Luca, Massimo Morelli, Antonio Nicolò, and Paolo Roberti, "The shift to commitment politics and populism: Theory and evidence," *BAFFI CAREFIN Centre Research Paper*, 2023, (204).

Bénabou, Roland, "Groupthink: Collective delusions in organizations and markets," *Review of economic studies*, 2013, *80* (2), 429–462.

——— , **Armin Falk, and Jean Tirole**, “Narratives, imperatives, and moral reasoning,” Technical Report, National Bureau of Economic Research 2018.

Besley, Tim and Torsten Persson, “The rise of identity politics,” Working paper, London School of Economics and Stockholm School of Economics 2021.

Besley, Timothy and Andrea Prat, “Handcuffs for the Grabbing Hand? Media Capture and Government Accountability,” *American Economic Review*, June 2006, *96* (3), 720–736.

Blouin, Arthur and Sharun W. Mukand, “Erasing Ethnicity? Propaganda, Nation Building, and Identity in Rwanda,” *Journal of Political Economy*, 2019, *127* (3), 1008–1062.

Bonomi, Giampaolo, Nicola Gennaioli, and Guido Tabellini, “Identity, Beliefs, and Political Conflict,” *The Quarterly Journal of Economics*, 09 2021, *136* (4), 2371–2411.

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

Bénabou, Roland and Jean Tirole, “Belief in a Just World and Redistributive Politics*,” *The Quarterly Journal of Economics*, 05 2006, *121* (2), 699–746.

Cheng, Haw and Alice Hsiaw, “Distrust in experts and the origins of disagreement,” *Journal of economic theory*, 2022, *200*, 105401.

Corasaniti, Nick and Trip Gabriel, “Trump Tells Supporters His Criminal Indictments Are About ‘You’,” The New York Times, `https://www.nytimes.com/2023/08/08/us/politics/trump-indictments-2024-campaign.html`, 2023.

Douglas, Karen M, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi, “Understanding conspiracy theories,” *Political Psychology*, 2019, *40*, 3–35.

Egorov, Georgy and Konstantin Sonin, “The Political Economics of Non-democracy,” *Journal of Economic Literature*, June 2024, *62* (2), 594–636.

Eliaz, Kfir and Ran Spiegler, “A Model of Competing Narratives,” *American Economic Review*, December 2020, *110* (12), 3786–3816.

——— , **Simone Galperti, and Ran Spiegler**, “False Narratives and Political Mobilization,” 2022.

Fajgelbaum, Pablo D, Pinelopi K Goldberg, Patrick J Kennedy, and Amit K Khandelwal, “The Return to Protectionism*,” *The Quarterly Journal of Economics*, 11 2019, *135* (1), 1–55.

Franklin, Charles, “Nixon, Watergate and Partisan Opinion,” `https://medium.com/@PollsAndVotes/nixon-watergate-and-partisan-opinion-524c4314d530`, 2018.

Friedman, Lisa, Brad Plumer, and Harry Stevens, "Trump Is Freezing Money for Clean Energy. Red States Have the Most to Lose.," The New York Times, https://www.nytimes.com/2025/02/10/climate/trump-clean-energy-republican-states.html, 2025.

Funke, Manuel, Moritz Schularick, and Christoph Trebesch, "Populist leaders and the economy," *American Economic Review*, 2023, *113* (12), 3249–3288.

Galperti, Simone, "Persuasion: The Art of Changing Worldviews," *American Economic Review*, March 2019, *109* (3), 996–1031.

Glaeser, Edward L., "The Political Economy of Hatred," *The Quarterly Journal of Economics*, 02 2005, *120* (1), 45–86.

Grossman, Gene M and Elhanan Helpman, "Identity politics and trade policy," *The Review of Economic Studies*, 2021, *88* (3), 1101–1126.

Guiso, Luigi, Herrera Helios, Massimo Morelli, and Tommaso Sonno, "Economic insecurity and the demand of populism in Europe," *Economica*, 2023.

Guriev, Sergei and Daniel Treisman, "A theory of informational autocracy," *Journal of Public Economics*, 2020, *186*, 104158.

___ **and** ___ , *Spin Dictators: The Changing Face of Tyranny in the 21st Century*, Princeton University Press, 2022.

___ **and Elias Papaioannou**, "The Political Economy of Populism," *Journal of Economic Literature*, 2022.

Horovitz, David, "Victim of a left-wing coup? Why Netanyahu's conspiracy theory is foul and absurd," The Times of Israel, https://www.timesofisrael.com/victim-of-a-left-wing-coup-why-netanyahus-conspiracy-theory-is-foul-and-absurd/, 2020.

Hotez, P.J., *The Deadly Rise of Anti-science: A Scientist's Warning*, Johns Hopkins University Press, 2023.

hvg.hu, "A magyarok csaknem fele nem is hisz a Soros-tervben," hvg.hu, https://hvg.hu/itthon/20171020_A_magyarok_kozel_fele_nem_is_hisz_a_Sorostervben, 2017.

Kamenica, Emir and Matthew Gentzkow, "Bayesian Persuasion," *American Economic Review*, October 2011, *101* (6), 2590–2615.

Koçak, Korhan, "Sequential updating: A behavioral model of belief change," in "Tech. Rep., Technical report, Working Paper" 2018.

Kocsis, Eva, "Orban Viktor a Kossuth Radio '180 perc' cimu musoraban," [Radio broadcast transcript] Website of the Hungarian Government, https://2015-2019.kormany.hu/hu/a-miniszterelnok/beszedek-publikaciok-interjuk/orban-viktor-a-kossuth-radio-180-perc-cimu-musoraban-20171006, 2017.

Kuziemko, Ilyana, Nicolas Longuet Marx, and Suresh Naidu, "'Compensate the Losers?'Economy-Policy Preferences and Partisan Realignment in the US," 2022.

Le Yaouanq, Yves, "A model of voting with motivated beliefs," *Journal of Economic Behavior & Organization*, 2023, *213*, 394–408.

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

Mays, Jeffrey C., "Mayor's Public Defense Leans on Conspiracy Theories and Race," The New York Times, `https://www.nytimes.com/2024/09/26/nyregion/eric-adams-defense-conspiracy-theories-race.html`, 2024.

McFadden, Alyce and Jeffery C. Mays, "69 Percent of New Yorkers Think Eric Adams Should Resign, Poll Shows," The New York Times, `https://www.nytimes.com/2024/10/04/nyregion/eric-adams-resign-poll.html`, 2024.

Mcmillan, John and Pablo Zoido, "How to Subvert Democracy: Montesinos in Peru," *Journal of Economic Perspectives*, December 2004, *18* (4), 69–92.

Mudde, Cas, "The populist zeitgeist," *Government and opposition*, 2004, *39* (4), 541–563.

Navot, Doron, "Corruption in Israel," in "The Palgrave International Handbook of Israel," Springer, 2022, pp. 1–14.

Norris, Pippa and Ronald Inglehart, *Cultural backlash: Trump, Brexit, and authoritarian populism*, Cambridge University Press, 2019.

Nyhan, Brendan, "Facts and myths about misperceptions," *Journal of Economic Perspectives*, 2020, *34* (3), 220–236.

——, "Why the backfire effect does not explain the durability of political misperceptions," *Proceedings of the National Academy of Sciences*, 2021, *118* (15), e1912440117.

Schwartzstein, Joshua and Adi Sunderam, "Using Models to Persuade," *American Economic Review*, January 2021, *111* (1), 276–323.

Shabecoff, Philipe, "A Secondary Defense of Nixon," The New York Times, `https://www.nytimes.com/1974/07/16/archives/a-secondary-defense-of-nixon-ziegler-presents-thesis.html`, 1974.

SSRS, "CNN Poll: July 1-31, 2023," `https://www.documentcloud.org/documents/23895856-cnn-poll-on-biden-economy-and-elections`, 2023.

Sunstein, Cass R and Adrian Vermeule, "Conspiracy theories: Causes and cures," *Journal of political philosophy*, 2009, *17* (2), 202–227.

Swan, Jonathan, Ruth Igielnik, Shane Goldmacher, and Maggie Haberman, "How Trump Benefits From an Indictment Effect," The New York Times, `https://www.nytimes.com/2023/08/13/us/politics/trump-indictment-effect.html`, 2023.

Szeidl, Adam and Ferenc Szucs, "Media Capture Through Favor Exchange," *Econometrica*, 2021, *89* (1), 281–310.

Uscinski, Joseph E., Karen Douglas, and Stephan Lewandowsky, "Climate Change Conspiracy Theories," 09 2017.

Wallace, Jacob, Paul Goldsmith-Pinkham, and Jason L Schwartz, "Excess death rates for Republicans and Democrats during the COVID-19 pandemic," Technical Report, National Bureau of Economic Research 2022.

Yanagizawa-Drott, David, " Propaganda and Conflict: Evidence from the Rwandan Genocide," *The Quarterly Journal of Economics*, 11 2014, *129* (4), 1947–1994.

YouGov, "CBS News Pol," `https://docs.cdn.yougov.com/3aamn30mjr/cbsnews_20230611_1.pdf`, 2023.

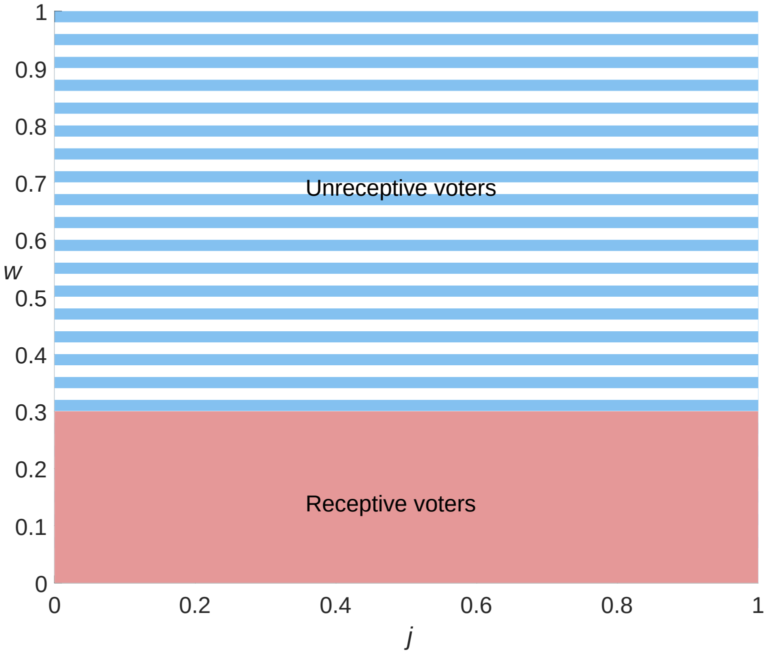
Zhang, Dong, "Draining the Swamp? Populist leadership and corruption," *Governance*, 2024, *37* (4), 1141–1161.

Appendix for Online Publication

A Definitions and proofs

A.1 Formal model of audiences

Figure 3: Media 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 3, the first α 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

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

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; λ is an additional preference component of the voter about the incumbent, which reflects ideological alignment; and ϵ and η_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 θ_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 ϵ, 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 η_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

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

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 θ_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 θ_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 θ_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 $\sigma_{rec}(a^2_{rec}|\theta_{rec}, \hat{a}^1)$. 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 $\sigma_{un}(a^2_{un}|\hat{a}^1)$. We let $\hat{\sigma}$ denote perturbed strategies that incorporate Nature's trembles. We denote the prior belief of actor k of type θ_k by $\mu^0_k(\theta|\theta_k)$, and the posterior belief after history $\hat{h}^t$ by $\mu^t_k(\theta|\theta_k, \hat{h}^t)$. 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:

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

where $\mu^0_{rec}(\theta_{-rec}|\theta_{rec})$ is the prior of the receptive voter of type θ_{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 θ. 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 σ'_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 χ below δ there exists an equilibrium in which all mixing probabilities are within ε 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 α. 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'. *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 θ_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 π.

Consider next the share α 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 β in the denominators, since β 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 π large (holding fixed β) 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 α value

We turn to characterize the condition on α 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 π approaching 1, the term multiplying $1-\alpha$ on the left hand side approaches 1, while the term multiplying α on the right hand side approaches $\hat{q}_c<1$, implying that for π large enough, $\bar{\alpha}(\pi)>0.5$. As a result, when π 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 π 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 π 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 π 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 α who observe propaganda. For these voters, if the elite sends a bad message (with probability π), 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 π 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 π 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 π, in the AR it sends a bad message with probability 1. However, it remains true that in the limit as π 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.[22] 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

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

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

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

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

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 π goes to one r goes to zero.

With this result in hand, we can establish the optimality of the politician's proposed behavior for π large. Consider the limit, as π approaches one, of the mean voter belief after a bad report and propaganda. Given that r goes to zero uniformly in α, 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 π 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 π goes to one), while with propaganda he pays a non-negligible cost f. We conclude that the mixed equilibrium exists for π 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 χ, 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 22 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 π, 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 π sufficiently high, the good R politician never sends propaganda. This is because with a high π 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 π 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 α of voters are receptive to propaganda.

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

- *Partially naive electorate.* A share α_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 6. *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 6.

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 π 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 π 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, α_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 π), 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 π goes to one, converges to $\alpha\hat{q}_c$, so under Assumption 2, for π 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 π goes to one, to $\alpha\hat{q}_c$. Thus, Assumption 2 implies that for π 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 π sufficiently high, the R elite is truthful and the good R politician never sends propaganda. It also follows that for π 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 α, 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 α 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

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

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 (θ_r,θ_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 $1-q_r q_c$. 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 π approaching one the expected posterior will equal the above expression. This is because for π 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 π, 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 π 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.[23] Given this, the AR elite, who prefers to keep the good AR politician, will send a good message after a good signal. Thus, the AR politician will get a payoff near his first best for π 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.

[23] 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 π large

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

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 π large enough the difference is increasing in q_c. For a more direct argument, note that for π 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 π 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 ρ as a function of $\hat{q}_{ar}$, which is valid since q_{ar} is a strictly monotonic transformation of $\hat{q}_{ar}$; (ii) as π 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 π goes to one, all derivatives of ρ 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 π approaches one ρ converges to one and ρ' 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 π 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 ρ 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 π 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 π large there is a unique q_{ar}. Moreover, the arguments also imply that as π 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 3. Assumption 2 holds with the unique solution q^*_{ar} of $P(0,1)cq_c = C'(\hat{q}_{ar})$.

Proposition 4. *Suppose that Assumptions 1 and 3 hold and $\alpha < 0.5$. For π 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 λ and in the voter's prior probability of a good politician q_c.*

Proof of Proposition 4. Consider the proposed equilibrium profile. In that profile, for π 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 π 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 π 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 π large enough $\hat{q}_{ar}$ is increasing in λ. And then q_{ar} is also increasing in λ because q_{ar} is an increasing transformation of $\hat{q}_{ar}$.

The intuition for the result is that λ 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 4. 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-π equilibrium. Since the R elite's reputation costs are prohibitively large, the R elite is truthful in any profile. Given this, for π 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 $\bar{\mu}$ 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 π 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 π large enough elite member j—who cares about reducing $\bar{\mu}$ but has a cost χ_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 π approaches 1), when reporting bad rather than good after a good signal, gains $1-\alpha$ from unreceptive voters but loses at most α 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 π 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 π 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 π 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 π 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 π, if the AR2 politician prefers to send AR2 propaganda, then so does the AR1 politician, a contradiction. It follows that for π 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 χ_h is prohibitively large implies that the R elite is truthful. Therefore, for π large the good R politician does not send propaganda. For π large the AR1 elite is also truthful. This is because, in the limit as π 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 χ_f. However, for π 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 α 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 π 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 π large, both when $e = 0$ and when $e = 1$, the elite's signal is almost perfectly informative. As a result, the large-π 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-π 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 π approaches one, approaches $1-\alpha$, because the $1-\alpha$ share of unreceptive types believe (for π large) that the elite's message is almost perfectly informative. The loss from sending a bad rather than a good message is at most α because in the worst case the share α of receptive voters react in the exact opposite way to her message. Since $\alpha < 0.5$, for π large enough the AR elite always sends a bad message.

Now consider the good R politician. For π 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 π 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 μ' 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 β) expects, for π 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 π'. 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 π', precisely in the β 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, πq_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 π and π' approach one, the condition collapses to $\alpha' > 1/(1+\hat{q}_c)$. Thus, for any such α', we can find π large enough that the result holds.

Equilibrium selection. We show that for π 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 π 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.

A.10 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 with donations data from the Federal Elections Commission. We estimate

$$y_i = \text{const} + \beta \Delta DVM_i + \gamma DVM_i^{old} + \varepsilon_i, \tag{A24}$$

where y_i measures donations received by candidate i in the quarter of the 2022 midterm elections; DVM_i^{old} is the predicted Democratic vote margin of candidate i in their electoral district in the

	Trump donors	Trump donors	Other donors
	Share	Amount (1000 dollars)	
Δ 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)
Constant	0.109***	49.7***	346.4***
	(0.017)	(14.1)	(38.2)
Observations	266	296	296

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

period 2011-2020; and $\Delta DVM_i = DVM_i^{new} - DVM_i^{old}$ is the change in predicted Democratic vote margin between the new and the old district.

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.