The Dictator's Dilemma: A Theory of Propaganda and Repression

A. Arda Gitmez* Konstantin Sonin†

June 2022

Abstract

Repression and information manipulation are two main tools of any authoritarian regime. Most theories, classic and recent, consider them as substitutes. Our theoretical model demonstrates how repression and propaganda complement each other: when the harshest regimes' critics are repressed, the effect of persuasion on the rest of the society is stronger and propaganda is used more heavily. We also demonstrate that the dictator cannot do better than resorting to public messaging even if he has an opportunity to tailor his message to citizens with different attitudes towards the regime.

Keywords: Authoritarian regimes, propaganda, Bayesian persuasion.

JEL Classification: D85, L82.

*Bilkent University. E-mail: arda.gitmez@bilkent.edu.tr.
†University of Chicago. E-mail: ksonin@uchicago.edu

Introduction

At all times, repression and propaganda have been considered the primary tools of autocratic control (Svolik, 2012). In modern times, propaganda took a central place in studies of totalitarian dictatorships such as Hitler's Germany, Stalin's Russia, and Mao's China, in which the state tried to control all aspects of subjects' lives (Arendt, 1951; Friedrich and Brzezinski, 1956; Cassinelli, 1960). With the demise of totalitarian dictatorships, propaganda is no longer considered as means of ideological indoctrination, but rather as a leader's tool of maintaining reputation as a strong and competent hand (Guriev and Treisman, 2022). Yet repression is still a critical tool for a dictator. In a year that followed the summer 2020 protests, Belarus' Alexander Lukashenko had more than 30,000 people arrested and hundreds given long jail terms, a more than ten-fold increase over the average number of political prisoners during the previous decade.

The tradition to consider repression and propaganda as substitutes rather than complements goes back centuries. In *The Prince*, Niccolo Machiavelli writes on whether it is better to be feared than loved: "The answer is that one would like to be both the one and the other; but because it is difficult to combine them, it is far safer to be feared than loved if you cannot be both" (Machiavelli, 2019). In the early formal theory of nondemocratic government (Wintrobe, 1990, 1998) focused on a simple trade-off: the dictator was deciding how to optimally allocate resources between "repression" and "benefits" to population aimed to make the dictator to be more popular. Recently, Guriev and Treisman (2019) argued that subtle propaganda is a substitute for brutal repression.

In this paper, we demonstrate that propaganda is a natural complement to repression. The basic logic of our argument is as follows: when the repression is targeted towards the harshest critics of the regime, it changes the distribution of attitudes towards the dictator within the society. With the harshest critics purged out of the society, the remaining citizens tend to have more favorable attitudes towards the government. Without repression, they would have been "underpersuaded" under the optimal strategy of the dictator. With repression, propaganda towards these citizens are more effective and therefore, optimally, more intense. In other words, once the most disloyal elements of the society are taken out,

the rest can be manipulated more. Thus, the dictator does not face the choice of repression versus propaganda, but rather looks for an optimal bundle of the two.

The repressions in our model are targeted, but they do not need to be precise: realistically, the government might not be able to make a perfect distinction between supporters and skeptics.[1] Yet, the leader will resort to imperfectly targeted repression as long as the cost of repression is relatively small. In Section 4 we consider the possibility of targeted propaganda, in which messaging can be tailored to agents' attitudes, and show that the government cannot do better than with public messaging.

To model information manipulation, we use the basic model of Bayesian persuasion (Kamenica and Gentzkow, 2011; Gehlbach and Sonin, 2014) towards an audience with heterogeneous priors (Alonso and Câmara, 2016; Gitmez and Molavi, 2022). Compared to other communication protocols, the model of Bayesian persuasion assumes fuller commitment on behalf of the sender. This makes sense in an applied model: dictators do not edit news in the real time. Instead, they pass laws, establish institutions of censorship, and appoint editors to control the flow of information. The choice of the institutional bias or the editor of known views corresponds to the choice of the main parameter in the model.

Yet there are theoretical advantages of using the Bayesian persuasion model as well. First and most importantly, the model allows one to study the *maximum propaganda:* it provides the upper limit on the amount of persuasion that can be done via any information exchange between a sender and a receiver. At the same time, our qualitative results easily translate to other information-exchange models such as such as cheap talk in Crawford and Sobel (1982), verifiable messaging in Milgrom (1981),[2] and signaling in Spence (1973). Though the machinery of the respective models would be different, the main intuition is the same.

An important part of our model is that the government organizes information manipulation as a public communication: it establishes an institution, which learns the true state of the world, and then makes a public report. A natural question is whether the government

[1] Arendt (1951) makes a distinction between the *dictatorial* terror, aimed against the well-identified opponents of the regime, from an all-pervasive *totalitarian* terror of purges, mass executions, and concentration camps, which harms many people who are loyal to the regime. Modern theories of repressions with strategic targeting and selection include Myerson (2015), Tyson (2018), and Dragu and Przeworski (2019).

[2] See Titova (2022) for a an argument on how Bayesian persuasion can be embedded in a model of verifiable disclosure with a rich state space.

could do better if it were possible to target different groups with different messages. Is it possible to do more persuasion if persuasion based on private characteristics were possible? In Section 4, we demonstrate that the possibility of private persuasion does not add to the government's persuasion power. Substantively, this explains why many authoritarian regimes use blank, one-size-fits-all messaging rather than target groups with different attitudes individually. As a technical matter, Theorem 2 justifies our assumption that the government sticks to the public persuasion mechanism. This result mirrors the main result in Kolotilin et al. (2017); the difference is that while in Kolotilin et al. (2017) the receivers have heterogeneous preferences, in our model they have heterogeneous priors.

In our abstract model of repression, we do not specify what happens to those who are repressed. At a cost for the dictator, they are no longer a threat.[3] This might be physical elimination as in Esteban, Morelli and Rohner (2015), but might be other forms of political disenfranchisement as well. In addition to mass executions, Stalin relocated hundreds of thousands from the places where they were a political threat to distant regions of Russia. In most cases, Stalin's mass repression campaigns were organized about broad ethnic or social categories (Gregory, Schröder and Sonin, 2011); in our model, this corresponds to the dictator repressing people basing on imperfect information about their attitudes. In the realm of democratic politics, Glaeser and Shleifer (2005) show that the incumbent politician might deliberately choose policies that drive voters who oppose him out of the district. Our theory applies to such situations as well.

The rest of the paper is organized as follows. Section 2 studies the case when the leader cannot repress, only persuade, the citizens. Section 3 analyzes the main case, when the leader optimally combines repression and persuasion. Finally, Section 4 deals with private persuasion.

[3] Montagnes and Wolton (2019) and Rozenas (2020) use communist purges in Stalin's Russia and Mao's China to demonstrate the effect of repression on *behavior* of dictator's subjects. In our model, there is no such effect: the repressions change the distribution of attitudes towards the leader by eliminating certain members of the population.

2 Propaganda without Repression

We begin with a discussion of our model with information manipulation only, and we introduce the possibility of repression in Section 3.

2.1 Setup

There is a sender s (the *leader*) and a continuum of receivers $I = [0, 1]$ (the *citizens*). A state of the world is denoted by $\omega \in \{0, 1\}$. Here, $\omega = 1$ is the state of the world where the citizens' preferences align with the leader (e.g., the state where the leader is competent), and $\omega = 0$ is the state where there is a misalignment.

Share $\alpha \in (0, 1)$ of citizens have prior $\mu_L = \Pr\{\omega = 1\}$, and share $1 - \alpha$ has prior $\mu_H = \Pr\{\omega = 1\}$ with

$$\mu_L < \mu_H < \frac{1}{2}.$$

We call those with prior μ_L *skeptics*. The leader has prior μ_H,[4] and knows the distribution of priors in the society.

Given her information about ω, each citizen $i \in I$ takes an action $a_i \in [0, 1]$. This can be interpreted as the level of *support* citizen i provides to the leader. Receiver $i \in I$'s payoff as a function of the action taken and the state is

$$u_i(a_i, \omega) = -|a_i - \omega| \tag{1}$$

Therefore, the optimal choice is $a_i = 1$ only if the posterior assigned to state $\omega = 1$ by receiver i exceeds $\frac{1}{2}$. Otherwise, $a_i = 0$. Since $\mu_H < \frac{1}{2}$, absent any information, all citizens choose $a_i = 0$.

The leader's payoff is the total support he receives from the citizens

$$u_s\left(\{a_i\}_{i \in I}\right) = \int_{i \in I} a_i \cdot di$$

[4]The main insights of the model will not change with a different value of the leader's prior.

Propaganda. Citizens do not have any information about ω beyond the information conveyed by the leader. The leader uses a public persuasion mechanism that sends messages from M. That is, the leader commits to an information structure $\{\sigma(\cdot|\omega)\}_{\omega\in\{0,1\}}$ where

$$\sigma(\cdot|\omega) \in \Delta(M) \qquad \text{for all } \omega \in \{0,1\},$$

and the message drawn, $m \in M$, is publicly observable to each citizen.

In an environment where each media source is accessible to citizens, restriction to public persuasion is without loss of generality. Indeed, consider a setup where there are multiple information sources $1,\ldots,n$ with message spaces $M_1,\ldots,M_n$. Let source $j \in \{1,\ldots,n\}$ use information structure $\{\sigma_j(\cdot|\omega)\}_{\omega\in\{0,1\}} \in \Delta(M_i)$. As long as the citizens can observe messages from various sources, one can define

$$M \equiv M_1 \times \ldots \times M_n$$

and, for each $m = (m_1,\ldots,m_n) \in M$, let

$$\sigma(m|\omega) = \sigma_1(m_1|\omega) \cdot \ldots \cdot \sigma_n(m_n|\omega) \qquad \text{for all } \omega \in \{0,1\},$$

so that the same outcome can be implemented via a public persuasion mechanism. We assume that $|M|$ is large enough, so that there is a sufficient number of action recommendations for each receiver. As we will show later, in this model the leader uses at most two messages.

The assumption that each citizen can access many information sources is a reasonable starting point for a media application. However, we will go further, and demonstrate that the leader cannot do better even if there were a possibility of private persuasion. In Section 4, we present an alternative setup where each citizen can access to at most one media source, and she picks the media source that gives the highest (subjective) expected payoff. This setup imposes a natural *incentive compatibility* constraint. We will see that the leader cannot do better with private persuasion than with the public one: the payoff from an incentive compatible private persuasion mechanism can always be achieved via a public persuasion mechanism.

2.2 Analysis

We begin by drawing the **value function** of the leader as a function of his posterior belief $\mu = \Pr_s(\omega = 1|m)$, and then use the concavification approach of Kamenica and Gentzkow (2011).

Suppose the leader's posterior is $\mu \in [0, 1]$. Since share $1 - \alpha$ of citizens, the non-skeptics, start with the same prior as the leader, they end up with the same posterior μ. By Proposition 1 of Alonso and Câmara (2016), the skeptics have the following posterior:

$$\mu' = \frac{\mu \frac{\mu_L}{\mu_H}}{\mu \frac{\mu_L}{\mu_H} + (1-\mu)\frac{1-\mu_L}{1-\mu_H}} \tag{2}$$

Note that μ' is monotonic in μ: if some public information makes the sender more optimistic, it makes every receiver more optimistic as well. The necessary condition for skeptics to support the leader is $\mu' \geq \frac{1}{2}$.[5] If this is true, then, by (2), the leader's posterior must be at least

$$\bar{\mu} \equiv \frac{1}{1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}} > \frac{1}{2}$$

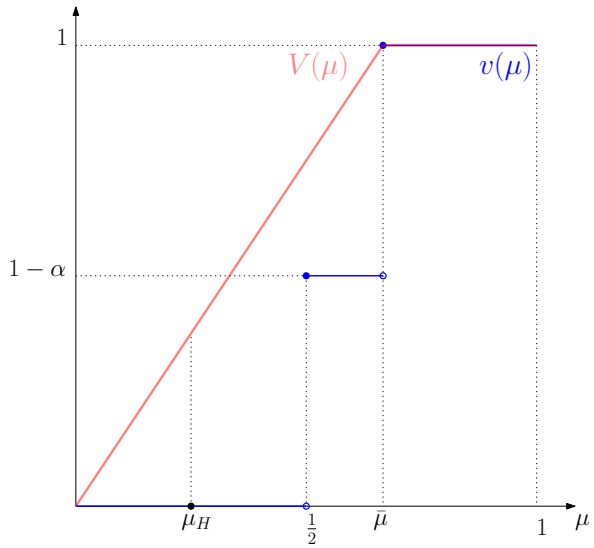
The intuitive reason why the inequality is strict is as follows. The skeptics start with pessimistic beliefs, and thus convincing them to support the leader requires strong evidence in favor of $\omega = 1$. With such strong evidence, the other citizens would be more than sufficiently convinced. That is, *when the skeptics are marginal, the non-skeptics are inframarginal.*

Using this observation, the value function of the sender (as a function of sender's posterior μ) is

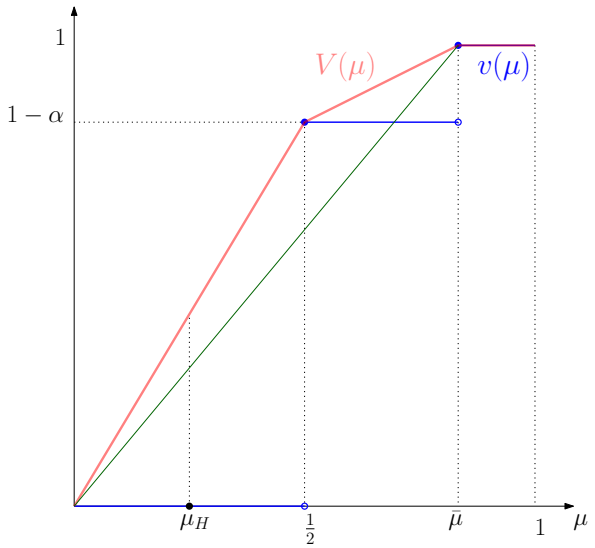
$$v(\mu) = \begin{cases} 0, & \text{if } \mu \in \left[0, \frac{1}{2}\right), \\ 1 - \alpha, & \text{if } \mu \in \left[\frac{1}{2}, \bar{\mu}\right), \\ 1, & \text{if } \mu \in [\bar{\mu}, 1]. \end{cases}$$

[5] As usual in the Bayesian persuasion literature, we are assuming that the receivers take sender's favorite action when indifferent. This is justified by observing that the sender designs the information structure, and can arbitrarily approximate this decision rule.

Figure 1: The leader's Value Function and its Concavification.



(a) Numerous skeptics $\left(\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \geq 1-2\alpha\right)$;



(b) Few skeptics $\left(\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \leq 1-2\alpha\right)$.

The optimal solution relies on the characterization of the concave closure of $v(\mu)$, $V(\mu)$. Figures 1a and 1b illustrate the value functions and their concave closures for two possible cases. Intuitively, Figure 1a corresponds to the "high α" case, where there are numerous skeptics. On the contrary, Figure 1b illusrates the "low α" case where there are few skeptics.

Visual examination of Figures 1a and 1b demonstrates that the optimal information policy invokes two posteriors. This can be achieved by using two messages: $m \in \{0,1\}$. Moreover, one of the posteriors in the support is $\mu = 0$, i.e., one of the messages perfectly reveals $\omega = 0$. This can be achieved by setting $\sigma(m=1|\omega=1) = 1$ in the optimal policy. Therefore, the optimal policy is characterized by a single-dimensional object:

$$\beta = \sigma(m=1|\omega=0) \in [0,1].$$

The most natural interpretation of β is the level of *propaganda*: it is the likelihood that the leader will send the message "things are good" when the leader is, in fact, incompetent. Our analysis so far can be summarized by the following proposition.

Proposition 1. *The propaganda level chosen by the leader is*

$$\beta^*(\alpha) = \begin{cases} \frac{\mu_H}{1-\mu_H}, & \text{if } \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \leq 1-2\alpha, \\ \frac{\mu_L}{1-\mu_L}, & \text{if } \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \geq 1-2\alpha, \end{cases}$$

and the leader's subjective payoff from the optimal policy is

$$V^*(\alpha) = \max\left\{\mu_H \cdot \left(1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}\right), \mu_H \cdot 2(1-\alpha)\right\}$$

The first part of Proposition 1 implies that the level of propaganda decreases with α, i.e., the share of skeptics in the population. Intuitively, with a small enough share of skeptics, the leader ignores them and cater to the non-skeptics. Since non-skeptics are more malleable towards being convinced, this results in a higher level of propaganda. It is also worth noting that $V^*(\alpha)$ is non-increasing in α: the leader is worse off when the share of skeptics in the society is higher.

3 Propaganda with Repression

We now include the possibility of repression in our model. The repression takes place before the propaganda and involves purging a particular group of citizens. The leader does not observe the citizens' priors perfectly; instead, he observes an informative signal of their priors. As in Gregory, Schröder and Sonin (2011), there is a label $\ell \in \{L, H\}$ associated with each citizen. Here, $\ell = L$ stands for a label of "skeptic" (an agent with a low prior), and $\ell = H$ stands for a label of "non-skeptic" (an agent with a high prior). The association between the prior and the label is

$$\Pr\{\ell = L|\mu_L\} = \Pr\{\ell = H|\mu_H\} = \rho,$$

where $\rho \in \left[\frac{1}{2}, 1\right]$ measures the quality of information available to the leader.[6] When $\rho = \frac{1}{2}$, the leader does not have access to any information about the prior of a citizen; when $\rho = 1$,

[6] Clearly, when $\rho < \frac{1}{2}$, the labels can be swapped and the same analysis applies.

the leader observes each citizen's prior perfectly.

The measure of citizens with label $\ell = L$ is

$$\bar{\lambda}_L \equiv \rho \cdot \alpha + (1 - \rho) \cdot (1 - \alpha).$$

Similarly, a measure of $\bar{\lambda}_H \equiv \rho \cdot (1 - \alpha) + (1 - \rho) \cdot \alpha$ citizens have label $\ell = H$. Let λ_L be the set of citizens with label $\ell = L$ who are repressed, and let λ_H be the set of citizens with label $\ell = H$ who are repressed. The leader chooses the measure of citizens to repress:[7]

$$(\lambda_L, \lambda_H) \in [0, \bar{\lambda}_L] \times [0, \bar{\lambda}_H]$$

and pays the cost $c \cdot (\lambda_L + \lambda_H)$, where $c > 0$.

The repressed citizens are eliminated (*purged*) from the society, so that when the leader chooses (λ_L, λ_H), the total measure of citizens is $1 - \lambda_L - \lambda_H$. The leader's payoff is the fraction of receivers who take $a_i = 1$ among those who were not repressed. That is, if the leader chooses (λ_L, λ_H), letting $I' \subset I$ denote the set of citizens who were not repressed, the leader's payoff is

$$u_S(\{a_i\}_{i \in I'}, \lambda_L, \lambda_H) = \frac{\int_{i \in I'} a_i di}{1 - \lambda_L - \lambda_H} - c \cdot (\lambda_L + \lambda_H).$$

The assumption that the leader maximizes the share of support is standard in the political economy literature; in many models, the remaining share determines, e.g., the probability of a revolution that removes the leader from power.

[7] Because the leader does not observe any information beyond the label, she does not distinguish among citizens with the same label. Consequently, we assume that the leader targets an arbitrarily chosen fraction of the receivers with the same label.

3.1 Analysis

After the leader chooses (λ_L, λ_H) to repress, the share of skeptics remaining in the society is

$$\alpha(\lambda_L, \lambda_H) \equiv \frac{1}{1-\lambda_L-\lambda_H}\left(\alpha - \frac{\rho\cdot\alpha}{\rho\cdot\alpha+(1-\rho)\cdot(1-\alpha)}\lambda_L - \frac{(1-\rho)\cdot\alpha}{\rho\cdot(1-\alpha)+(1-\rho)\cdot\alpha}\lambda_H\right) \tag{3}$$

Following Proposition 1, the propaganda level to accompany (λ_L, λ_H) is β^* characterized in Proposition 1 with $\alpha(\lambda_L, \lambda_H)$. The leader's subjective payoff from repressing (λ_L, λ_H), therefore, is

$$u(\lambda_L, \lambda_H) \equiv V^*\left(\alpha(\lambda_L, \lambda_H)\right) - c\cdot(\lambda_L+\lambda_H)$$

In the following part, we characterize the level of repression $(\lambda_L^*, \lambda_H^*)$ that maximizes $u(\lambda_L, \lambda_H)$.

Purging Non-Skeptics. We begin with a simple observation: a leader never represses citizens with label $\ell = H$.

Proposition 2. *The chosen level of repression satisfies* $\lambda_H^* = 0$.

This is an intuitive result: the citizens with $\ell = H$ are *positively selected group*, with a share of skeptics lower than α. Therefore, eliminating these people would lead to a larger α. By Proposition 1, a leader is worse off under a larger α.

With Proposition 2 in hand, the leader's choice reduces to a single-dimensional optimization problem with choice variable $\lambda_L \in [0, \bar{\lambda}_L]$.

Purging Skeptics. Our next result characterizes the leader's chosen level of repression towards citizens labeled with $\ell = L$.

Proposition 3. *Let:*

$$c^*(\rho,\alpha) \equiv \frac{\mu_H}{\rho\alpha+(1-\rho)(1-\alpha)}\left(\frac{2\rho(1-\alpha)}{1-\rho\alpha-(1-\rho)(1-\alpha)} - \max\left\{2(1-\alpha), 1+\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}\right\}\right)$$

The chosen level of repression satisfies:

$$\lambda_L^* = \begin{cases} \bar{\lambda}_L, & \text{if } c \leq c^*(\rho, \alpha) \\ 0, & \text{if } c > c^*(\rho, \alpha) \end{cases}$$

That is, the leader represses all citizens with $\ell = L$ if the cost of repression is low enough, and does not repress anyone otherwise.

The proof of Proposition 3 goes through showing that the objective function is convex in λ_L, and thus the chosen level must be a corner solution. The remainder of the proof is a straightforward exercise of checking the value of objective function at the boundaries.

A couple of notes about Proposition 3 is in order.

- Consider the case where $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \leq 1 - 2\alpha$. In this case,

$$c^*(\rho, \alpha) = \frac{\mu_H}{\rho\alpha + (1-\rho)(1-\alpha)} \frac{2(1-\alpha)\alpha(2\rho - 1)}{1 - \rho\alpha - (1-\rho)(1-\alpha)} \geq 0$$

  Thus, there is a range of c where the leader indeed chooses repression. When $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \geq 1 - 2\alpha$, the parameters may be such that $c^*(\rho, \alpha) \leq 0$.

- When $\rho = \frac{1}{2}$,

$$c^*(\frac{1}{2}, \alpha) = 2\mu_H \left(2(1-\alpha) - \max\left\{ 2(1-\alpha), 1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \right\} \right) \leq 0$$

  Naturally, when the leader does not have any information about the citizens' beliefs, repression is extremely ineffective and is never used. In contrast, when $\rho = 1$,

$$c^*(1, \alpha) = \frac{\mu_H}{\alpha} \left(2 - \max\left\{ 2(1-\alpha), 1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \right\} \right) > 0$$

Now, we are ready to investigate the joint repression and propaganda levels by the leader. When the cost of repression is c, the leader chooses a propaganda level $\beta^*(c)$ associated with the repression level $\lambda^*(c)$. The following Theorem, our main result, summarizes our analysis.

Theorem 1. *If* $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \leq 1-2\alpha$*, the propaganda and repression levels are given by:*

$$\beta^*(c) = \frac{\mu_H}{1-\mu_H} \qquad \lambda^*(c) = \begin{cases} \bar{\lambda}_L, & \text{if } c \leq c^*(\rho,\alpha) \\ 0, & \text{if } c > c^*(\rho,\alpha) \end{cases}$$

If $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \geq 1-2\alpha$*, propaganda and repression levels are given by:*

$$\beta^*(c) = \begin{cases} \frac{\mu_H}{1-\mu_H}, & \text{if } c \leq c^*(\rho,\alpha) \\ \frac{\mu_L}{1-\mu_L}, & \text{if } c > c^*(\rho,\alpha) \end{cases} \qquad \lambda^*(c) = \begin{cases} \bar{\lambda}_L, & \text{if } c \leq c^*(\rho,\alpha) \\ 0, & \text{if } c > c^*(\rho,\alpha) \end{cases}$$

Proof. The calculation of $\lambda^*(c)$ follows from Proposition 3 and the calculation of $\beta^*(c)$ follows from Proposition 1. □

An implication of Theorem 1 is regarding the $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \geq 1-2\alpha$ case. This is the case where skeptics have a sufficiently large share in the population so that, in the absence of repression, the leader caters towards them. Now, suppose ρ is high enough so that $c^*(\rho,\alpha) > 0$. When the cost of repression decreases to $c \leq c^*(\rho,\alpha)$, the leader represses all the receivers labeled "skeptic", and the propaganda caters to non-skeptics. Therefore, *when the cost of repression is lower, higher repression is accompanied with a higher level of propaganda*. Propaganda and repression are complements.

4 Targeted Propaganda

The crucial step in the complementarity of propaganda and repression relies on repression being *targeted* towards citizens. A natural question to ask is whether the leader also benefits from targeted propaganda. In this section, we argue that the answer is no. In particular, allowing for *private persuasion,* i.e., the opportunity to design type-specific propaganda, does not actually expand the menu of leader's tools. We show that it is never optimal for the leader to create two different information sources that appeal to different groups of citizens. Therefore, under the optimal policy, the leader sticks with public propaganda. This result is closely related to the "impossibility of private persuasion" result in Kolotilin et al. (2017);

the difference is that our result is in a setup with heterogeneous priors, rather than with heterogeneous preferences as in Kolotilin et al. (2017).

We consider a general group of persuasion mechanisms where the leader designs an information structure for each group. These information structures can be interpreted as different media sources targeted towards citizens with different priors. For citizens with priors μ_L, the sender sends messages $m^L \in M^L = \{0,1\}$. For citizens with priors μ_H, the sender sends messages $m^H \in M^H = \{0,1\}$.[8]

Definition 1. *A* **persuasion mechanism** *is a pair of information structures* (σ_0^L, σ_1^L) *and* (σ_0^H, σ_1^H)*, where:*

$$\sigma_\omega^\tau = \Pr(m^\tau = 1|\omega) \in [0,1] \qquad \textit{for } \tau \in \{L,H\}, \omega \in \{0,1\}$$

Throughout, we will let $\sigma = (\sigma_0^L, \sigma_1^L, \sigma_0^H, \sigma_1^H)$ denote a persuasion mechanism. Fix a persuasion mechanism σ. Once a citizen with prior $\mu_t, t \in \{L,H\}$, observes a message $m^\tau \in M^\tau$ from an information structure $(\sigma_0^\tau, \sigma_1^\tau)$, she forms the posterior:

$$\Pr_t\{\omega = 1|m^\tau = 0\} = \frac{\mu_t(1-\sigma_1^\tau)}{\mu_t(1-\sigma_1^\tau) + (1-\mu_t)(1-\sigma_0^\tau)} \tag{4}$$

$$\Pr_t\{\omega = 1|m^\tau = 1\} = \frac{\mu_t\sigma_1^\tau}{\mu_t\sigma_1^\tau + (1-\mu_t)\sigma_0^\tau} \tag{5}$$

By (1), the citizen's action following the message is:

$$a_t(m^\tau) = \begin{cases} 1, & \text{if } \Pr_t\{\omega = 1|m^\tau\} \geq \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Assume that each citizen can observe messages drawn from one, and only one, information structure. A citizen with prior μ_t cannot be forced to follow the information rule designed for her, (σ_0^t, σ_1^t). Instead, she must have the correct incentives to choose her designated information structure. We capture this through an *incentive compatibility* constraint.[9]

[8] The sufficiency of two messages in the support is immediate from a revelation principle argument. Each message in support stands for the respective action recommendation.

[9] One way to justify the "one information source for each citizen" assumption is cognitive constraints.

To formally introduce the incentive compatibility constraint, define the (subjective) payoff of a citizen with prior μ_t from observing messages drawn from $(\sigma_0^\tau, \sigma_1^\tau)$:

$$\begin{aligned} U(\mu_t, \tau) &= \mathbb{E}_{\omega \sim \mu_t, m^\tau \sim (\sigma_0^\tau, \sigma_1^\tau)} \left[-|a_t(m^\tau) - \omega| \right] \\ &= \mu_t \cdot (\sigma_1^\tau \cdot a_t(m^\tau = 1) + (1 - \sigma_1^\tau) \cdot a_t(m^\tau = 0) - 1) \\ &\quad + (1 - \mu_t) \cdot (-\sigma_0^\tau \cdot a_t(m^\tau = 1) - (1 - \sigma_0^\tau) \cdot a_t(m^\tau = 0)) \end{aligned} \tag{7}$$

We are now ready to introduce the notion of incentive compatibility.

Definition 2. *A persuasion mechanism* $\sigma = \left(\sigma_0^L, \sigma_1^L, \sigma_0^H, \sigma_1^H\right)$ *is* **incentive compatible** *if*

$$U(\mu_H, H) \geq U(\mu_H, L), \qquad U(\mu_L, L) \geq U(\mu_L, H)$$

Under an incentive compatible persuasion mechanism, each citizen has incentives to follow the news conveyed by the information source targeted to her. The leader's subjective payoff from an incentive compatible persuasion mechanism $\sigma = \left(\sigma_0^L, \sigma_1^L, \sigma_0^H, \sigma_1^H\right)$ is

$$\begin{aligned} v_\sigma(\mu_s) \equiv\ & \mu_s \cdot \alpha \cdot \left(\sigma_1^L \cdot a_L(m^L = 1) + (1 - \sigma_1^L) \cdot a_L(m^L = 0)\right) \\ & + \mu_s \cdot (1 - \alpha) \cdot \left(\sigma_1^H \cdot a_H(m^H = 1) + (1 - \sigma_1^H) \cdot a_H(m^H = 0)\right) \\ & + (1 - \mu_s) \cdot \alpha \cdot \left(\sigma_0^L \cdot a_L(m^L = 1) + (1 - \sigma_0^L) \cdot a_L(m^L = 0)\right) \\ & + (1 - \mu_s) \cdot (1 - \alpha) \cdot \left(\sigma_0^H \cdot a_H(m^H = 1) + (1 - \sigma_0^H) \cdot a_H(m^H = 0)\right) \end{aligned} \tag{8}$$

We now present the main finding of this section.

Theorem 2. *For any incentive compatible persuasion mechanism* $\sigma = \left(\sigma_0^L, \sigma_1^L, \sigma_0^H, \sigma_1^H\right)$,

Alternatively, if citizens can follow more than one source, any additional information conveyed by an information structure will be taken into account by each citizen. In this case, any persuasion mechanism will be equivalent to a public persuasion mechanism which combines the information conveyed by both information structures. Then, restricting attention to public information structures, as we did in Section 2, is naturally justified. Finally, one can consider a setup where each citizen can "sell" the information obtained from one source to the others. Such a possibility introduces an incentive compatibility constraint of the type we have here.

there exists another persuasion mechanism $\hat{\sigma} = \left(\hat{\sigma}_0^L, \hat{\sigma}_1^L, \hat{\sigma}_0^H, \hat{\sigma}_1^H\right)$*, where*

$$\hat{\sigma}_0^L = \hat{\sigma}_0^H \qquad \hat{\sigma}_1^L = \hat{\sigma}_1^H$$

and

$$v_{\hat{\sigma}}(\mu_s) \geq v_{\sigma}(\mu_s)$$

Theorem 2 implies that the leader can always maximize his payoff by using a **public persuasion mechanism**, where he offers the same information structure for each citizen. Intuitively, this is because the incentive compatibility constraints are extremely binding for the leader, to the extent that a public mechanism (which satisfies incentive compatibility trivially) can achieve the same payoff. Substantively, this provides an explanation why many authoritarian regimes prefer standardized approach to censorship and propaganda.

5 Conclusion

We offer a model of information manipulation and repression, two main tools in any autocrat's arsenal. The possibility of repression enhances the dictator's ability to persuade citizens via information manipulation. The mechanism is as follows. The optimal survival strategy of the dictator calls for repressing those who are most likely to have the least positive attitude towards him. Then, the rest could be manipulated, via propaganda, more than before: targeting persuasion towards skeptics, without repression, would have left the non-skeptics "underpersuaded". The possibility to repress skeptics allows to extract more of the non-skeptics informational rent.

We also demonstrate that there is a structural reason why many dictators resort to uniform, one-size-fits-all messaging in their propaganda. If people could consume information from many media sources than, being rational, they would consume it from all sources – that is, the communication channel is public. Now, if people' media consumption is limited, they will need to self-select into targeted media sources. Indeed, if the dictator is able to differentiate his subject by their attitude, then repressing the most skeptical ones is the best

strategy. Absent such ability to differentiate, private persuasion have to rely on incentive compatible self-selection, which turns out to be impossible.

Our model explains why George Orwell's state is not content with flooding his subjects with propaganda, but has to use repression to make propaganda work. In Oceania, people are forced to use the *newspeak*, a special language designed to limit their ability to articulate anti-government concepts, cannot switch off radio that translates propaganda, and are forced to participate in ideological indoctrination meetings. Yet the overall logic of the novel is that it is torture applied to the skeptics that makes the regime stable.

References

Alonso, Ricardo and Odilon Câmara. 2016. "Bayesian Persuasion with Heterogeneous Priors." *Journal of Economic Theory* 165:672–706.

Arendt, Hannah. 1951. *The Origins of Totalitarianism*. New York: Harcourt and Brace.

Cassinelli, C. W. 1960. "Totalitarianism, Ideology, and Propaganda." *The Journal of Politics* 22(1):68–95.

Crawford, Vincent P. and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50(6):1431–1451.

Dragu, Tiberiu and Adam Przeworski. 2019. "Preventive Repression: Two Types of Moral Hazard." *American Political Science Review* 113(1):77–87.

Esteban, Joan, Massimo Morelli and Dominic Rohner. 2015. "Strategic Mass Killings." *Journal of Political Economy* 123(5):1087–1132.

Friedrich, Carl J. and Zbigniew K. Brzezinski. 1956. *Totalitarian Dictatorship and Autocracy*. Cambridge, MA: Harvard University Press.

Gehlbach, Scott and Konstantin Sonin. 2014. "Government Control of the Media." *Journal of Public Economics* 118:163–171.

Gitmez, A. Arda and Pooya Molavi. 2022. "Polarization and Media Bias." *Working Paper* .

Glaeser, Edward L. and Andrei Shleifer. 2005. "The Curley Effect: The Economics of Shaping the Electorate." *Journal of Law, Economics, and Organization* 21(1):1–19.

Gregory, Paul R., Philipp J. H. Schröder and Konstantin Sonin. 2011. "Rational Dictators and the Killing of Innocents: Data from Stalin's Archives." *Journal of Comparative Economics* 39(1):34–42.

Guriev, Sergei and Daniel Treisman. 2019. "Informational Autocrats." *Journal of Economic Perspectives* 33(4):100–127.

Guriev, Sergei and Daniel Treisman. 2022. *Spin Dictators: The Changing Face of Tyranny in the 21st Century*. Princeton, New Jersey: Princeton University Press.

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

Kolotilin, Anton, Tymofiy Mylovanov, Andriy Zapechelnyuk and Ming Li. 2017. "Persuasion of a Privately Informed Receiver." *Econometrica* 85(6):1949–1964.

Machiavelli, Niccolo. 2019. *Machiavelli: The Prince*. Cambridge Texts in the History of Political Thought 2 ed. Cambridge: Cambridge University Press.

Milgrom, Paul R. 1981. "Good News and Bad News: Representation Theorems and Applications." *The Bell Journal of Economics* pp. 380–391.

Montagnes, Pablo and Stephane Wolton. 2019. "Mass Purges: Top-Down Accountability in Autocracy." *American Political Science Review* 113(4):1045–1059.

Myerson, Roger B. 2015. "Moral Hazard in High Office and the Dynamics of Aristocracy." *Econometrica* 83(6):2083–2126.

Rozenas, Arturas. 2020. "A Theory of Demographically Targeted Repression." *Journal of Conflict Resolution* 64(7-8):1254–1278.

Spence, Michael. 1973. "Job Market Signaling." *Quarterly Journal of Economics* 87(3):355–374.

Svolik, Milan W. 2012. *The Politics of Authoritarian Rule*. New York: Cambridge University Press.

Titova, Maria. 2022. "Persuasion with Verifiable Information." *Working Paper* .

Tyson, Scott A. 2018. "The Agency Problem Underlying Repression." *The Journal of Politics* 80(4):1297–1310.

Wintrobe, Ronald. 1990. "The Tinpot and the Totalitarian: An Economic Theory of Dictatorship." *American Political Science Review* 84(3):849–872.

Wintrobe, Ronald. 1998. *The Political Economy of Dictatorship*. Cambridge: Cambridge University Press.

Appendix

A1 Proofs

Proof of Proposition 1. Given the value function, the optimal policy can be derived through Corollary 2 of Kamenica and Gentzkow (2011). The sender's subjective payoff from optimal policy is $V(\mu_H)$, where $V(\mu)$ is the concave closure of $v(\mu)$. The optimal policy can be derived through visual inspection of Figures 1a and 1b.

- If $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \geq 1-2\alpha$, the optimal information structure induces posteriors $\mu \in \{0, \frac{1}{1+\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}}\}$.
  This is achieved by having two signals in the support: $M = \{0, 1\}$, and:

$$\sigma^*(m=1|\omega=1) = 1$$
$$\sigma^*(m=1|\omega=0) = \frac{\mu_L}{1-\mu_L}$$

  The leader's payoff from optimal policy is

$$V(\mu_H) = \mu_H\left(1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}\right)$$

  Intuitively, the condition $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \geq 1-2\alpha$ corresponds to the case where there are sufficiently many skeptics in the population (relative to their pessimism). In this case, the leader chooses the optimal policy of skeptics. The remaining citizens are inframarginal under the optimal policy.

- If $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \leq 1-2\alpha$, the optimal information structure induces posteriors $\mu \in \{0, \frac{1}{2}\}$.
  This is achieved by having two signals in the support: $M = \{0, 1\}$, and:

$$\sigma^*(m=1|\omega=1) = 1$$
$$\sigma^*(m=1|\omega=0) = \frac{\mu_H}{1-\mu_H}$$

  The leader's payoff from optimal policy is

$$V(\mu_H) = \mu_H \cdot 2 \cdot (1-\alpha)$$

Intuitively, the condition $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L} \leq 1-2\alpha$ corresponds to the case where there are too few skeptics in the population. In this case, the leader ignores the skeptics and adopts the optimal policy for the remaining citizens.

□

Proof of Proposition 2. Suppose, towards a contradiction, that the leader chooses $(\lambda_L^*, \lambda_H^*)$ with $\lambda_H^* > 0$. The leader's payoff is

$$u(\lambda_L^*, \lambda_H^*) = V^*\left(\alpha(\lambda_L^*, \lambda_H^*)\right) - c \cdot (\lambda_L^* + \lambda_H^*)$$

Consider the alternative choice of $(\lambda_L^*, 0)$, coupled with the corresponding propaganda level β_0^* given by Proposition 1 applied to $\alpha(\lambda_L^*, 0)$. This yields the payoff:

$$u(\lambda_L^*, 0) = V^*\left(\alpha(\lambda_L^*, 0)\right) - c \cdot \lambda_L^*$$

We first note that $\alpha(\lambda_L^*, 0) \leq \alpha(\lambda_L^*, \lambda_H^*)$. To see this, let:

$$\overline{\theta} \equiv \frac{\rho \cdot \alpha}{\rho \cdot \alpha + (1-\rho) \cdot (1-\alpha)} \tag{A1}$$

$$\underline{\theta} \equiv \frac{(1-\rho) \cdot \alpha}{\rho \cdot (1-\alpha) + (1-\rho) \cdot \alpha} \tag{A2}$$

Note that, since $\rho \geq \frac{1}{2}$, $\overline{\theta} \geq \alpha \geq \underline{\theta}$. Moreover,

$$\alpha(\lambda_L^*, \lambda_H^*) = \frac{\alpha - \overline{\theta}\lambda_L^* - \underline{\theta}\lambda_H^*}{1 - \lambda_L^* - \lambda_H^*}$$

$$\alpha(\lambda_L^*, 0) = \frac{\alpha - \overline{\theta}\lambda_L^*}{1 - \lambda_L^*}$$

Then, $\alpha(\lambda_L^*, 0) \leq \alpha(\lambda_L^*, \lambda_H^*)$ if and only if:

$$\underline{\theta} \cdot (1 - \lambda_L^*) \leq \alpha - \overline{\theta}\lambda_L^* \iff \left(\overline{\theta} - \underline{\theta}\right)\lambda_L^* \leq \alpha - \underline{\theta}$$

Substituting (A1) and (A2), this is equivalent to:

$$\frac{\alpha(1-\alpha)(2\rho-1)}{(\rho\cdot\alpha+(1-\rho)\cdot(1-\alpha))\,(\rho\cdot(1-\alpha)+(1-\rho)\cdot\alpha)}\lambda_L^* \leq \frac{\alpha(1-\alpha)(2\rho-1)}{\rho\cdot(1-\alpha)+(1-\rho)\cdot\alpha}$$

$$\iff$$

$$\frac{1}{\rho\cdot\alpha+(1-\rho)\cdot(1-\alpha)}\lambda_L^* \leq 1$$

Finally, using $\lambda_L^* \leq \bar{\lambda}_L = \rho\cdot\alpha+(1-\rho)\cdot(1-\alpha)$ yields the desired result, and we conclude that $\alpha(\lambda_L^*,0) \leq \alpha(\lambda_L^*,\lambda_H^*)$. Since V^* is nonincreasing in α, then,

$$V^*\left(\alpha(\lambda_L^*,0)\right) \geq V^*\left(\alpha(\lambda_L^*,\lambda_H^*)\right)$$

This, along with $c>0$ and $\lambda_H^*>0$, implies:

$$V^*\left(\alpha(\lambda_L^*,0)\right) - c\cdot\lambda_L^* \geq V^*\left(\alpha(\lambda_L^*,\lambda_H^*)\right) - c\cdot(\lambda_L^*+\lambda_H^*) \implies u(\lambda_L^*,0) > u(\lambda_L^*,\lambda_H^*)$$

which contradicts the optimality of $(\lambda_L^*,\lambda_H^*)$ with $\lambda_H^*>0$. □

Proof of Proposition 3. Recall that the leader's choice of repressing λ_L citizens with $\ell=L$ results in the subjective payoff of: $u(\lambda_L,0) = V^*\left(\alpha(\lambda_L,0)\right) - c\cdot\lambda_L$. By Proposition 1, then, the leader's payoff is

$$u(\lambda_L,0) = \max\left\{\mu_H\cdot\left(1+\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}\right), \mu_H\cdot 2(1-\alpha(\lambda_L,0))\right\} - c\cdot\lambda_L$$

Substituting Equation (3) and rearranging, the leader's chosen level of repression satisfies:

$$\lambda_L^* = \arg\max_{\lambda\in[0,\bar{\lambda}_L]} \mu_H\cdot\max\left\{1+\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}, 2\cdot\frac{1-\alpha-\frac{(1-\rho)(1-\alpha)}{\rho\alpha+(1-\rho)(1-\alpha)}\lambda}{1-\lambda}\right\} - c\cdot\lambda$$

The first thing to realize is that the objective function is convex in λ. This is because:

$$\frac{\partial}{\partial\lambda}\left(\frac{1-\alpha-\frac{(1-\rho)(1-\alpha)}{\rho\alpha+(1-\rho)(1-\alpha)}\lambda}{1-\lambda}\right) = \frac{\alpha(1-\alpha)(2\rho-1)}{\rho\alpha+(1-\rho)(1-\alpha)}\frac{1}{(1-\lambda)^2} \geq 0$$

and

$$\frac{\partial^2}{\partial \lambda^2}\left(\frac{1-\alpha-\frac{(1-\rho)(1-\alpha)}{\rho\alpha+(1-\rho)(1-\alpha)}\lambda}{1-\lambda}\right)=2\frac{\alpha(1-\alpha)(2\rho-1)}{\rho\alpha+(1-\rho)(1-\alpha)}\frac{1}{(1-\lambda)^3}\geq 0$$

The convexity of the objective function then follows from the max operator preserving convexity, and the cost of repressing entering as an additive linear term. Below, we provide two representative pictures, Figures 2 and 3, illustrating the objective function.

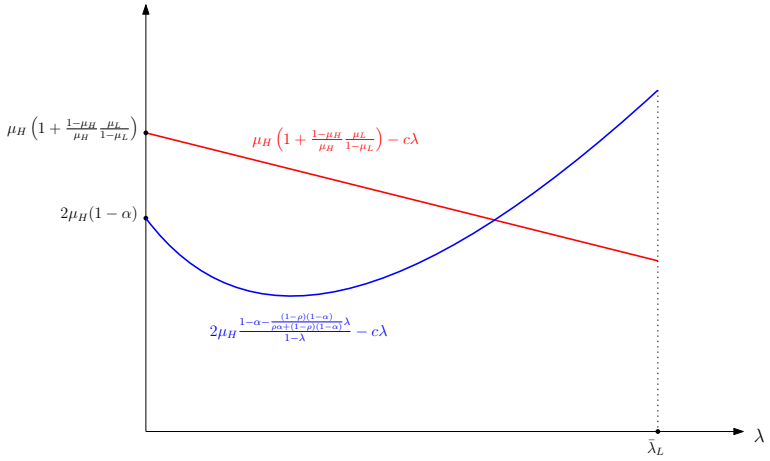


Figure 2: The case where $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}\geq 1-2\alpha$ and $c<c^*(\rho,\alpha)$. The objective function is the upper envelope of the red line and the blue curve.

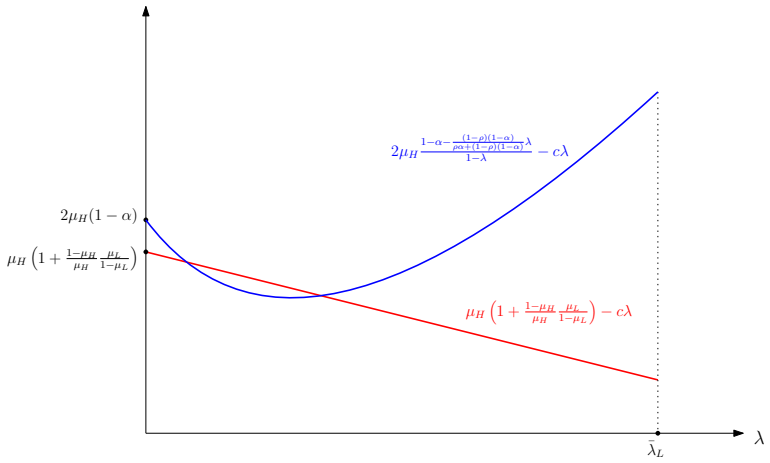


Figure 3: The case where $\frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}\leq 1-2\alpha$ and $c<c^*(\rho,\alpha)$. The objective function is the upper envelope of the red line and the blue curve.

The convexity of the objective function implies that it is sufficient to compare the bounds of parameter space: $\lambda^*\in\{0,\bar{\lambda}_L\}$. Then, $\lambda^*=\bar{\lambda}_L$ if and only if $u(\bar{\lambda}_L,0)\geq u(0,0)$. Substi-

tuting, $\lambda^* = \bar{\lambda}_L$ if and only if:

$$\mu_H \cdot \max\left\{1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}, 2 \cdot \frac{1-\alpha-\frac{(1-\rho)(1-\alpha)}{\rho\alpha+(1-\rho)(1-\alpha)}\bar{\lambda}_L}{1-\bar{\lambda}_L}\right\} - c \cdot \bar{\lambda}_L \geq \mu_H \cdot \max\left\{1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}, 2 \cdot (1-\alpha)\right\}$$

Because $c > 0$ and $\bar{\lambda}_L > 0$, the condition for $\lambda^* = \bar{\lambda}_L$ holds if and only if:

$$\mu_H \cdot 2 \cdot \frac{1-\alpha-\frac{(1-\rho)(1-\alpha)}{\rho\alpha+(1-\rho)(1-\alpha)}\bar{\lambda}_L}{1-\bar{\lambda}_L} - c \cdot \bar{\lambda}_L \geq \mu_H \cdot \max\left\{1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}, 2 \cdot (1-\alpha)\right\}$$

Substituting $\bar{\lambda}_L = \rho\alpha + (1-\rho)(1-\alpha)$, this condition is equivalent to:

$$\mu_H \cdot \frac{2\rho(1-\alpha)}{1-\rho\alpha-(1-\rho)(1-\alpha)} - c \cdot (\rho\alpha + (1-\rho)(1-\alpha)) \geq \mu_H \cdot \max\left\{1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}, 2 \cdot (1-\alpha)\right\}$$

which is equivalent to:

$$c \leq \frac{\mu_H}{\rho\alpha + (1-\rho)(1-\alpha)}\left(\frac{2\rho(1-\alpha)}{1-\rho\alpha-(1-\rho)(1-\alpha)} - \max\left\{2(1-\alpha), 1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}\right\}\right)$$

The result follows. □

Proof of Theorem 2. Fix an incentive compatible persuasion mechanism $\sigma = \left(\sigma_0^L, \sigma_1^L, \sigma_0^H, \sigma_1^H\right)$. Without loss of generality, we assume that $\sigma_0^L \leq \sigma_1^L$ and $\sigma_0^H \leq \sigma_1^L$; otherwise, one can obtain the same result by swapping the labels of messages $m^\tau = 0$ and $m^\tau = 1$.

By (4), a citizen with prior μ_t has the following posterior after a message $m^\tau = 0$:

$$\frac{\mu_t(1-\sigma_1^\tau)}{\mu_t(1-\sigma_1^\tau) + (1-\mu_t)(1-\sigma_0^\tau)} = \frac{\mu_t}{\mu_t + (1-\mu_t)\frac{1-\sigma_0^\tau}{1-\sigma_1^\tau}} \leq \mu_t$$

where the inequality follows because $\sigma_0^\tau \leq \sigma_1^\tau$. Since $\mu_L < \mu_H < \frac{1}{2}$, we conclude that the

posteriors following $m^\tau = 0$ from any information source are below $\frac{1}{2}$. By (6), this implies:

$$a_L(m^L = 0) = a_H(m^H = 0) = 0$$

That is, a pessimistic message always induces $a_i = 0$. Then, (8) becomes:

$$\begin{aligned} v_\sigma(\mu_s) = \mu_s \cdot \alpha \cdot \sigma_1^L \cdot a_L(m^L = 1) + \mu_s \cdot (1-\alpha) \cdot \sigma_1^H \cdot a_H(m^H = 1) \\ + (1-\mu_s) \cdot \alpha \cdot \sigma_0^L \cdot a_L(m^L = 1) + (1-\mu_s) \cdot (1-\alpha) \cdot \sigma_0^H \cdot a_H(m^H = 1) \end{aligned}$$

Rest of the proof proceeds in considering four different cases.

Case 1: $a_L(m^L = 1) = a_H(m^H = 1) = 0$**.** In this case, $v_\sigma(\mu_s) = 0$. The same payoff can be achieved with a fully uninformative mechanism $\hat{\sigma} = \left(\hat{\sigma}_0^L, \hat{\sigma}_1^L, \hat{\sigma}_0^H, \hat{\sigma}_1^H\right)$ where:

$$\hat{\sigma}_0^L = \hat{\sigma}_1^L \qquad \hat{\sigma}_0^H = \hat{\sigma}_1^H$$

Case 2: $a_L(m^L = 1) = 0, a_H(m^H = 1) = 1$**.** In this case,

$$v_\sigma(\mu_s) = (1-\alpha) \cdot \left(\mu_s \cdot \sigma_1^H + (1-\mu_s) \cdot \sigma_0^H\right)$$

Consider the mechanism $\hat{\sigma} = \left(\hat{\sigma}_0^L, \hat{\sigma}_1^L, \hat{\sigma}_0^H, \hat{\sigma}_1^H\right)$ such that:

$$\hat{\sigma}_0^L = \hat{\sigma}_0^H = \sigma_0^H \qquad \hat{\sigma}_1^L = \hat{\sigma}_1^H = \sigma_1^H$$

By construction, $a_H(m^H = 1) = 1$ under $\hat{\sigma}$. By (5) and (6), under $\hat{\sigma}$,

$$a_L(m^H = 1) = \begin{cases} 1, & \text{if } \frac{\sigma_H^0}{\sigma_1^H} \leq \frac{\mu_L}{1-\mu_L}, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore,

$$v_{\hat{\sigma}}(\mu_s) = \begin{cases} \mu_s \cdot \sigma_1^H + (1-\mu_s) \cdot \sigma_0^H, & \text{if } \frac{\sigma_H^0}{\sigma_1^H} \leq \frac{\mu_L}{1-\mu_L}, \\ (1-\alpha) \cdot \left(\mu_s \cdot \sigma_1^H + (1-\mu_s) \cdot \sigma_0^H\right), & \text{otherwise.} \end{cases}$$

which implies: $v_{\hat{\sigma}}(\mu_s) \geq v_\sigma(\mu_s)$.

Case 3: $a_L(m^L = 1) = 1, a_H(m^H = 1) = 0$**.** In this case,

$$v_\sigma(\mu_s) = \alpha \cdot \left(\mu_s \cdot \sigma_1^L + (1 - \mu_s) \cdot \sigma_0^L\right)$$

Consider the mechanism $\hat{\sigma} = \left(\hat{\sigma}_0^L, \hat{\sigma}_1^L, \hat{\sigma}_0^H, \hat{\sigma}_1^H\right)$ such that:

$$\hat{\sigma}_0^L = \hat{\sigma}_0^H = \sigma_0^L \qquad \hat{\sigma}_1^L = \hat{\sigma}_1^H = \sigma_1^L$$

Since $a_L(m^L = 1) = 1$, by (5) and (6), $\frac{\sigma_L^0}{\sigma_1^L} \leq \frac{\mu_L}{1-\mu_L}$. Since $\mu_L < \mu_H$, $\frac{\sigma_L^0}{\sigma_1^L} < \frac{\mu_H}{1-\mu_H}$. Therefore, under $\hat{\sigma}$, $a_H(m^H = 1) = 1$ and

$$v_{\hat{\sigma}}(\mu_s) = \mu_s \cdot \sigma_1^H + (1 - \mu_s) \cdot \sigma_0^H > v_\sigma(\mu_s)$$

Case 4: $a_L(m^L = 1) = a_H(m^H = 1) = 1$**.** In this case,

$$v_\sigma(\mu_s) = \mu_s \cdot \alpha \cdot \sigma_1^L + \mu_s \cdot (1 - \alpha) \cdot \sigma_1^H + (1 - \mu_s) \cdot \alpha \cdot \sigma_0^L + (1 - \mu_s) \cdot (1 - \alpha) \cdot \sigma_0^H$$

Since $a_L(m^L = 1) = 1$, by (6), $\Pr_L\{\omega = 1 | m^L = 1\} \geq \frac{1}{2}$. By (5), then,

$$\frac{\sigma_1^L}{\sigma_0^L} \geq \frac{1 - \mu_L}{\mu_L} \tag{A3}$$

Similarly, since $a_H(m^H = 1) = 1$, (6) and (5) imply:

$$\frac{\sigma_1^H}{\sigma_0^H} \geq \frac{1 - \mu_H}{\mu_H} \tag{A4}$$

Moreover, note that $\mu_L < \mu_H$ implies: $\frac{1-\mu_H}{\mu_H} < \frac{1-\mu_L}{\mu_L}$. Therefore, (A3) implies:

$$\frac{\sigma_1^L}{\sigma_0^L} > \frac{1 - \mu_H}{\mu_H} \tag{A5}$$

We consider two mutually exhaustive cases:

Case 4.1: $\frac{\sigma_1^H}{\sigma_0^H} > \frac{1-\mu_L}{\mu_L}$**.** In this case, let:

$$(\hat{\sigma}_0, \hat{\sigma}_1) \in \arg \max_{(\sigma_0^\tau, \sigma_1^\tau), \tau \in \{L,H\}} \mu_s \cdot \sigma_1^\tau + (1 - \mu_s) \cdot \sigma_0^\tau$$

By (A4) and (A5), $a_H(\hat{m}=1)=1$. Similarly, by (A3) and because $\frac{\sigma_1^H}{\sigma_0^H} > \frac{1-\mu_L}{\mu_L}$, $a_L(\hat{m}=1)=1$. Therefore,

$$\begin{aligned} v_{\hat{\sigma}}(\mu_s) &= \mu_s \cdot \alpha \cdot \hat{\sigma}_1 + \mu_s \cdot (1-\alpha) \cdot \hat{\sigma}_1 + (1-\mu_s) \cdot \alpha \cdot \hat{\sigma}_0 + (1-\mu_s) \cdot (1-\alpha) \cdot \hat{\sigma}_0 \\ &= \alpha \cdot (\mu_s \cdot \hat{\sigma}_1 + (1-\mu_s) \cdot \hat{\sigma}_0) + (1-\alpha) \cdot (\mu_s \cdot \hat{\sigma}_1 + (1-\mu_s) \cdot \hat{\sigma}_0) \\ &\geq \alpha \cdot \left(\mu_s \cdot \sigma_1^L + (1-\mu_s) \cdot \sigma_0^L\right) + (1-\alpha) \cdot \left(\mu_s \cdot \sigma_1^H + (1-\mu_s) \cdot \sigma_0^H\right) \\ &= v_\sigma(\mu_s) \end{aligned}$$

Case 4.2: $\frac{\sigma_1^H}{\sigma_0^H} \leq \frac{1-\mu_L}{\mu_L}$**.** In this case, first note that (A5) implies: $\Pr_H\{\omega=1|m^L=1\} \geq \frac{1}{2}$. Then, by (6), $a_H(m^L=1)=1$. Combining this with $a_H(m^L=0)$, we have

$$U(\mu_H, L) = \mu_H \cdot \left(\sigma_1^L - 1\right) + (1-\mu_H) \cdot \left(-\sigma_0^L\right)$$

whereas:

$$U(\mu_H, H) = \mu_H \cdot \left(\sigma_1^H - 1\right) + (1-\mu_H) \cdot \left(-\sigma_0^H\right)$$

Incentive compatibility requires $U(\mu_H, H) \geq U(\mu_H, L)$, or,

$$\mu_H \cdot \left(\sigma_1^H - 1\right) + (1-\mu_H) \cdot \left(-\sigma_0^H\right) \geq \mu_H \cdot \left(\sigma_1^L - 1\right) + (1-\mu_H) \cdot \left(-\sigma_0^L\right)$$

Rearranging, we conclude that a necessary condition for incentive compatibility is

$$\sigma_1^L \leq \sigma_1^H - \frac{1-\mu_H}{\mu_H} \cdot (\sigma_0^H - \sigma_0^L) \tag{A6}$$

Therefore, an upper bound on the sender's subjective payoff is

$$\begin{aligned} v_\sigma(\mu_s) \leq \mu_s \cdot \alpha \cdot &\left(\sigma_1^H - \frac{1-\mu_H}{\mu_H} \cdot (\sigma_0^H - \sigma_0^L)\right) \\ &+ \mu_s \cdot (1-\alpha) \cdot \sigma_1^H + (1-\mu_s) \cdot \alpha \cdot \sigma_0^L + (1-\mu_s) \cdot (1-\alpha) \cdot \sigma_0^H \end{aligned}$$

Noting that $\mu_s = \mu_H$,

$$\begin{aligned} v_\sigma(\mu_s) &\leq \mu_H \cdot \alpha \cdot \sigma_1^H - (1-\mu_H) \cdot \alpha \cdot (\sigma_0^H - \sigma_0^L) \\ &\quad + \mu_H \cdot (1-\alpha) \cdot \sigma_1^H + (1-\mu_H) \cdot \alpha \cdot \sigma_0^L + (1-\mu_H) \cdot (1-\alpha) \cdot \sigma_0^H \\ &= \mu_H \cdot \sigma_1^H + (1-\mu_H) \cdot (1-2\alpha) \cdot \sigma_0^H + (1-\mu_H) \cdot 2\alpha \cdot \sigma_0^L \end{aligned} \tag{A7}$$

By (A3), $\sigma_0^L \leq \frac{\mu_L}{1-\mu_L}\sigma_1^L$. Combining this with (A6):

$$\sigma_0^L \leq \frac{\mu_L}{1-\mu_L}\left(\sigma_1^H - \frac{1-\mu_H}{\mu_H} \cdot (\sigma_0^H - \sigma_0^L)\right)$$

Rearranging, we have

$$\sigma_0^L \leq \frac{\frac{\mu_L}{1-\mu_L}\sigma_1^H - \frac{\mu_L}{1-\mu_L}\frac{1-\mu_H}{\mu_H}\sigma_0^H}{1 - \frac{\mu_L}{1-\mu_L}\frac{1-\mu_H}{\mu_H}} \tag{A8}$$

Substituting this into (A7), we have

$$\begin{aligned} v_\sigma(\mu_s) &\leq \mu_H \cdot \sigma_1^H + (1-\mu_H) \cdot (1-2\alpha) \cdot \sigma_0^H + (1-\mu_H) \cdot 2\alpha \cdot \frac{\frac{\mu_L}{1-\mu_L}\sigma_H^1 - \frac{\mu_L}{1-\mu_L}\frac{1-\mu_H}{\mu_H}\sigma_H^0}{1 - \frac{\mu_L}{1-\mu_L}\frac{1-\mu_H}{\mu_H}} \\ &= \left(\mu_H + (1-\mu_H) \cdot 2\alpha \cdot \frac{\frac{\mu_L}{1-\mu_L}}{1 - \frac{\mu_L}{1-\mu_L}\frac{1-\mu_H}{\mu_H}}\right)\sigma_1^H + (1-\mu_H) \cdot \left(\frac{1 - 2\alpha - \frac{\mu_L}{1-\mu_L}\frac{1-\mu_H}{\mu_H}}{1 - \frac{\mu_L}{1-\mu_L}\frac{1-\mu_H}{\mu_H}}\right)\sigma_0^H \end{aligned}$$

and $\frac{\sigma_1^H}{\sigma_0^H} \in [\frac{1-\mu_H}{\mu_H}, \frac{1-\mu_L}{\mu_L}]$.

- If $1-2\alpha-\frac{\mu_L}{1-\mu_L}\frac{1-\mu_H}{\mu_H} \geq 0$, the maximum value this expression can take is when $\sigma_1^H = 1$, $\sigma_0^H = \frac{\mu_H}{1-\mu_H}$. Then, the maximum value of the leader's subjective payoff is $v_s(\mu_H) = \mu_H \cdot 2(1-\alpha)$.
- If $1 - 2\alpha - \frac{\mu_L}{1-\mu_L}\frac{1-\mu_H}{\mu_H} \leq 0$, the maximum value this expression can take is when $\sigma_1^H = 1$, $\sigma_0^H = \frac{\mu_L}{1-\mu_L}$. Then, the maximum value of the leader's subjective payoff is $v_s(\mu_H) = \mu_H \cdot \left(1 + \frac{1-\mu_H}{\mu_H}\frac{\mu_L}{1-\mu_L}\right)$.

As shown in Proposition 1, in either case, these payoffs can be achieved with a public persuasion mechanism.

□