Global Manipulation by Local Obfuscation [*]

Fei Li[†] Yangbo Song[‡] Mofei Zhao[§]

August 26, 2020

Abstract

We study information design in a regime change context. A continuum of agents simultaneously choose whether to attack the current regime and will succeed if and only if the mass of attackers outweighs the regime's strength. A designer manipulates information about the regime's strength to maintain the status quo. The optimal information structure exhibits local obfuscation, some agents receive a signal matching the true strength of the status quo, and others receive an elevated signal professing slightly higher strength. Public signals are strictly suboptimal, and in some cases where public signals become futile, local obfuscation guarantees the collapse of agents' coordination.

Keywords: Bayesian persuasion, coordination, information design, obfuscation, regime change

JEL Classification: C7, D7, D8.

[*] We thank Yu Awaya, Arjada Bardhi, Gary Biglaiser, Daniel Bernhardt, James Best, Jimmy Chan, Yi-Chun Chen, Liang Dai, Toomas Hinnosaar, Tetsuya Hoshino, Ju Hu, Yunzhi Hu, Chong Huang, Nicolas Inostroza, Kyungmin (Teddy) Kim, Qingmin Liu, George Mailath, Laurent Mathevet, Stephen Morris, Xiaosheng Mu, Peter Norman, Mallesh Pai, Alessandro Pavan, Jacopo Perego, Xianwen Shi, Joel Sobel, Satoru Takahashi, Ina Taneva, Can Tian, Kyle Woodward, Xi Weng, Ming Yang, Jidong Zhou, and Zhen Zhou for comments.

[†] Department of Economics, University of North Carolina, Chapel Hill. Email: lifei@email.unc.edu.

[‡] School of Management and Economics, The Chinese University of Hong Kong, Shenzhen. Email: yangbosong@cuhk.edu.cn.

[§] School of Economics and Management, Beihang University. Email: zhaomf.06@gmail.com.

1 Introduction

The revolution of information and communication technology raises growing concerns about digital authoritarianism.[1] Despite their tremendous effort to establish information censorship and to spread disinformation, full manipulation of information remains outside autocrats' grasp in the modern age. To optimistic liberals, it keeps the hope alive. This paper begins the formal investigation of whether this hope is justified. We study information design in a canonical regime-change game *à la* Morris and Shin (2003) and derive the optimal information structure to stabilize the regime. The result paints a bleak picture: the most effective policy merely requires creating a minor obfuscation for a fraction of agents. A natural and practical scenario is that an information designer, being, for instance, the propaganda department of an autocratic regime, spreads messages with close but distinct implications across various social media. The designer does not need to control over the identity of receivers of each message, but randomly sows confusion and doubt by dividing message recipients by media coverage.[2]

In our model, an information designer faces a unit mass of agents who simultaneously decide whether to coordinate on an attack. Attacking is costly, and each attacker will be rewarded if and only if the status quo is overthrown. The strength of the status quo, namely the state, is randomly selected from an interval by nature and unknown to the agents. The status quo persists if and only if the total measure of attackers does not exceed its state. If the state is above one, it is *invincible* because the status quo persists under the attack of all agents; otherwise, it is called *vincible*. The information designer commits to a state-dependent information policy that sends a signal, which can be public or private, to each agent, and his objective is to maximize the probability of preserving the regime in its least-preferred (adversarial) equilibrium among agents.

A natural starting point is to consider public persuasion. In some vincible states, the designer sends all agents the same signal as in invincible states, so that agents

[1] See, for example, Yuval Noah Harari, "Why Technology Favors Tyranny," *The Atlantic*, October 322(3), 2018.

[2] It is an open secret that these strategies are widely employed in many authoritarian regimes. For example, King et al. (2017) empirically study online comments by the notorious "50-cent gang" who post fabricated social media comments as if they were the genuine opinions of ordinary citizens. Also, Ong and Cabanes (2018) report interviews of "disinformation architects", the high-level strategists and digital support workers, behind fake news and "digital black operation" in Philippines.

are scared off from attacking when they believe with sufficiently high probability that the true state is invincible. This idea of leveraging on the invincible states has been extensively studied by the literature, which is a natural analogy of the classical single-receiver persuasion (e.g. the jury example in Kamenica and Gentzkow (2011)). However, a coordination game allows the designer, when not constrained to public propaganda, to manipulate information in a more subtle fashion: a state does not have to be truely invincible to be leveraged on, but only needs to convince agents of no sufficient coordination.

In this spirit, the designer may not only leverage on the invincible states, but also on weaker states. This is an iterated, possibly infinite-step process enabled by the coordination nature of the base game: through obfuscating signals, some invincible states are the first tier of leveraged states to save certain weaker states; once these vincible states will never face sufficient coordinated attacks, they in turn become a "conditionally invincible" tier and may be leveraged on to save more vincible states, and so on. This observation echoes the classical Traveller's Dilemma in that agents' deeper epistemological reasoning (on recognizing the conditionally invincible states) grants the designer more advantage. The linkages between these tiers are endogenously characterized by the designer's information policy, so the optimal policy must determine the number of such tiers, the states to be included in every tier, and how to interconnect them in agents' beliefs via obfuscating signals.

Our first contribution is to identify an optimal policy that takes a surprisingly simple form despite the scope and complexity of possible designs. It has an important and novel property that we call *local obfuscation*. Specifically, the first tier contains all invincible states and sends signal s_1 to all agents; the second tier is weaker than the first, and it sends s_1 to a (randomly selected) proportion of agents and another signal s_2 to others; the third is weaker than the second, and it sends s_2 to a proportion of agents, and another signal s_3 to others; and so on. Finally, there is a single weakest tier, characterized by an endogenously determined threshold, that sends a self-identifying signal s_a. In other words, except the invincible and the weakest states, the information designer under each state executes a *local obfuscating* policy, essentially revealing its tier to some agents but deceiving other agents by a slightly stronger tier. Heuristically speaking, the designer treats some agents with loosely defined honesty but others with "alternative facts" that marginally distort the truth.

The optimal local obfuscation collapses *global coordination* among agents by creating both fundamental and belief uncertainty. Under this policy, an agent seeing s_1 will refrain from attacking since the probability of facing an invincible status quo is sufficiently high; realizing this, she will not attack upon s_2 either, because she believes a fraction of her fellow agents are likely to receive s_1 and be deterred, resulting in an insufficient mass of attackers in expectation. Further still, she will not attack upon s_3, because she believes a (larger) fraction of her fellow agents are likely to receive s_2 and be deterred; the iteration goes on. The use of s_1, therefore, generates a ripple effect to squelch agents' attack on lower-tier signals. The iteration then unravels to the limit case that no single agent seeing any s_n will choose to attack, and thus all states beyond the above threshold for sending s_a persist.

Our second contribution is to develop an iterated method to study adversarial information design in classic regime-change games which cannot be solved by any off-the-shelf method. Recall that the standard Bayes correlated equilibrium approach (Bergemann and Morris (2016) and Taneva (2019)) implicitly selects the designer's favorite equilibrium, and our model has a continuum of states and agents, so the recently developed methods relying on concavification over belief hierarchies (Mathevet et al. 2019) or introducing sequential obedience constraints (Morris et al. 2019) which originate from finite games do not immediately apply as well. Our method does not explicitly deal with belief hierarchies but decomposes the information design into a sequence of simple optimizations, each of which is subject to the solutions to the precedent programmings. This approach allows us to quantify the size of the above-mentioned ripple effect, both for every round of iteration and at the limit, as a function of the state distribution.

The third contribution of this paper is to understand how the depth of agents' reasoning determines the implementable outcome of the optimal information design. A higher level of reasoning benefits the information designer by creating more "conditionally invincible" states. We find that when the designer is only capable of manipulating agents' higher-order reasoning up to a finite level k, a local obfuscating policy producing $k+1$ tiers in total is the unique optimal information structure. Thus our result not only highlights the advantage the designer enjoys from agents' higher-level reasoning, but also explicitly identifies the magnitude of this advantage as agents' depth of reasoning improves. Moreover, as far as the implementable outcome is concerned, there exists a one-to-one relation between depth of reasoning and signal complexity: the level-k local obfuscating policy also makes the designer's optimum when agents are fully rational but the designer only has $k+1$ distinct signals at his disposal.

Following up on this point, local obfuscation has a unique advantage over public information structures that send each agent the same signal conditional on the state, whose outcome is the same as when agents are only capable of level-1 reasoning. We demonstrate this advantage in two distinct ways. On the one hand, given a target set of persisting states, optimal local obfuscation allows for a lower threshold of attacking cost to achieve the target than optimal public disclosure, and the difference between the cost thresholds coincides with the conditionally expected strength of the persisting states below one. On the other hand, when the measure of invincible states converges to zero, the optimal public information structure becomes futile, while optimal local obfuscation still manages to save a significant measure of vincible states. A sharp implication of this result is that when the attacking cost is sufficiently high but the measure of invincible states becomes almost negligible, virtually no state persists under public information disclosure, but all states persist under optimal local obfuscation.

When the cost of attacking is sufficiently high, the status quo always persists and the information designer is even relieved of the usual *commitment* concern, which is often a questioned assumption when the regime faces a one-shot, life-or-death change. In fact, despite the name "regime change", the most relevant applications of our analysis are prevalent settings where changes are non-fatal to the designer. A typical scenario is repeated interaction, where a government wishes to implement various policies such as regulations or information censorship. A regime change would be abandonment of the current policy, and the regime's strength would be the threshold of opposing citizens' fraction that can force the government to give up.[3] Holding commitment power thus benefits the government in the long run.[4] The practical solution may take the form of, for instance, a handbook for bureaucracy in

[3] One example would be the censorship of mourning articles for Dr. Li Wenliang, an ophthalmologist who warned about the outbreak of COVID-19. When such articles first appeared at a small scale on a number of Chinese self-media, they were automatically deleted only hours after publication. However, realizing after a few days that a vast number of social media users had been constantly and voluntarily forwarding these contents, the authorities lifted the ban and recognized Dr. Li publicly as martyr in fighting COVID-19.

[4] See Best and Quigley (2017) and Mathevet et al. (2018) who justify the commitment assumption by long-term interaction.

the propaganda department.

Optimal local obfuscation is robust in several ways. For instance, the identification of the threshold of persisting states, and thus the maximum probability of the status quo's persistence, results from the converging iterated process and does not require any prior knowledge of the optimum. Besides, this information structure remains optimal when arbitrary correlation across signals is allowed.

Related literature. This paper contributes to the growing literature on information manipulation in global games (see, e.g., Edmond (2013), Goldstein and Huang (2016, 2018), Inostroza and Pavan (2018) and Basak and Zhou (2018)). Instead of focusing on public or partial information design subject to agents' private signals, we solve for an *unconstrained* optimal information structure. Our optimal information structure preserves some feature of a global game – each agent receives a noisy signal that forces him to take account of more than one possible games, as well as higher-order uncertainty of his opponents. However, compared to an exogenous global game structure that encompasses the entire class of games in an agent's belief, we show that it is optimal to maintain both the fundamental and belief uncertainty locally, i.e. an agent knows that he is in one of at most two sub-classes of games characterized by adjacent strength levels of the regime, and he knows at the same time that all his peers receive one of at most two adjacent signals. An immediate implication is that unlike in standard global games, the incomplete information will not be rich enough to generate signals that make every action sometimes dominant.

In general, information design in games inevitably involves higher-order beliefs manipulation.[5] We show that the endogenous belief hierarchy in canonical regime-change games is manageable and the optimal information structure takes a simple and intuitive form. As an example of their aforementioned belief-base method, Mathevet et al. (2019) study a two-player and binary-state coordination game and illustrate that introducing more informational states than actual ones, thereby creating belief hierarchy among agents, benefits the information designer. In their example, each action is dominant in one state, so the coordination problem emerges only under incomplete information. An immediate consequence is that information manipulation

[5] For example, Hoshino (2019) shows that, using the leverage of strategic uncertainty, agents can be persuaded to take an action profile which satisfies a generalization of risk dominance given any nondegenerate prior. See Bergemann and Morris (2019) for more discussion on the connection between adversarial information design and the literature on higher-order beliefs.

above the agents' second-order strategic reasoning is unnecessary. On the contrary, most states in our model has no dominant action, and so the designer finds it profitable to locally *obfuscates nearby states* to manipulate agents' higher-order strategic reasoning. Also, a key ingredient of local obfuscation policy is to "decieve" more agents by elevated signals at stronger states, which is impossible to be discussed in a two-agent model.

More broadly, our paper belongs to the literature of information design with multiple audiences. See e.g., persuasion in voting games (Alonso and Câmara (2016), Bardhi and Guo (2018), Chan et al. (2019)), team production (Halac et al. 2020), and social network (Galperti and Perego 2019), etc. In this literature, the outstanding performance of discriminatory information structure typically requires the designer to manage the statistical correlation between target signals of agents. On the contrary, the optimal information structure in the current paper is completely anonymous. One exception is Mathevet and Taneva (2020) who study implementable outcome by some familiar indirect information structure in a finite game with strategic complementarity. They find that in certain circumstances "spreading the words" to a selected group of receivers dominates public persuasion.

Organization. The rest of the paper is organized as follows. Section 2 lays out the model. Section 3 presents the main result, which explicitly characterizes the optimal information structure, and compares optimal local obfuscation with optimal public propaganda. Section 4 examines the robustness of our proposed information structure. Section 5 concludes. Proofs, unless otherwise specified, are in the Appendices.

2 Model

Base game. The society is populated by a unit mass of agents, indexed by $i \in [0,1]$. There are two possible regimes, the status quo, and an alternative. Agent i decides to attack the current regime $(a_i = 1)$ or not $(a_i = 0)$.

Regime change needs coordination. Denote the aggregate mass of population that attacks by A such that

$$A = \int_0^1 a_i di.$$

The strength of the status quo is represented by a random variable θ. The status quo persists if and only if $\theta \geq A$. The state is drawn from a commonly known probability distribution on $\Theta \subseteq \mathbb{R}$. The cumulative probability function (CDF) of the distribution $F(\cdot)$ is differentiable for every θ, and let $f(\theta)$ denote its density function.

If an agent does not attack, her payoff is zero. If she attacks, her payoff depends on her action and the regime status: she incurs positive cost $c \in (0, 1)$ regardless of the regime status, and if the regime is overthrown, she receives a benefit, which is normalized to be 1.[6] An agent's utility function is therefore

$$u(a_i, A, \theta) = a_i(\mathbb{1}\{\theta < A\} - c)$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. To avoid a trivial case, we assume that

$$\Theta \equiv (0, \bar{\theta}], \text{ and } \bar{\theta} > 1.$$

In other words, the regime never fails if no agent attacks, and there are states ($\theta > 1$) in which the corresponding base game is dominance solvable with no attack.

Information structure. An information designer *commits* to disclosing information to the agents about the state θ. This is modeled as an information structure consisting of a signal space S and a state-dependent distribution over the signal profile $S^{[0,1]}$, where S contains at least countably infinite distinct signals. For our purpose, it is sufficient to specify the resulting state-dependent distribution $\pi : \Theta \to \Delta(S)$ where $\pi(s|\theta)$ corresponds to the measure of agents receiving signal $s \in S$. Hence, the received signal s also corresponds to the agent's *type*. We assume the information structure is *anonymous*,[7] i.e., agent i receives signal s in state θ with probability $\pi(s|\theta)$ for any i, s, and θ. Put differently, the designer is allowed to send differential signals to agents, but he is unable to discriminate agents based on their identities. Without loss of generality, we further focus on the class of distributions where the density is almost everywhere well-defined and integrable, and thereby restrict our attention to

[6] The benefit can be interpreted as ideology or pecuniary benefits that help to overcome the classic free-rider problem (Olson 1965). For example, agents may view their participation in an attack as beneficial for the society and therefore it directly adds to their utility. See chapter 2.3 of Acemoglu and Robinson (2005) for a comprehensive discussion.

[7] Our configuration of π, which determines with certainty the measure of agents receiving each signal, provides a tractable framework for the exposition of economics in our main results. Meanwhile, our analysis readily extends to the general and more complex cases without determinate measures or agent anonymity. We relegate the generalization to Section 4.

policies under which the regime outcome is measurable in the information designer's information. In the rest of the paper, we use information structure and its resulting distribution of agents' types (S, π) interchangeably unless otherwise noted.

Bayesian game and solution concept. The combination of information structure and base game constitutes a Bayesian game, which proceeds as follows. First, θ is drawn by nature. Then, given an information structure indexed by (S, π), each agent i receives signal $s \in S$ according to π, and all agents simultaneously choose their actions. Agent i's *strategy* $a_i : S \to [0, 1]$ specifies the probability of attack. In a Bayesian Nash equilibrium, given a_{-i} and her own signal s, agent i attacks if and only if she strictly prefers to attack.[8]

For a given information structure, there may be multiplicity due to the coordination nature of the base game. We solve for the information designer's *worst* Bayesian Nash equilibrium to capture the idea of adversarial/robust information design. That is, for each information structure, agents coordinate on a strategy profile such that the largest measure of agents attacks. In the remainder of the article, we refer to the information designer's worst Bayesian Nash equilibrium as (adversarial) *equilibrium*.

The information designer's problem is to choose (S, π) to induce an adversarial Bayesian Nash equilibrium which maximizes the regime's expected probability of persistence.

3 Analysis

We begin with the equilibrium characterization for an arbitrary information structure.

Proposition 1. *For every (S, π), the induced Bayesian game has a unique (adversarial) equilibrium.*

We follow the familiar argument of iterated elimination of strictly dominated strategies (IESDS) to construct an equilibrium. Fix an information structure (S, π), we begin with the most aggressive strategy where all agents attack regardless of their signals. We identify a set of no-attack signals S_1 such that an individual agent finds

[8] The requirement of an agent's strict preference for attacking on defining a Bayesian Nash equilibrium is only technical but without loss of any generality. In this way, the information designer's optimum in preserving the regime can be exactly achieved, rather than only approximated.

attack to be dominated when receiving a signal in S_1. Then we examine an agent's incentive when she believes all other agents play a less aggressive strategy: attack if and only if their signals are outside of S_1. We identify another set of no-attack signals S_2 such that an agent finds it sub-optimal to attack when receiving signals in S_2. Since agents' actions are strategic complementary, the best response to a less aggressive strategy must be less aggressive, making $S_2 \supseteq S_1$. This iteration proceeds further for $S_3, S_4 \cdots$. As k goes to infinity, we obtain the maximal set of no-attack signals $S^* = \lim_{n\to\infty} S_n \subseteq S$. In doing so, we construct an equilibrium where an agent attacks if and only if his signal lies in $S \setminus S^*$.

The equilibrium probability of the status quo being overthrown is unique because we solve for the designer's worst equilibrium. But there may be a multiplicity in the set of non-attack signals. We further show there is a unique equilibrium. Note that one cannot apply the standard iterated dominance argument (Morris and Shin 1998) to prove the uniqueness. This is because we allow for an arbitrary information structure, and so there may not be signals serving as take-offs for the iterated domination for both actions. The uniqueness is indeed driven by our equilibrium selection rule. Suppose two distinct equilibria with different sets of no-attack signals S^* and S^{**}. Since two equilibria induce an identical probability of regime changes, both S^* and S^{**} must contain some exclusive signals respectively. We show that there must be another equilibrium where agents play weakly more aggressive than the following strategy: attack if and only if receiving signals from $S \setminus (S^* \cap S^{**})$. This is, once again, due to the strategic complementarity: a more aggressive strategy leads to a more aggressive best response. However, this equilibrium induces a strictly larger probability of regime change, which leads to a contradiction.

3.1 Main Result

Our main result is the characterization of the optimal information structure, which maximizes the probability of the status quo's persistence. First, we introduce a class of information structures.

Definition 1. *An information structure (S, π) is a **local obfuscator** if*

1. *there is a cutoff state $\theta^* \in [0, \bar{\theta}]$ that partitions the state space into a sequence of intervals $\{(\theta_{k+1}, \theta_k]\}_{k=0}^{\infty} \cup (0, \theta^*]$, where $\theta_0 = \bar{\theta}$, and $\lim_{k\to\infty} \theta_k = \theta^*$,*

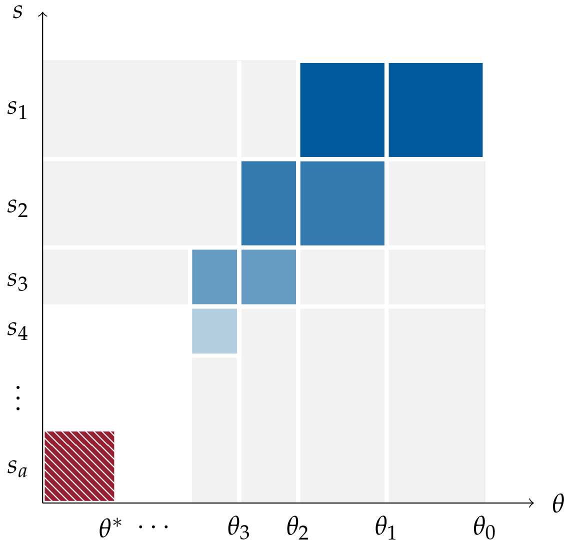

Figure 1: Illustration of local obfuscator. The horizontal axis represents states and the vertical axis represents signals and their face values. We use differential shades to distinguish information sets following different signals. When the state $\theta \in (\theta_1, \theta_0]$, every agent is truly informed by signal s_1 whose face value coincides with the interval containing the state. When $\theta \in (\theta_2, \theta_1]$, some agents receive signal s_2, but others receive signal elevated s_1. Similarly, when $\theta \in (\theta_{k+1}, \theta_k]$, some agents are receive signal s_{k+1}, but others receive the elevated signal s_k. When $\theta \in [0, \theta^*]$, every agent receives signal s_a.

2. *the signal space S is such that $\{s_k\}_{k=1}^{\infty} \cup \{s_a\} = S$, and*
3. *the state-dependent distribution π is such that*

$$\begin{cases} \pi(s_1|\theta) = 1 & \text{if } \theta \in (\theta_1, \theta_0] \\ \pi(s_{k+1}|\theta) + \pi(s_k|\theta) = 1 & \text{if } \theta \in (\theta_{k+1}, \theta_k], \forall k \geq 1 \\ \pi(s_a|\theta) = 1 & \text{if } \theta \in (0, \theta^*] \end{cases}.$$

In other words, if an information structure locally obfuscates agents, a set of adjacent states is categorized into a number of intervals, each of which corresponds to a unique signal. We interpret interval $(\theta_{k+1}, \theta_k]$ as the *face value* of signal s_{k+1}. When the state is $(\theta_{k+1}, \theta_k]$, an agent receives either signal s_{k+1} or a *slightly elevated* signal, s_k. When the state does not belong to any such interval, each agent receives the same signal s_a which conclusively reveals $\theta \in (0, \theta^*]$. Figure 1 visualizes an information structure that exhibits local obfuscation.

We refer to the obfuscation induced by the aforementioned information structure as *local* for two reasons. First, an agent can never distinguish states that belong to the same interval. Second, when an agent is misinformed about the true interval, the signal she receives just marginally elevates the true state interval. Obfuscation makes the agent skeptical about the face value of signals. When receiving signal s_k, instead of taking the signal at face value, the agent believes that the true state is in either $(\theta_{k+1}, \theta_k]$ or $(\theta_k, \theta_{k-1}]$, so her unresolved uncertainty about the fundamental state is local. Moreover, the obfuscation creates belief uncertainty among agents, making the coordination harder. Thanks to the optimal information structure, such a belief uncertainty is also local. An agent who receives signal s_k is uncertain whether other agents receive signals $\{s_{k-1}, s_k\}$ or $\{s_k, s_{k+1}\}$. The information designer can manage agents' posterior beliefs about other agents' signals, beliefs and therefore action profiles by manipulating the information structure.

We are ready to present our main result.

Theorem 1. *The designer's optimum is achieved by a local obfuscator* (S, π^*) *where*

1. *when* $\theta \in (\theta_1, \theta_0]$*, the signal always matches its face value, i.e.*$\pi(s_1|\theta) = 1$*. For each* $k = 1, 2, ...,$ *if* $\theta \in (\theta_{k+1}, \theta_k] \cap \Theta$*, the signal matches its face value with probability* θ *and slightly elevates with the complementary probability, i.e.,*

$$\pi^*(s_{k+1}|\theta) = 1 - \pi^*(s_k|\theta) = \theta.$$

2. *the sequence* $\{\theta_k\}_{k=1}^{\infty}$ *is such that* $\theta_1 = 1$, $\theta_2 = \max\{0, \hat{\theta}_2\}$ *where* $\hat{\theta}_2$ *solves*

$$-c \underbrace{\int_1^{\bar{\theta}} f(\theta)d\theta}_{\theta>1,\ \textit{receive } s_1} + (1-c) \underbrace{\int_{\hat{\theta}_2}^1 (1-\theta)f(\theta)d\theta = 0}_{\theta\in(\hat{\theta}_2, 1],\ \textit{receive } s_1}, \tag{1}$$

$\theta_k = \max\{0, \hat{\theta}_k\}$ *where* $\hat{\theta}_k$ *recursively solves*

$$-c \underbrace{\int_{\theta_{k-1}}^{\theta_{k-2}} \theta f(\theta)d\theta}_{\theta\in(\theta_{k-1}, \theta_{k-2}],\ \textit{receive } s_{k-1}} + (1-c) \underbrace{\int_{\hat{\theta}_k}^{\theta_{k-1}} (1-\theta)f(\theta)d\theta = 0}_{\theta\in(\hat{\theta}_k, \theta_{k-1}],\ \textit{receive } s_{k-1}}, \tag{2}$$

for $k = 3, 4, ...,$ and θ^* is uniquely characterized by

$$\theta^* = \inf\left\{\theta' \in \Theta : \frac{\int_1^{\bar{\theta}} f(\theta)d\theta + \int_{\theta'}^1 \theta f(\theta)d\theta}{\int_{\theta'}^1 (1-\theta)f(\theta)d\theta} \geq \frac{1-c}{c}\right\}. \tag{3}$$

Given (S, π^*), an agent attacks if and only if receiving signal s_a, and the status quo persists if and only if $\theta \in (\theta^*, \bar{\theta}]$.

The equilibrium outcome under optimal design. Theorem 1 says that there is an optimal information structure that exhibits local obfuscation. In other words, to maintain the status quo, the information designer needs only to *slightly exaggerate* the true state to *some* agents. The state set is partitioned into tiers by what signal to send: the invincible tier 1, or $(1, \bar{\theta}]$, always sends s_1 to all agents. When $\theta \leq 1$, state θ in tier k sends a face-value-matching signal s_k to exactly fraction θ of agents (because signals are private i.i.d.) and a slightly elevated signal s_{k-1} to the remaining agents. The fraction θ coincides with the maximum measure of attack that the regime would be able to tolerate, assuming that agents receiving s_{k-1} refrained from attacking. The partition of states is characterized by (1) and (2). The two equations indicate that an agent receiving signal s_k would be indifferent between attacking and not attacking, if she believed that all others would refrain if and only if receiving signals s_{k-1}.

To see why no agent attacks given private signal s_k, $k = 1, 2, ...,$ one may begin with an agent who receives signal s_1. Given her knowledge about π^*, she infers that the true state θ is either in $(\theta_1, \theta_0]$ or in $(\theta_2, \theta_1]$. If $\theta \in (\theta_1, \theta_0]$, the status quo persists regardless of the agents' coordinated action, making attack strictly sub-optimal. If $\theta \in (\theta_2, \theta_1]$, the regime changes only if a sufficiently large amount of agents attack. Since θ_2 balances equation (1), given s_1, the conditional expected benefit of attack does not exceed the cost even if *all other* agents attack for sure. Consequently, the agent does not attack. Given that no agent attacks at signal s_1, consider an agent's belief when receiving s_2. On the one hand, the agent knows for sure that $\theta < 1$ (s_2 implies that either $\theta \in (\theta_3, \theta_2]$ or $\theta \in (\theta_2, \theta_1]$, while $\theta_1 = 1$); on the other hand, she is also aware that attacking will never succeed when $\theta \in (\theta_2, \theta_1]$ because fraction $1 - \theta$ of agents will receive s_1 and never attack. Therefore the best scenario for attacking is when all other agents coordinate to attack given s_2 or s_3, overthrowing the regime when $\theta \in (\theta_3, \theta_2]$. However, since θ_2 and θ_3 balance (2), the agent's expected net payoff from attacking remains non-positive even under the best scenario. Therefore the agent does not attack either given s_2. We can then apply mathematical induction to generate the sequence $\{\theta_k\}_{k=1}^{\infty}$ and associated signals $\{s_k\}_{k=1}^{\infty}$.

The equilibrium regime status is fully determined by θ^*, which is pinned down by (3). If $\theta^* = 0$, the status quo essentially persists for sure; otherwise, in which case θ^* is in fact the limit of θ_k, the regime status is state-dependent. When $\theta \leq \theta^*$, every agent receives signal s_a and attacks, and the status quo collapses. When $\theta > \theta^*$, agents are locally obfuscated and the status quo persists. The optimal local obfuscator has a *global* impact. First, when $\theta > \theta^*$, it collapses all agents' attack by sending disinformation to a proportion of agents only. Second, it suppresses agents' attack in a large set of states through obfuscating nearby states.

For the sake of simplicity in notation, we will henceforth suppress the signal space S and refer to the optimal local obfuscator as π^*. A formal proof of Theorem 1 is in the Appendix; we devote the rest of this section to heuristically explaining the optimality of π^* as characterized above.

Endogenized iterated reasoning. Before explaining the optimality of π^*, we would like to highlight the novel aspect and main challenge for our theoretical analysis. For better exposition, we develop a "credit-discredit" system to describe the hierarchy of endogenously induced beliefs among the agents, endowing the aforementioned email-game-like (higher-order) belief contagions argument an interpretation based on neoclassical consumption theory and classic information design.

We illustrate how the system works by using the example in Figure 1 and considering the process of IESDS that determines the agent equilibrium. To make any agent restrain from attacking given signal s_1 (the first round of IESDS), the signal must induce a sufficiently high belief that she is facing an invincible state ($\theta \geq 1$); hence the invincible states provide the initial endowment of "credit" and the other states sending s_1 consume the credit or create "discredit". To deter agents' attack upon receiving s_1, the credit consumption $\int_{\theta_2}^{1}(1-\theta)f(\theta)d\theta$ must be limited by the credit endowment $\int_{1}^{\bar{\theta}} f(\theta)d\theta$ adjusted by the "relative price" $c/(1-c)$. The budget balance of credit and discredit in equation (1) corresponds to the standard incentive compatible or obedience constraint in canonical information design (Bergemann and Morris (2016) and Taneva (2019)). However, credit consumption does not stop here: for some state $\theta < 1$ sending s_1 (only to a fraction of agents), when the measure of agents receiving s_1 exceeds $1 - \theta$, the state becomes "conditionally invincible" to the rest of agents. That is, the regime persists even if all of these agents manage to coordinate in attacking. The designer can therefore produce additional credit by sending the unbiased signal s_2 to these agents when $\theta \in (\theta_2, \theta_1]$. More < 1 states can then consume this credit, avoid being attacked and further create credit themselves, and the process moves on as IESDS proceeds. This process of credit production and consumption implies that agents' obedience constraints upon receiving each signal, and therefore each step of IESDS, are *endogenously interconnected*. The mix between credit and discredit to restrain an agent's attack upon receiving signal s_k generates a positive externality on the obedience constraint for signal s_{k+1} thanks to the coordination friction.[9]

This system applies to every information structure. In the first round of IESDS under an arbitrary information structure, the agents refrain from attacking given any signal suggesting that the state is sufficiently likely to be invincible. The signal may not be a single determinate one as in the above example, but again the invincible states represent the initial credit endowment. Every $\theta < 1$ in this round, which has convinced at least $1 - \theta$ of agents not to attack, becomes "conditionally invincible" and can continue credit production in the next round. The process then continues analogously. Still, in each round of IESDS, the ratio between the total credit created by the states that persists the former round of IESDS and the total discredit created by the states that persists this round of IESDS must be at least $c/(1-c)$. Note that in an arbitrary information structure, it is not necessary that stronger states create credit earlier than weaker states during IESDS; Figure 2 presents such an example.

The optimality of π^*. Consider an arbitrary information structure and an arbitrary round of IESDS, at the beginning of which a certain amount of (net) credit, leftover from all previous rounds, is available at the information designer's disposal. The information designer's problem is then to select a subset of currently still vulnerable states to exploit the existing credit via creating discredit and thus persist, and at the same time create new credit on their own. Note that for any state with strength $\theta <$

[9] The coordination feature of the base game plays a key role here. When $\theta \in (0, 1)$, neither attacking nor refraining is dominant – attacking is optimal if and only if enough others also attack. Hence in some round of IESDS, a state that will survive even under the currently most adversarial coordination possible creates credit, while another state that only survives by mimicking the former one's signal creates discredit. In the next round, however, the latter state becomes one to create credit with weakened coordination among agents.

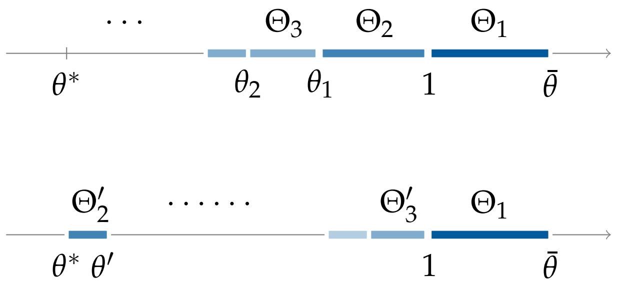

Figure 2: The upper panel corresponds to IESDS for the optimal local obfuscator. Denote $\Theta_1 = (1, \bar{\theta}]$ and Θ_k as the set of states being leveraged in the $k+1$th round of IESDS. As $k \to \infty$, every $> \theta^*$ state is leveraged. In the lower panel, the procedure is similar except in the first round. $\Theta'_2 = (\theta^*, \theta']$ where θ' is chosen to balance the credit constraint, $c \int_1^{\bar{\theta}} f(\theta) d\theta = (1-c) \int_{\theta_*}^{\theta'} (1-\theta) f(\theta) d\theta$. Obviously, the resulting measure of Θ'_2 is less than Θ_2. The choice of Θ'_2 further tightens the credit constraints in subsequent rounds, i.e. $\int_{\Theta'_k} \theta d f(\theta) < \int_{\Theta_k} \theta d f(\theta)$, making the measure of Θ'_{k+1} less than Θ_{k+1} for $k = 3, 4, 5...$

1 to be included in this round, it surely persists from attack as long as it sends a self-identifying signal (credit) to no more than a θ fraction of the agents, while the rest $1-\theta$ fraction of agents receive some signal mimicking a stronger state from the previous round (discredit). This argument reveals a nice duality: on the one hand, the maximum measure of credit it can offer is exactly θ, which is increasing in θ; on the other hand, the minimum measure of discredit it needs to create is $1-\theta$, which is decreasing in θ. In other words, the information designer's conditional optimal choice — which maximizes the additional states that can be saved after this round — is to select the highest states possible. We thus obtain a recursive characterization of the unconditional optimum, summarized by (1) and (2). It then implies that at optimum the persisting states form one unique interval $(\theta^*, \bar{\theta}]$. Figure 2 demonstrates the optimality to leverage states monotonically in IESDS.

By this property, we thus obtain an explicit upper bound for the status quo's probability of persistence, which also identifies a lower bound for a persisting state at optimum, by a straightforward necessary condition which leads to (3) at optimum:

$$\frac{\int_1^{\bar{\theta}} f(\theta) d\theta + \int_{\theta^*}^1 \theta f(\theta) d\theta}{\int_{\theta^*}^1 (1-\theta) f(\theta) d\theta} \geq \frac{1-c}{c}.$$

This inequality is an aggregate budget constraint, implying that the ratio between the total measures of credit and discredit created by $(\theta^*, \bar{\theta}]$ must be at least $\frac{1-c}{c}$. Intuitively, a sufficiently large proportion of credit in the set of no-attack signals need to be provided, to hold an agent at least indifferent between attacking and not.

Finally, we verify that π^* achieves exactly the maximum probability of the status quo's persistence by direct calculation. This can be easily seen by summing up (1) and (2) over k to arrive at (3) at the limit. In π^*, the ratio between credit and discredit in every round of IESDS is kept at precisely $\frac{1-c}{c}$, which automatically preserves the same ratio between the total measures.

Notice that agents coordinate perfectly under π^*. Despite the fundamental and beliefs uncertainty, the outcome of the coordination game is always deterministic. This is intuitive. If an agent receiving signal s has the incentive to attack, the regime must fail with a sufficiently high probability to compensate his attacking cost. In this event, the designer will be better off by encouraging every agent to attack to avoid wasting credit.

The necessity of multiple signals. Although the optimal information structure essentially produces a set of attack signals and another set of no-attack signals, the maximum probability of the status quo's persistence cannot be reached by pooling all signals into binary recommendation signals. To see the logic, first notice that the classic revelation principle/Bayes-correlated-equilibrium approach (See Bergemann and Morris (2016) and Taneva (2019)) implicitly selects the designer's favorite equilibrium. It does not apply if one focuses on the designer's worst equilibrium; and thus it is no longer without loss to focus on "recommendation signals." More importantly, the multiple (and possibly infinite) rounds of IESDS are necessary to maximize the status quo's probability of persistence. The binary recommendation is at best equivalent to the first round of IESDS under the optimal information structure.

3.2 Unique Optimum with Finite Signals

An immediate implication of Theorem 1 is that the outcome induced by local obfuscation, or any information policy, relies on the level of reasoning that the designer can manipulate: the higher the level, the better outcome for the designer. Intuitively, there exists a one-to-one correspondence between the maximum manipulable level of reasoning and the maximum number of available signals: manipulation of up to level-k higher-order reasoning is equivalent, in terms of the optimal outcome, to a restricted set of $k+1$ signals. When k is finite, we show that an optimal information structure must exhibit local obfuscation.

Proposition 2. *For $n = 2, 3, \cdots$, let π_n denote the following state-dependent signal distribution:*

$$\begin{cases} \pi_n(s_1|\theta) = 1 & \text{if } \theta \in (\theta_1, \theta_0] \\ \pi_n(s_k|\theta) = 1 - \pi_n(s_{k-1}|\theta) = \theta & \text{if } \theta \in (\theta_k, \theta_{k-1}] \cap \Theta, \forall k = 2, \cdots, n-1 \\ \pi_n(s_a|\theta) = 1 - \pi_n(s_{n-1}|\theta) = \theta & \text{if } \theta \in (\theta_n, \theta_{n-1}] \cap \Theta \\ \pi_n(s_a|\theta) = 1 & \text{if } \theta \in [0, \theta_n] \cap \Theta \end{cases}.$$

where $S = \{s_k\}_{k=1}^{n-1} \cup \{s_a\}$. Suppose that the information designer is restricted to using S that contains at most n elements; then either

1. *π_n is the unique optimal information policy, or*
2. *under an optimal information policy, no agent ever attacks and the status quo always persists.*

The argument underlying Proposition 2 is centered on maximizing the ripple effect created by the initial credit from $\theta \in (1, \bar{\theta}]$. When only finite signals are available, the agents only go through finite rounds of IESDS. In terms of credit creation, the iterated reasoning process among agents resembles money creation in the banking system to a certain extent. Intuitively, a certain amount of credit created in an earlier round proves more "useful" to the information designer than the same amount of credit in a later round, because it generates a larger sum of additional credit through the remaining rounds. By induction, the optimal information structure must seek to create maximum possible credit in each round sequentially, which uniquely corresponds to π_n.

It is worth noting that, although the optimal local obfuscator is the unique optimal policy when k is finite, uniqueness is not guaranteed for $k = \infty$. In other words, the optimal local obfuscator in Theorem 1 may not be the *only* information structure securing the status quo's persistence for $\theta > \theta^*$.

To understand the multiplicity of optimum, recall the "credit-discredit" interpretation. The optimal local obfuscator not only maximizes the credit production in every round of IESDS, but also uses stocking credit most economically, i.e. saves

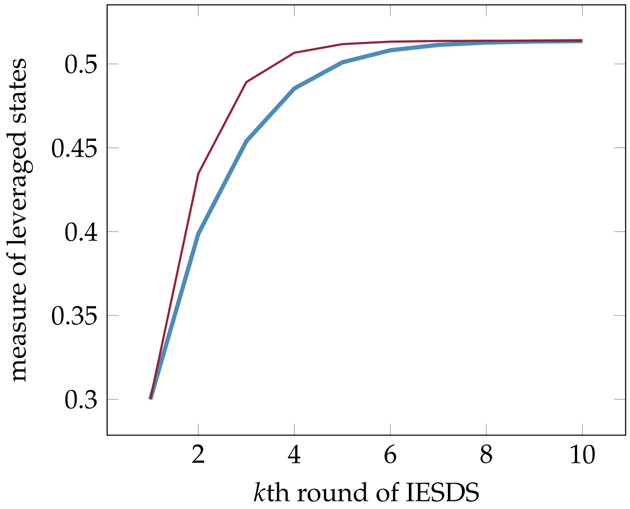

Figure 3: The horizontal axis represents the round of IESDS, and the vertical axis represents the cumulative measure of states being leveraged until each round. The thin red curve corresponds to the optimal local obfuscator π^*, while the thick blue curve corresponds to the alternative information structure π'. The total measure of states being leveraged under π' falls behind that under π^* since the second round of IESDS, but it eventually starts to catch up. When $k = 8$, the difference already shrinks to 0.0003.

the most states given the credit constraint in each round. Nevertheless, alternative designs may exist under which the same *overall* amounts of credit and discredit are created as under the optimal local obfuscator, but different amounts occur in *each round* of IESDS. In such a design, the probability of the status quo's persistence after the first k rounds of IESDS is strictly smaller than in π^* regardless of k; only as the process of IESDS takes infinitely many rounds and the marginal production of credit diminishes to zero, the gap becomes negligible as the procedure forwards.

We give a numerical example below, where $c = 1/6$, and θ is uniformly distributed on $\Theta = [0, 1.1]$. We consider a design that differs from π^* in that it identifies the newly persisting states in the second round of IESDS from θ^* upwards instead of from those in the first round downwards. The result is depicted in Figure 3: after the deviation in the second round, the probability of status quo's persistence under π' is always strictly smaller than under π^* for any k, but will converge to the same limit as $k \to \infty$.

Our discussion above clearly indicates that local obfuscation dominates simple information structures such as public propaganda; after all, a public information policy produces at best the outcome from level-1 manipulation. In the next section, we will highlight this advantage of local obfuscation via comparative static analysis.

3.3 Comparative Statics

We now take a closer look at the optimal local obfuscator π^*. We conduct comparative static analysis on two primitives, the cost of attack, c and the likelihood that attack being dominated, $F(1)$, and examine the advantage of local obfuscation relative to public propaganda.

Public signals. As a benchmark, we first derive the optimal public information structure, i.e., for every state θ, signals received by any two agents i, j must be identical. Straightforwardly, it is optimal to set the signal space to be binary, $S = \{s_a, s_n\}$, and broadcast an attack signal s_a if $\theta \leq \theta^\dagger$ and a no-attack signal s_n otherwise for some cutoff $\theta^\dagger$ solving

$$c = \frac{F(1) - F(\theta^\dagger)}{1 - F(\theta^\dagger)}. \tag{4}$$

The right-hand side of equation (4) is an agent's expected benefit if she attacks given that $\theta > \theta^\dagger$ and all other agents attack. Given the no-attack signal s_n, the agent believes that $\theta > \theta^\dagger$, and finds not to attack to be weakly dominant. This is because when $\theta \in (1, \bar{\theta}]$, attack is a strictly dominated strategy. Obfuscating states on $(\theta^\dagger, \bar{\theta}]$ makes attack an unwise choice given s_n.

To ease the discussion of comparative statics, we rewrite equation (4) as

$$1 - c = \frac{1 - F(1)}{1 - F(\theta^\dagger)}, \tag{5}$$

which is plotted in Figure 4. Naturally, the cutoff value $\theta^\dagger$ is decreasing in c. When $c \geq F(1)$, we have $\theta^\dagger = 0$: agents never attack, and the status quo always persists. When the cost of attack falls, the coordination becomes easier, and the status quo persists in a smaller set of states. As $c \to 0$, $\theta^\dagger \to 1$, and the status quo fails whenever $\theta \notin (1, \bar{\theta}]$. In this case, the leverage caused by the local domination in $(1, \bar{\theta}]$ on lower states vanishes. We summarize the comparative statics results in the following proposition.

Proposition 3.A. *In an optimal public information structure, the ex ante probability that the status quo persists, $1 - F(\theta^\dagger)$ has the following properties.*

1. *It increases in c, converges to $1 - F(1)$ as $c \to 0$, and equals one if $c \geq F(1)$.*

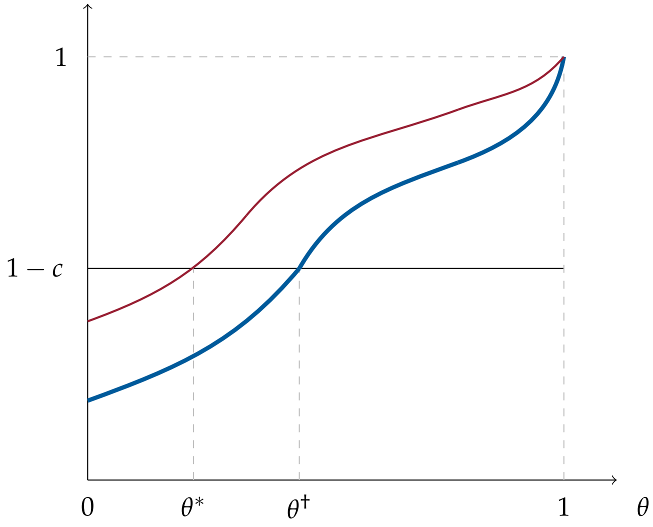

Figure 4: The comparative statics. The thick blue curve represents the right-hand side of equation (5), and the thin red curve represents the right-hand side of equation (6). When c increases, the black curve $1-c$ is shifted down for every θ, and so both θ^* and $\theta^\dagger$ decrease.

2. *When $1-F(1)$ increases, and $f(\cdot)$ decreases arbitrarily and accordingly for $\theta < 1$, $1-F(\theta^\dagger)$ increases. When $1-F(1) \to 0$ and $f(\cdot)$ increases arbitrarily and accordingly for $\theta < 1$, $1-F(\theta^\dagger)$ converges to 0.*

It is worth noting that the second statement immediately implies that the status quo's probability of persistence increases in F in the sense of first-order stochastic dominance, i.e. if the distribution of θ becomes G which first-order stochastic dominates F, the status quo persists with a higher probability under an optimal public information structure.

Local obfuscation. Now we turn to the comparative statics on optimal local obfuscation. Rewrite equation (3) as

$$1-c = \frac{1-F(1)}{1-F(\theta^*)} + \frac{\int_{\theta^*}^{1} \theta f(\theta) d\theta}{1-F(\theta^*)}. \tag{6}$$

Compared to equation (5), equation (6) has a new term on the right-hand side. It captures the total benefit of using local obfuscation through a sequence of signals. Notice that its numerator equals the total credit from the states being leveraged by the local dominance interval $(1, \bar{\theta}]$.

Once again, the cutoff value θ^* is depicted in Figure 4. Higher cost of attack makes the coordination more difficult, and therefore lowers the cutoff state θ^*. Hence, θ^* decreases in c, and converges to 1 as $c \to 0$. If

$$c \geq F(1) - \int_0^1 \theta f(\theta) d\theta, \tag{7}$$

the agents never attack and the status quo never fails. Notice that in this case, the ex ante optimal local obfuscator is also *ex post optimal* to the designer, so it remains credible even if the designer has no commitment power.

The monotonicity of probability of persistence under first-order stochastic dominance is preserved. Indeed, when the state distribution becomes more skewed towards stronger states, more credit and less discredit are created for every given measure of persisting < 1 states. Thus the information designer may prevent more states from being attacked by enrolling them into the iterated process.

Under an optimal information structure $1 - F(\theta^*)$ is bounded away from 0 even if the dominance interval converges to measure 0. The intuition is that a non-public information structure can leverage much more states — those in the dominance interval, as well as those that persist in the subsequent rounds of IESDS. Note that the states below but sufficiently close to 1 actually produce more leverage for subsequent states than consumed from a previous round of IESDS to save them: in particular, every state θ satisfying $\theta > 1 - c$ lies in this category. Then no matter how small $1 - F(1)$ is, it will start the iterated reasoning process that keeps saving lower states, and the process will never stop before $\theta < 1 - c$. Therefore $1 - c$ presents an explicit upper bound for θ^*, meaning that as long as $\theta \in [1 - c, 1]$ with a significant probability, the status quo persists also with a significant probability however small the measure of invincible states is.

The comparative statics is summarized as follows.

Proposition 3.B. *Under an optimal information structure, the ex ante probability that the status quo persists,* $1 - F(\theta^*)$ *has the following properties.*

1. *It increases in* c*, converges to* $1 - F(1)$ *as* $c \to 0$*, and equals one if* $c \geq c^*$*.*

2. *Suppose that* G *first-order stochastically dominates* F*, and let* θ^{**} *denote the lower bound of persisting states under the corresponding optimal local obfuscator given* G*. We have* $1 - G(\theta^{**}) \geq 1 - F(\theta^*)$*.*

3. *Consider $\{F_n\}_{n\in\mathbb{N}^+}$ (with f_n and θ_n^* defined correspondingly) such that $\lim_{n\to\infty} 1 - F_n(1) = 0$, and suppose that $\liminf_{n\to\infty} f_n(\theta) > 0$ for all $\theta \in \hat{\Theta}$, for some non-empty $\hat{\Theta} \subset [1-c, 1]$. Then $\liminf_{n\to\infty} 1 - F_n(\theta_n^*) > 0$.*

Public vs private signals. We are now ready to discuss the advantage of the local obfuscation compared to the public signal (propaganda). One way to examine the advantage is to look at $F(\theta^\dagger) - F(\theta^*)$, the measure of the set of states that coordination is crushed under local obfuscation only.

Proposition 4. *The advantage of local obfuscation relative to public propaganda $F(\theta^\dagger) - F(\theta^*)$ has the following properties:*

1. *It is non-negative for every c, and strictly positive when $c < F(1)$.*
2. *It is increasing in c.*
3. *Suppose that $\{F_n\}_{n\in\mathbb{N}^+}$ (with f_n, $\theta_n^\dagger$ and θ_n^* defined correspondingly) satisfy the conditions in Proposition 3.B. Then $\liminf_{n\to\infty} F_n(\theta_n^\dagger) - F_n(\theta_n^*) > 0$.*

Under the optimal public information structure, even fewer states persist than under π^* after the first round of IESDS. The reason is that the public information structure inevitably wastes some credit provided by $(1, \bar{\theta}]$. For the sake of argument, consider a hypothetical measure 1 of some state $\theta < 1$. The public information structure can save θ from a regime change only by designing for it the same signal as some > 1 state, therefore inducing *all* agents to refrain from attacking. In other words, θ creates discredit of measure 1 as well. Under π^*, however, θ only mimics some > 1 state towards $1-\theta$ fraction of the agents, reducing the measure of discredit produced to only $1-\theta$. The remaining measure of θ then leaves room for more < 1 states to fill with their discredit and persist. Hence as long as the optimal public information structure saves a proportion of states < 1, π^* must be strictly preferred by the information designer (Property 1). It then follows directly from this argument that the additional probability of persistence induced by π^* over the optimal public information structure in the first round of IESDS, as well as that in every subsequent round under π^*, is increasing in c, which leads to Property 2. Note also that both $\theta^\dagger$ and θ^* approach 1 as $c \to 0$; that is, even when non-public information structures are available, an infinitesimal cost always renders information design futile.

Property 3 highlights a significant difference between public and non-public information structures in an extreme scenario. Although $F(\theta^{\dagger}) - F(\theta^*)$ may not be monotone in $1 - F(1)$, the probability measure of invincible states, it does remain bounded away from 0 as the measure gradually becomes negligible. This result implies that using non-public signals indeed bears a unique advantage, which does not vanish even when the optimal public signal becomes almost ineffective. However small the measure of invincible states is, it creates a significant ripple effect by the infinite rounds of IESDS under π^*. The starkest contrast arises when $c > 1 - \int_0^1 \theta f(\theta) d\theta$ and $1 - F(1) \to 0$: almost no state persists under the optimal public information structure, but all states persist under optimal local obfuscation! In this case, the ex ante optimal information policy is also ex post optimal, making our model immune to the usual criticism of perfect commitment assumption.

4 Robustness of Local Obfuscation

In this section, we turn to the robustness of using the optimal local obfuscator. We first show that the optimal local obfuscator remains optimal even if the designer is allowed to target agents according to their identities. Then we argue that a local obfuscator is robust to perturbation, such as private communication and private signal of agents, which undermines the full control of the information structure.

4.1 Relaxing Determinate Measures and Anonymity

Our main results extend to the following general environment. Let S be a compact metric space; the information designer's proposed information structure π is now a mapping from Θ to $\Delta(M(S))$, where $M(S) \subset \{S^{[0,1]}\}$ is a set of integrable functions with codomain S. This configuration allows for (1) arbitrary correlation — given measurability across states, agents and signals — among signals and (2) information structures that target particular agent groups.

With a slight abuse of notation, we adopt the notation π^* for the local obfuscator specified in Theorem 1. We show that its optimality is preserved.

Corollary 1. *π^* remains optimal in the above environment.*

We leave the proof of this result to the Online Appendix. The main intuition is based upon an alternative characterization of the information designer's problem. Since the information designer aims to maximize the ex-ante probability of the status quo's persistence, i.e. the probability measure of the persisting states, we can without loss of generality re-label each state θ as a multiple replica of itself bearing a total density of $f(\theta)$, each representing the same state under a realized measure distribution of signals. Given such a distribution, the state either persists or falls with certainty, in which case we can readily apply the proof of Theorem 1. On a more abstract level, it is only reasonable that as agents are coordinating on the information designer's least preferred equilibrium, introducing no correlation among signals will never hurt. Therefore, compared to the simplest i.i.d. information structure, the ability to target specific agents or to create arbitrary correlation yields no extra leverage for the information designer.

4.2 Exogenous Private Signals

A key distinction between our work and the literature of information design in coordination games is that the latter often assumes a structure of exogenous private signals among agents, so that heterogeneous private beliefs exist even without the designer's input. In our paper, the designer is interpreted as an informational autocratic, and agents are interpreted as citizens. In modern age, it is unrealistic to think that citizens have zero access to alternative news sources. It is therefore reasonable to consider the robustness of the optimal local obfuscator when agents receive exogenous private signals.

Needless to say, when agents' private signals are sufficiently informative, the designer's information manipulation will become fruitless. Here we provide a sketched argument that introducing not-very-informative exogenous private signals does not remove the optimality of local obfuscation. A general and formal analysis requires committing to specific a private information structure, which is left for future research.

Informed elites. One of the simplest but meaningful ways of imposing exogenous private signals is to assume that a fraction $\Delta > 0$ of agents, either randomly or deterministically selected, knows the true state with certainty. In applications, these truth-knowing agents can be regarded as a group of “informed elite” as in Guriev and Treisman (2019). Our Theorem 1 can be directly applied to characterize the optimal information structure, with the minor alteration that every < 1 state above θ^* (which is endogenously determined) sends an elevated signal with probability $\frac{1-\theta}{1-\Delta}$ instead of $1 - \theta$. It is easy to verify that θ^* is increasing in Δ; thus a large number of informed elites is not a good news to the informational autocratic.

General specifications. An environment with more general exogenous private signals bears similar essence in logic. With potentially heterogeneous private signals, a fraction of agents will have better or more optimistic (in the sense that θ is more likely to be low) information about θ and thus become harder to discourage from attacking when $\theta < 1$. Our iterated credit-discredit system remains valid, but the designer has to deliberately shrink the range of θ in each round of IESDS and at the same time make those θ send an elevated signal more often, to once again guarantee that there is never a sufficient measure of agents who may coordinate on attacking. Of course, additional complication arises when the distribution of exogenous signals imposes implicit and non-standard constraints on credit creation in the iterated process, which may render the characterization of optimum less tractable. See Inostroza and Pavan (2018) for a discussion on information design with normally distributed exogenous signals.

4.3 Private Communication

Our final remark regards the robustness of local obfuscation when the designer cannot fully control the information structure. In the multi-agent information design settings, it is well known that using private signals can strictly improve the persuasion outcome as discussed in the introduction. It is often criticized that private communication of agents makes it impossible for the designer to perfectly differentiate agents’ information in a significant amount. It is worth pointing out that our information structure is robust to private communication: collapsing local obfuscation requires a large proportion of agents to exchange a substantial amount of information to resolve both fundamental and belief uncertainty.

To fix the idea, we use a stylized model extension to heuristically illustrate how the main economics of local obfuscation is preserved under limited private communication. The formal analysis is similar to the baseline model, so it is omitted. Suppose that, after receiving the signal sent by the designer, each agent is randomly paired with another agent, with whom she shares her signal. Under local obfuscation, this implies that an agent will end up in one of three possible information sets: two identical low signals that induce a optimistic belief; two identical high signals that induce a pessimistic belief; or two different signals that reveal the true state. To maintain the iterated reasoning process produced by local obfuscation as before, the designer only needs to appropriately increase the probability of sending an elevated signal by each < 1 state – and decrease the range of states that do so – in every round of IESDS, so that enough pessimistic agents will refrain from attacking. In this way, even a truth-knowing agent will not attack because she realizes that the fraction of peers that can possibly coordinate is never sufficient.

The optimal information structure in this setting remains an open question, as communication makes it potentially worthwhile for some state to send more than two signals. However, on the one hand, our above argument suggests that if the probability of each agent meeting another is arbitrarily small, the designer can always use an adjusted local obfuscator to reach a probability of persistence arbitrarily close to the one at optimum without communication. On the other hand, if each agent can share her information with more and more peers, it becomes harder and harder for a < 1 state to create credit through an iterated reasoning process. At the limit, if an agent meets a positive measure of other agents, common knowledge on the signal distribution arises. The optimal information structure then coincides with the optimal public propaganda.

To summarize, as long as the private communication among agents is limited, a local obfuscator is virtually optimal. In our opinion, the private information exchange is insufficient to overturn local obfuscation in many political economy settings. The reasons are twofold.

Indistinguishable nearby states. When the obfuscation is local, only signals representing nearby states are sent simultaneously. It is natural to believe that distinguishing nearby states is more difficult/costly to the agent than distant states. (See Hébert and Woodford (2017), Pomatto et al. (2018), Morris and Yang (2019), and Guo and Shmaya (2019) for formal discussion.) Thus, a slightly exaggerated signal is unlikely to be detected even if agents are allowed to privately verify the signals through a communication or private signals. To fully address this issue is beyond the scope of this paper; we therefore leave it for future research.

Echo chamber. Second, people are more likely to communicate with those they interact with, which in turn creates "echo chambers" that prevent people from being exposed to information that contradicts their preexisting beliefs (see Levy and Razin (2018), Lipnowski and Sadler (2019), and Li and Tan (2019)). In our model, imagine that agents are divided into several chambers, and information exchange is allowed only within a chamber.Since agents from the same chamber tend to share political views and be exposed to similar information sources, it is realistic to allow the designer to send target signals based on agents' chambers. Some randomly selected chambers receive the true signal and others receive the elevated signal. The designer is restricted to sending identical signal to agents from the same chamber, and so the regime's maximum probability of persistence cannot be achieved, but the structure of local obfuscation remains.

5 Conclusion

Our analysis has shown that when the information designer has extensive power in information design, in particular when it can endogenously determine the structure of noise in the agents' information, the optimal persuasion scheme takes a simple and intuitive form. The information designer randomizes between honesty and deceit, which takes the particular form of local obfuscation. We believe that our stylized framework can be enriched to build a research agenda on many related topics, including competitive information designers, dynamic persuasion and communication among agents.

A Proofs of Main Results

A.1 Proof of Proposition 1

We prove the proposition through a number of Lemmas. We begin with an order on the strategy space.

Definition 2. *For i's two strategies a_i and a'_i, we denote that $a_i \geq a'_i$ if $a_i(s) \geq a'_i(s)$ for every $s \in S$, and that $a_i > a'_i$ if $a_i(s) \geq a'_i(s)$ for every $s \in S$ and $a_i(s) > a'_i(s)$ for some*

*$s \in S$. We say a_i is **(weakly) more aggressive** than a_i'.*

The following Lemma regards i's best response given s. It is an immediate consequence of strategic complementarity among agents' actions. It says that when every other agents' strategies become more aggressive, an agent's best response is either unchanged or more aggressive.

Lemma 1. *Consider two strategy profiles of agents other than i, a_{-i} and a'_{-i}. Suppose that $a_j \geq a_j'$ for every $j \neq i$, and that $a_j > a_j'$ for all j in a subset of $[0,1] \setminus \{i\}$ with positive measure. If it is optimal for agent i to attack given $s \in S$ and a'_{-i}, it is also optimal to attack given s and a_{-i}. Similarly, if it is optimal for agent i not to attack given s and a_{-i}, it is also optimal not to attack given s and a'_{-i}.*

Proof. We prove the first part of the lemma. The proof of the second part is almost identical and therefore omitted. Fix $s \in S$, the signal of agent i. Suppose that it is optimal for agent 1 to attack given a'_{-i} and signal s, and suppose that $a_j \geq a_j'$ for every $j \neq i$, and that $a_j > a_j'$ for all j in a subset of $[0,1] \setminus \{i\}$ with positive measure. We must have

$$\begin{aligned} c &< \int_\Theta \left(\frac{f(\theta)\pi(s|\theta)}{\int_\Theta f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < \int_{[0,1]\setminus\{i\}} \int_S a_j'(v)\pi(v|\theta)dvdj\} \right) d\theta \\ &\leq \int_\Theta \left(\frac{f(\theta)\pi(s|\theta)}{\int_\Theta f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < \int_{[0,1]\setminus\{i\}} \int_S a_j(v)\pi(v|\theta)dvdj\} \right) d\theta. \end{aligned}$$

where the first inequality holds because of the optimality of attack given s and a'_{-i}, and the second inequality holds because $a_j \geq a_j'$ for every $j \neq i$, and that $a_j > a_j'$ for all j in a subset of $[0,1] \setminus \{i\}$ with positive measure. Thus, agent i finds it optimal to attack given signal s and a_{-i}. □

Now we are ready to address the equilibrium existence.

Lemma 2. *For any (S, π), there exists an equilibrium.*

Proof. Fix (S, π), we construct an equilibrium through *iterated elimination of strictly dominated strategies (IESDS)*. We begin with the strategy profile that everyone attacks, denoted by a_i^0, $a_i^0(s) \equiv 1$ for every i, and $s \in S$. Define $S_1 \subseteq S$ as the set of signal s

such that

$$\int_{\Theta} \frac{f(\theta)\pi(s|\theta)}{\int_{\Theta} f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < \int_{[0,1]} \int_{S} a_j^0(v)\pi(v|\theta)dvdj\}d\theta \leq c. \tag{8}$$

The left-hand side of (8) is the probability of $\theta < 1$ given s. Hence the condition means that, if agent i receives signal $s \in S_1$, he weakly prefers not to attack even if all other agents attack for certain. Define a_i^1

$$a_i^1(s) = \begin{cases} 0 & \text{if } s \in S_1 \\ 1 & \text{otherwise,} \end{cases}$$

which is weakly less aggressive than $\mathbf{a}_i^0$. By Lemma 1, an agent i weakly prefers not to attack if all other agents play a_i^1.

For $k = 2, 3, \cdots$, define $S_k \subseteq S$ as the set of signal s such that

$$\int_{\Theta} \frac{f(\theta)\pi(s|\theta)}{\int_{\Theta} f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < \int_{[0,1]} \int_{S} a_j^{k-1}(v)\pi(v|\theta)dvdj\}d\theta \leq c.$$

and define

$$a_i^k(s) = \begin{cases} 0 & \text{if } s \in S_k \\ 1 & \text{otherwise.} \end{cases}$$

Notice that a_i^k becomes weakly less aggressive as k increases. By Lemma 1, $S \supseteq S_k \supseteq S_{k-1}$ for every k. At the limit as $k \to \infty$, the set $S^* = \lim_{k\to\infty} S_k$ exists, and $S^* \subseteq S$. Also, define:

$$a_i^*(s) = \begin{cases} 0 & \text{if } s \in S^* \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

for each agent i. Notice that S_1 may be empty. In that case, $S_k, S^* = \emptyset$.

Next we show that the strategy profile specified in (9) is indeed an equilibrium. First, by the construction of S^*, an agent prefers attacking when receiving every signal in $S \backslash S^*$ given other agents follow the strategy specified in (9). Second, we show that for any non-empty S^*, given that the other agents follow the strategy in (9), an individual agent i strictly prefers not attacking for every signal in S^*. The proof is straightforward. Pick any $s \in S^*$, there exists a unique k such that $s \in S_k \backslash S_{k-1}$. By the definition of S_k, given that the other agents $j \neq i$ follow a_j^{k-1} and do not attack if and only if receiving signals in S_{k-1}, an individual agent prefers not attacking when receiving signals in $S_k \backslash \bar{S}_{k-1}$. Then by Lemma 1, if other agents $j \neq i$ follow a less aggressive strategy $a_j^* < a_j^{k-1}$ and do not attack if and only if receiving signals in $S^* \supseteq S_{k-1}$, an agent must prefer not attacking when receiving signals in $S_k \backslash S_{k-1}$. Thus, for every $s \in S^*$, $a_i^*(s) = 0$. Hence, we have the desired result. □

The definitions above guarantee a unique series of $\{S_k\}$ and a unique S^*. In what follows, we show that $a_i^*(s)$ is the unique equilibrium as well.

Lemma 3. *For any (S, π), there is a unique equilibrium.*

Proof. We first introduce a useful function form representing the agent's strategy. Define the measure of agents who attack in state θ as

$$A_\theta = \int_S a_i(s)\pi(s|\theta)ds, \forall \theta \in \Theta.$$

Pick any two agents i, j, in every equilibrium, the expected payoff of attacking when receiving signal s is

$$\int_\Theta \frac{f(\theta)\pi(s|\theta)}{\int_\Theta f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < A_\theta\}d\theta - c,$$

which is invariant to the identity of the agent. As we solve for the information designer's worst BNE, every equilibrium strategy must be *symmetric*, i.e. for every s, $a_i(s) = a_j(s) = a(s)$, which takes value either 0 or 1. For the sake of contradiction, suppose that for (S, π), there are two distinct equilibria a, a'. Let $\{s|a(s) = 0\}$ denote the set of signals the agents do not attack in equilibrium a and $\{s|a'(s) = 0\}$ denote the set of signals the agents do not attack in equilibrium a'. By the hypothesis that a and a' are distinct equilibria, $\{s|a(s) = 0\} \neq \{s|a'(s) = 0\}$. Moreover, they induce an identical probability of a regime being overthrown, so each set must contain some exclusive signals. Consider the following strategy a'' defined as follows:

$$a''(s) = \begin{cases} 0 & \text{if } s \in \{s|a(s) = 0\} \cap \{s|a'(s) = 0\} \\ 1 & \text{otherwise,} \end{cases}$$

which is strictly more aggressive than a and a'. By Lemma 1, an individual agent i

receiving a signal in $S\backslash(\{s|a(s)=0\}\cap\{s|a'(s)=0\})$ prefers attacking if every other agent is adopting strategy a''. Note that an equilibrium always exists, thus there must exist an equilibrium where the agents play at least as aggressively as a''. In such a case the regime changes with a greater probability than both in a and in a', which is a contradiction. □

The combination of Lemmas 1-3 yields Proposition 1.

A.2 Proof of Theorem 1

Step 1. *We define two series which will be useful in the following analysis. Given any information policy, these series are identified through IESDS; and they characterize the agents' iterative reasoning in coordination. Series $\{S_k\}_{k=1}^{\infty}$ is drawn from the proof of Proposition 1; it contains the signal sets which the agents refrain from attacking after the kth round of IESDS. Series $\{T_k\}_{k=1}^{\infty}$ satisfies the following condition: $\cup_{n=1}^{k}T_n$ contains the states that persist before the kth round of IESDS.*

Define state set $T_1=(1,\bar{\theta}]$. By the definition of S_1, for every $s\in S_1$, s induces the following posterior: the probability that the true state is in T_1 is larger than $1-c$, i.e.

$$\Pr(\theta\in T_1|s)\geq 1-c,\forall s\in S_1.$$

Next, we recursively define T_k as the set of states θ where more than $1-\theta$ measure of agents receive signals in S_{k-1} , i.e.

$$T_k\equiv\{\theta\in\Theta:\int_{s\in S_{k-1}}\pi(s|\theta)ds>1-\theta\}$$

for every $k=2,3,...$ Then, by the definition of S_k, for every $s\in S_k$, s induces the following posterior: the probability that the true state is in $\cup_{n=1}^{k}T_n$ is larger than $1-c$, i.e.

$$\Pr(\theta\in\cup_{n=1}^{k}T_n|s)\geq 1-c,\forall s\in S_k.$$

Finally, denote

$$T^*=\cup_{k=1}^{\infty}T_k.$$

For convenience, we also define $T_0=S_0=\emptyset$.

Note that for every k, S_k, S^*, T_k, and T^* are π specific, and we use $S_k|\pi$, $S^*|\pi$, $T_k|\pi$, and $T^*|\pi$ to denote the corresponding sets under information policy when necessary.

Step 2. *We prove that a necessary and sufficient condition for the regime to persist is $\theta \in T^*$.*

We first show the sufficiency. If $\theta \in T^*$, there exists k such that $\theta \in T_k$ and $\theta \notin T_l$ for $l = 1, 2, ..., k-1$. We show that the regime persists for any $k = 1, 2, ...$ Suppose the agents coordinate on attacking if their signals are in S; then by the definition of T_1 and S_1, an individual agent whose signal is in S_1 would prefer to deviate to not attacking. By the rule of coordination, no agent shall attack if her signal is in S_1, and every $\theta \in T_1$ always persists under information policy $\pi(\cdot|\theta)$. By a similar argument, suppose the agents coordinate on attacking if their signals are in $S \backslash S_1$; then an individual agent whose signal is in S_2 would prefer to deviate to not attacking, and every $\theta \in T_1 \cup T_2$ always persists. The rest of the proof follows by mathematical induction.

We prove the necessity by contrapositive. First, by the proof of Proposition 1, every agent shall attack if and only if her signal realization is not in S^*. Then by the definition of T^*, for every state θ not in T^*, the designer sends a signal in S^* with probability less than $1-\theta$; otherwise θ is in T^*. Thus, every state θ not in T^* is attacked by a mass greater than θ and eventually fails. This completes the proof of the necessity.

Step 3. *We identify an upper bound of the ex ante probability that the regime persists, $\int_{T^*} f(\theta)d\theta$.*

Fix any information structure (S, π), and define a function $T : T^* \to \mathbb{N}$ such that for every $\theta \in T^*$, we have $\theta \in \cup_{n=1}^{T(\theta)} T_n \backslash \cup_{n=1}^{T(\theta)-1} T_n$. By definition, $T(\theta)$ is unique for every θ. Intuitively, for every $\theta \in T^*$, $T(\theta)$ means that θ persists after and only after $T(\theta) - 1$ rounds of IESDS.

For $k = 1, 2, ...$, define "discredit D_k":

$$D_k = \int_{\cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n} f(\theta) \int_{S_{k-1}} \pi(s|\theta) ds d\theta,$$

which is the measure of signals in S_{k-1} being sent for all $\theta \in \cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n$. Similarly, for $k = 1, 2, \cdots$, $p = k+1, k+2, \cdots$, define "credit $C_{k,p}$":

$$C_{k,p} = \int_{\cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n} f(\theta) \int_{S_p \backslash S_{p-1}} \pi(s|\theta) ds d\theta,$$

which is the measure of signals in $S_p \backslash S_{p-1}$ being sent when $\theta \in \cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n$. Intuitively, in each round of the IESDS, to deter a coordinated attack, a signal must induce a posterior belief that the true state is sufficiently likely to be strong; hence the chance of defeating it is sufficiently low.

Consider an arbitrary round k. The states that persist after the $k-1$th round of IESDS are $\cup_{n=1}^{k} T_n$; these states are considered as the strong states. A strong state discourages the agents from attacking the signals that it sends with positive probability, increasing the probability that the underlying true state is strong. Analogously, it's like providing "credit" to these signals. The amount of credit a strong state provides to a signal is the ex-ante probability measure that it sends this signal. Then $C_{k,p}$ is the aggregate credit all the states in $\cup_{n=1}^{k} T_n \backslash \cup_{n=1}^{k-1} T_n$ provide to the signals saved in the pth round of the IESDS, $S_p \backslash S_{p-1}$.

The weak states, however, are the states that still fail after the $k-1$th round of IESDS. A weak state encourages the agents to attack the signals that it sends with positive probability, decreasing the probability that the underlying true state is strong. Analogously, it's like drawing out credit from (or injecting "discredit" to) those signals. The amount of credit a weak state charges from a signal is the ex-ante probability measure that it sends this signal. Then D_{k+1} is the aggregate discredit all the states in $\cup_{n=1}^{k+1} T_n \backslash \cup_{n=1}^{k} T_n$ charge from the signals saved in the previous rounds S_k.

In the kth round of the IESDS, to save the states in $\cup_{n=1}^{k+1} T_n \backslash \cup_{n=1}^{k} T_n$, the information policy charges credit D_{k+1} from the signals saved in this and the previous rounds, S_k. The credit must not be overdrawn (specified later); otherwise those signals become too weak and the agents shall attack them in previous rounds. In the pth round of the IESDS, however, these states, $\cup_{n=1}^{k+1} T_n \backslash \cup_{n=1}^{k} T_n$, are strong states; they provide credit $C_{k+1,p}$ to save the states in $\cup_{n=1}^{p+1} T_n \backslash \cup_{n=1}^{p} T_n$. In conclusion, in each round of the IESDS, the newly saved states "pollute" the strong signals endorsed by the states saved in the previous rounds; nevertheless it creates spaces for the states saved in the latter rounds to pollute.

The above intuition leads, for every round k, to two conditions that characterize an upper bound of the ex-ante probability that the regime persists: first, the credit must not be overdrawn; second, the states in $\cup_{n=1}^{k+1} T_n \backslash \cup_{n=1}^{k} T_n$ are saved in the kth round. Precisely, consider round k. By the definition of $T_{(\cdot)}$ and $S_{(\cdot)}$, for every $s \in S_k$, an individual agent receiving s shall not attack even if every agent receiving a signal not in S_{k-1} attacks; that is to say, if she attacks, the probability of winning is smaller than or equal to c. Consider the coordination pattern at the beginning of the kth round of the IESDS, by definition the regime fails if the true state is in T_{k+1} and persists if and only if the true state is in $\cup_{n=1}^{k} T_n$, thus a necessary condition for an individual agent to not attack when receiving any signal in S_k is:

$$c \geq \frac{D_{k+1}}{\sum_{m=1}^{k} C_{m,k} + D_{k+1}}.$$

Then consider all the previous rounds of the IESDS, a necessary condition for $\cup_{n=1}^{k+1} T_n$ to be saved is

$$c \geq \frac{\sum_{m=1}^{k+1} D_m}{\sum_{m=1}^{k} \sum_{p=m}^{k} C_{m,p} + \sum_{m=1}^{k+1} D_m} \Leftrightarrow c \sum_{m=1}^{k} \sum_{p=m}^{k} C_{m,p} \geq (1-c) \sum_{m=1}^{k+1} D_m.$$

Also, by the definition of $T_{(\cdot)}$ and $S_{(\cdot)}$, for $m = 1, 2, \cdots, k+1$, for every $\theta \in \cup_{n=1}^{m} T_n \setminus \cup_{n=1}^{m-1} T_n$, $\int_{S_{m-1}} \pi(s_i|\theta) ds_i \geq \min\{0, 1-\theta\}$; this further implies $\int_{S_k \setminus S_{m-1}} \pi(s_i|\theta) ds_i \leq \max\{1, \theta\}$.

Expanding the above condition yields

$$\begin{aligned}
& c \sum_{m=1}^{k} \sum_{p=m}^{k} C_{m,p} \geq (1-c) \sum_{m=1}^{k+1} D_m \\
\Leftrightarrow \quad & c \sum_{m=1}^{k} \int_{\cup_{n=1}^{m} T_n \setminus \cup_{n=1}^{m-1} T_n} f(\theta) \int_{S_k \setminus S_{m-1}} \pi(s|\theta) ds d\theta \\
\geq \quad & (1-c) \sum_{m=1}^{k+1} \int_{\cup_{n=1}^{m} T_n \setminus \cup_{n=1}^{m-1} T_n} f(\theta) \int_{S_{m-1}} \pi(s|\theta) ds d\theta \\
\Leftrightarrow \quad & c \int_{\cup_{n=1}^{k} T_n} f(\theta) \int_{S_k \setminus S_{T(\theta)-1}} \pi(s|\theta) ds d\theta \\
\geq \quad & (1-c) \int_{\cup_{n=1}^{k+1} T_n} f(\theta) \int_{S_{T(\theta)-1}} \pi(s|\theta) ds d\theta \\
\Rightarrow \quad & c(\int_{T_1} f(\theta) d\theta + \int_{\cup_{n=2}^{k} T_n \setminus T_1} \theta f(\theta) d\theta) \geq (1-c) \int_{\cup_{n=1}^{k} T_n \setminus T_1} (1-\theta) f(\theta) d\theta
\end{aligned}$$

Next we introduce a lemma that helps us to focus on information policies that induce a specific form of agent equilibrium. Intuitively, it shows that the information designer can always construct an information policy π' that weakly improves upon π by, loosely speaking, saving the strong states.

Lemma 4. *For every information policy π, there exists another information policy π' such that $T^*|\pi' \supseteq (F^{-1}(1-\int_{T^*|\pi} f(\theta)d\theta), \bar{\theta}]$.*

Now we are in the position to identify an upper bound of the ex ante probability that the regime persists, $\int_{T^*} f(\theta)d\theta$. Let $\tilde{\theta} = F^{-1}(1-\int_{T^*} f(\theta)d\theta)$, by Lemma 4, as $k \to \infty$ we have

$$\begin{aligned} & c(\int_{T_1} f(\theta)d\theta + \int_{\cup_{n=2}^{\infty} T_n \setminus T_1} \theta f(\theta)d\theta) \geq (1-c)\int_{\cup_{n=1}^{\infty} T_n \setminus T_1} (1-\theta)f(\theta)d\theta \\ \Rightarrow \quad & c(\int_1^{\bar{\theta}} f(\theta)d\theta + \int_{\tilde{\theta}}^1 \theta f(\theta)d\theta) - (1-c)\int_{\tilde{\theta}}^1 (1-\theta)f(\theta)d\theta \geq 0 \end{aligned}$$

Suppose that π improves and $\int_{T^*} f(\theta)d\theta$ increases, $\tilde{\theta}$ decreases, the left-hand side of the second inequality above either always increases, or increases at first, then decreases. Thus, there exists a unique lower bound of $\tilde{\theta}$. If the lower bound is lower than or equal to 0, there exists π such that every state persists; otherwise if the lower bound is strictly larger than 0, we use $\theta^{*\prime}$ to denote this lower bound, and $\theta^{*\prime}$ solves

$$c\int_1^{\bar{\theta}} f(\theta)d\theta + \int_{\theta^{*\prime}}^1 (\theta + c - 1)f(\theta)d\theta = 0 \tag{10}$$

It's straightforward that $\theta^{*\prime}$ is unique, and then the upper bound of the ex ante probability that the regime persists is $1 - F(\theta^*)$.

Step 4. *We show that the probability of the status quo's persistence under π^* exactly equals to the upper bound we proposed. π^* is therefore an optimal signal.*

As shown in the main text, the equilibrium outcome under π^* is that every agent who receives a signal in $\{s_k\}_{k=1}^{\infty}$ does not attack; as a result, the status quo persists whenever $\theta \in (\theta^*, \bar{\theta}]$. When receiving s_a, it is common knowledge that the state is in $(0, \theta^*]$, so all agents attack, and the status quo is overthrown. Also, by the definition of T_k, under π^*, for $k = 1, 2, \cdots$, we have $T_k = (\theta_k, \theta_{k-1}]$.

Then by (1), (2), and (3)

$$\begin{aligned} c\left(\int_{\theta_1}^{\theta_0} f(\theta)d\theta + \sum_{k=3}^{\infty}\int_{\theta_{k-1}}^{\theta_{k-2}} \theta f(\theta)d\theta\right) &= (1-c)\sum_{k=2}^{\infty}\int_{\theta_k}^{\theta_{k-1}} (1-\theta)f(\theta)d\theta \\ c\left(\int_{\theta_1}^{\theta_0} f(\theta)d\theta + \int_{\theta^{*\prime}}^{\theta_1} \theta f(\theta)d\theta\right) &= (1-c)\int_{\theta^{*\prime}}^{\theta_1} (1-\theta)f(\theta)d\theta \end{aligned}$$

Notably, θ^* indeed solves (10); as the solution is unique, we have $\theta^* = \theta^{*\prime}$. The measure of $\int_{T^*} f(\theta)d\theta$ exactly equals the upper bound we proposed; thus θ^* is optimal.

Lastly, all the steps above assume that not every state persists under π^*. If otherwise, for some k we have $\theta_k < 0$, then π^* has already achieved the best outcome possible and is thus optimal.

A.3 Miscellaneous Proofs

Proof of Lemma 4. Given an arbitrary information structure π, we construct π' based on the following idea: find an interval $[\hat{\theta}, \bar{\theta}]$ with the same probability measure as T^*, such that (1) every signal (distribution) sent by some state in T^* under π is also sent by some state in $[\hat{\theta}, \bar{\theta}]$ under π' and (2) the signal sent by a higher state in T^* under π is sent by a higher state in $[\hat{\theta}, \bar{\theta}]$ under π'.

Specifically, for arbitrary $\check{\theta}$ that decreases from $\bar{\theta}$, π' is defined by

$$\int_{T^*\cap(\check{\theta},\bar{\theta}]} f(\theta)\pi(\cdot|\theta)d\theta \equiv \int_{F^{-1}(1-\int_{T^*\cap(\check{\theta},\bar{\theta}]} f(u)du)}^{\bar{\theta}} f(\theta)\pi'(\cdot|\theta)d\theta$$

and

$$\int_{(\check{\theta},\bar{\theta}]\setminus T^*} f(\theta)\pi(\cdot|\theta)d\theta \equiv \int_{F^{-1}(1-\int_{(\check{\theta},\bar{\theta}]\cup T^*} f(v)dv)}^{F^{-1}(1-\int_{T^*} f(w)dw)} f(\theta)\pi'(\cdot|\theta)d\theta.$$

Next we prove that every state in $T' \equiv (F^{-1}(1 - \int_{T^*|\pi} f(\theta)d\theta), \bar{\theta}]$, which bears the same probability measure as T^*, persists under information policy π'. Intuitively, note that by the construction of π', every signal sent by some state in T^* under π is sent by some higher state under π. Then whenever an agent refrains from attacking given a signal under π in fear of facing a strong state, this incentive only becomes stronger given the same signal under π'.

Under π', an agent receiving any signal (realization) has a belief that the state is in T', which is the same as the belief she would have that the state is in T^* given the same signal under π. Now consider an infinitely small measure of states around $\theta \in T^*$. By the above construction, we know that there exists a corresponding set of states around $\theta' \in T'$, the ex ante probability of the two sets of states are identical, the latter set induces exactly the same signal distribution under π' as the former set induces under π, and $\theta' \geq \theta$. Given θ, any potential coordination that is not self-sustainable (i.e., some agents shall unilaterally deviate) under π is weakly less likely to successfully overthrow a regime of state θ' under π'; therefore, such a coordination yields weakly less expected payoff for every agent and is not self-sustainable under π'. Similarly, any coordinated attack that fails to overthrow a regime of state θ under π also fails to overthrow a regime of state θ' under π'. As θ persists under π, θ' persists under π'. Then as every state in T^* persists under π, every state in T' persists under π'. □

Proof of Proposition 2. Suppose that π_n' is an optimal policy and π_n' is different from π_n. Through the following analysis we assume $\theta_k \geq 0$ for every $k \leq n$.

It's straightforward that n signal realizations can induce at most $n-1$ rounds of IESDS. Suppose that π_n' induces $m \leq n-1$ rounds of IESDS.

For π_n, we have

$$\begin{aligned}
D_1|\pi_n &= 0, C_{1,1}|\pi_n = \bar{\theta} - 1 \\
D_2|\pi_n &= \frac{c}{1-c} C_{1,1}|\pi_n \\
D_2|\pi_n + D_3|\pi_n &= \frac{c}{1-c}(C_{2,2}|\pi_n + C_{1,1}|\pi_n) \\
&\cdots \\
\sum_{p=2}^{n} D_p|\pi_n &= \frac{c}{1-c} \sum_{p=2}^{n} C_{p-1,p-1}|\pi_n.
\end{aligned}$$

For π_n', we have

$$\begin{aligned}
D_1|\pi_n' &= 0 \\
D_2\pi_n' &\leq \frac{c}{1-c} C_{1,1}|\pi_n' \\
D_2|\pi_n' + D_3|\pi_n' &\leq \frac{c}{1-c}(C_{2,2}|\pi_n' + C_{1,1}|\pi_n' + C_{1,2}|\pi_n') \\
\sum_{p=2}^{4} D_p|\pi_n' &\leq \frac{c}{1-c}(C_{3,3}|\pi_n' + \sum_{p=2}^{3} C_{2,p}|\pi_n' + \sum_{p=1}^{3} C_{1,p}|\pi_n') \\
&\cdots \\
\sum_{p=2}^{m+1} D_p|\pi_n' &\leq \frac{c}{1-c}(C_{m,m}|\pi_n' + \sum_{p=m-1}^{m} C_{m-1,p}|\pi_n' + \ldots + \sum_{p=1}^{m} C_{1,p}|\pi_n').
\end{aligned}$$

If $C_{1,1}|\pi'_n < C_{1,1}|\pi_n$, then $D_2|\pi'_n < D_2|\pi_n$, then $C_{2,2}|\pi'_n < C_{2,2}|\pi_n$, also we know $C_{1,2}|\pi'_n + C_{1,1}|\pi'_n \leq C_{1,1}|\pi_n$, then $D_2|\pi'_n + D_3|\pi'_n < D_2|\pi_n + D_3|\pi_n$, then $C_{3,3}|\pi'_n < C_{3,3}|\pi_n, \cdots$ following a mathematical induction we have $\sum_{p=2}^{m+1} D_p|\pi'_n < \sum_{p=2}^{m+1} D_p|\pi_n$, as $m \leq n-1$, $\sum_{p=2}^{m+1} D_p|\pi'_n \leq \sum_{p=2}^{n} D_p|\pi_n$.

By the proof of Theorem 1, step 3, under any information policy, the minimum discredit a state θ that persists charges is $1-\theta$. Thus, fix $\sum_{p=2}^{n} D_p|\pi_n$, $\cup_{p=1}^{n} T_n|\pi_n$ uniquely maximizes the information designer's ex ante probability of persistence. As $\sum_{p=2}^{m+1} D_p|\pi'_n < \sum_{p=2}^{n} D_p|\pi_n$, the information designer's ex ante probability of persistence under π'_n is strictly smaller than under π_n, and we reach a contradiction. Thus, in every optimal design, $C_{1,1}|\pi'_n = C_{1,1}|\pi_n$, $D_2|\pi'_n = D_2|\pi_n$; then we also have $C_{1,p}|\pi'_n = 0$ for $p = 2, 3, \cdots, m$.

Similarly, given $C_{1,1}|\pi'_n = C_{1,1}|\pi_n$, $D_2|\pi'_n = D_2|\pi_n$, suppose that $C_{2,2}|\pi'_n < C_{2,2}|\pi_n$, then $D_3|\pi'_n < D_3|\pi_n$, then $C_{3,3}|\pi'_n < C_{3,3}|\pi_n$, also we know $C_{2,3}|\pi'_n + C_{2,2}|\pi'_n \leq C_{2,2}|\pi_n$, then $D_3|\pi'_n + D_4|\pi'_n < D_3|\pi_n + D_4|\pi_n$, then $C_{4,4}|\pi'_n < C_{4,4}|\pi_n, \cdots$ following a mathematical induction we have $\sum_{p=3}^{m+1} D_p|\pi'_n < \sum_{p=3}^{m+1} D_p|\pi_n \leq \sum_{p=3}^{n} D_p|\pi_n$. Note that we already have $D_2|\pi'_n = D_2|\pi_n$, thus we have $\sum_{p=2}^{m+1} D_p|\pi'_n < \sum_{p=2}^{n} D_p|\pi_n$, by the same argument as above, the information designer's ex ante probability of persistence under π'_n is strictly smaller than under π_n, and we reach a contradiction. Thus, in every optimal design, $C_{2,2}|\pi'_n = C_{2,2}|\pi_n$, $D_3|\pi'_n = D_3|\pi_n$, then we also have $C_{2,p}|\pi'_n = 0$ for $p = 3, 4, \cdots, m$.

Iterate the above process, by a mathematical induction, we conclude that in every optimal design, for $p = 1, 2, \cdots, m$, $C_{p,p}|\pi'_n = C_{p,p}|\pi_n$, $D_{p+1}|\pi'_n = D_{p+1}|\pi_n$, and $C_{p,q}|\pi'_n = C_{p,q}|\pi_n = 0$ for $q = p+1, p+2, \cdots, m$.

By the above analysis, in every optimal design π'_n, states in $T_1|\pi'_n$ send signals in $S_1|\pi'_n$ with probability 100%, and states in $\Theta \backslash (T_1|\pi'_n \cup T_2|\pi'_n)$ should not send any signal in $S_1|\pi'_n$ with positive probability. We show that if S_1 contains more than one element (denoted by s_1), it must not be optimum. First, we construct information policy π''_n; under this information policy, the states in $\Theta \backslash (T_1|\pi'_n \cup T_2|\pi'_n)$ behave the same as under π'_n, the states in $T_1|\pi'_n \cup T_2|\pi'_n$ send s_1 whenever they should send a signal in S_1 under π'_n. It's straight forward that π''_n uses at least one signal less than π'_n does, and the outcome is the same as π'_n. Then we can construct information policy π'''_n that improves upon π''_n by using one more signal, denoted by s'. Under π'''_n, let the states in T_n send s' whenever they should send a signal not in S_{n-1} under

π_n'', then let a sufficiently small state set in $\Theta\backslash(\cup_{p=1}^m T_p|\pi_n'')$ send s' with probability 100%, the other states behave the same as under π_n''. All the states that persist under π_n'' still persist under π_n'''; and the small state set that sends s' with probability 100% now persists. Thus the ex ante probability of persistence under π_n''' is strictly larger than under π_n''; this contradicts our assumption that π_n' is optimum. Thus, states in $T_1|\pi_n'$ send s_1 with probability 100%, and a state in $T_2|\pi_n'$ sends s_1 with probability that equals to its strength.

By similar arguments, in every optimal design, a state in $T_2|\pi_n'$ sends one single signal, s_2, with probability that equals to one minus its strength. This iteration proceeds, and we show that in every optimal design, $\pi_n' = \pi_n$. □

Proof of Proposition 3.B. The first statement is straightforward.

To prove the second statement, rewrite (3) for F and G to get

$$c(1-F(\theta^*)) = \int_{\theta^*}^1 (F(\theta)-F(\theta^*))d\theta$$
$$c(1-G(\theta^{**})) = \int_{\theta^{**}}^1 (G(\theta)-G(\theta^{**}))d\theta.$$

Consider θ' such that $G(\theta') = F(\theta^*)$ which implies that $\theta' \geq \theta^*$ by first-order stochastic dominance. As $G(\theta) \leq F(\theta)$ for all θ, we know that $\int_{\theta'}^1 (G(\theta)-G(\theta'))d\theta \leq \int_{\theta^*}^1 (F(\theta)-F(\theta^*))d\theta$, i.e. $c(1-G(\theta')) \geq \int_{\theta'}^1 (G(\theta)-G(\theta'))d\theta$. As the left-hand side of (3) must be negative for all $\theta < \theta^{**}$ and positive for all $\theta > \theta^{**}$, it must be that $\theta^{**} \leq \theta'$. Therefore $1-G(\theta^{**}) \geq 1-G(\theta') = 1-F(\theta^*)$.

To prove the third statement, reconsider (3). When $\int_1^{\bar{\theta}} f_n(\theta)d\theta$ goes to zero, (3) becomes

$$\liminf_{n\to\infty} \theta_n^* = \inf\left\{\theta' \in \Theta : \liminf_{n\to\infty} \frac{\int_{\theta'}^1 \theta f_n(\theta)d\theta}{\int_{\theta'}^1 (1-\theta)f_n(\theta)d\theta} \geq \frac{1-c}{c}\right\}.$$

Note that if $\theta > 1-c$, we have $\liminf_{n\to\infty} \theta f_n(\theta) > \liminf_{n\to\infty}(1-\theta)f_n(\theta)$, which implies $\liminf_{n\to\infty} \int_\theta^1 \theta f_n(\theta)d\theta > \liminf_{n\to\infty} \int_\theta^1 (1-\theta)f_n(\theta)d\theta$. Therefore, as far as the measure of $\theta \in [1-c,1]$ is bounded away from 0, there exists ϵ sufficiently small such that $\liminf_{n\to\infty} \frac{\int_{1-c}^1 \theta f_n(\theta)d\theta}{\int_{\theta'}^1 (1-\theta)f_n(\theta)d\theta} > 1-c+\epsilon$. To get the inequality balanced, we need $\theta' < 1-c$ and eventually we have $\liminf_{n\to\infty} \theta_n^* < 1-c$ as well. Then we have $\liminf_{n\to\infty} 1-F_n(\theta_n^*) > 0$.

The above criterion is satisfied by $f(\theta) > 0$ for all $\theta \in \hat{\Theta}$, for some non-empty $\hat{\Theta} \subset [1-c, 1]$. The result thus follows. □

Proof of Proposition 4. We first consider increasing c. Note that $\theta^{\dagger}$ and θ^* are characterized by

$$c(F(\bar{\theta}) - F(\theta^{\dagger})) - (F(1) - F(\theta^{\dagger})) = 0 \tag{11}$$

$$c(F(\bar{\theta}) - F(\theta^*)) - (F(1) - F(\theta^*)) + \int_{\theta^*}^{1} \theta f(\theta) d\theta = 0, \tag{12}$$

where (12) is a representation of (3). (11)-(12) gives

$$\begin{aligned}
(1-c)(F(\theta^{\dagger}) - F(\theta^*)) &= \int_{\theta^*}^{1} \theta f(\theta) d\theta \\
(1-c)(F(\theta^{\dagger}) - F(\theta^*)) &= \int_{\theta^*}^{1} \theta f(\theta) d\theta \\
F(\theta^{\dagger}) - F(\theta^*) &= \frac{\int_{\theta^*}^{1} \theta f(\theta) d\theta}{1-c}
\end{aligned}$$

It is clear that θ^* decreases as c increases. Hence $F(\theta^{\dagger}) - F(\theta^*)$ increases as c increases.

As $F(1) \to 1$, $F(\theta^{\dagger}) \to 1$. Then if $F(\theta^*) \to 1$ we have $\int_{\theta^*}^{1} \theta f(\theta) d\theta \to 1$. Then we require $0 = \frac{1}{1-c}$, contradiction. Thus θ^* is bounded away from 1. □

B Proofs of Robustness (For Online Publication)

Proof of Corollary 1. Fix any information policy $\pi(\cdot|\theta)$; we define a series $T_{(\cdot)}$ on the state space and, for every i, a series $S^i_{(\cdot)}$ on the signal space. The idea is similar to the proof of Proposition 1 and the proof of Theorem 1, with minor modifications as follows.

The difference between this specification and Theorem 1 is that now, for each state, the ex post distribution of signals is not determinate; instead, it can be any distribution over all possible ex post distributions. Nevertheless, we show that this added degree of freedom does not increase the maximum of the status quo's probability of persistence, i.e. the original design remains optimal.

For every i, define $S^i_0 = \emptyset$ and $a^0_i(s) \equiv 1$. Define $S^i_1 \subseteq S$ as the set of states satisfying the following condition:

$$\int_\Theta \frac{f(\theta')\pi(s|\theta')}{\int_\Theta f(\theta'')\pi(s|\theta'')d\theta''} \Pr(\theta' < \int_{[0,1]} a_j^0(s_j)dj|\theta = \theta', s_i = s)d\theta' \leq c.$$

Define

$$a_i^1(s) = \begin{cases} 0 \text{ if } s \in S_1^i \\ 1 \text{ otherwise.} \end{cases}$$

We then perform an iteration similar to the proof of Proposition 1. For every i and for $k = 2, 3, ..., +\infty$, we define a_i^k, S_k^i, S^{i*}. We omit the proof that $\{S^{i*}\}_{i\in[0,1]}$ characterizes the unique agent equilibrium.

To save notations and to simplify our discussion, we modify π without changing the outcome of the equlibrium. Notice that for every bijective mapping A from S to S and every $s \in S$, let π' send $A(s)$ to agent i whenever π sends s, then the agent i's action when receiving $A(s)$ under policy π' is the same as his action when receiving s under policy π. Then, for every $\pi(\cdot|\theta)$, we can always find $\pi'(\cdot|\theta)$ such that $\bar{a}_i^k(\cdot) \equiv \bar{a}_j^k(\cdot)$ for every k and every $i, j \in [0,1]$, and the status quo's ex ante probability of persistence under $\pi'(\cdot|\theta)$ is identically equal to its ex ante probability of persistence under $\pi(\cdot|\theta)$.

Without loss of generality, from here on, we focus on policies under which $\bar{a}_i^k(\cdot) \equiv \bar{a}_j^k(\cdot)$ for every n and every $i, j \in [0,1]$, then $S_k^i = S_k^j$ for every k and every $i, j \in [0,1]$. To save notations, we still call this policy $\pi(\cdot|\theta)$, and define $\bar{a}^k(\cdot) \equiv \bar{a}_i^k(\cdot)$, $S_k = S_k^i$, for arbitrary i and for every k.

Next, define type set $T_0 = \emptyset$ and function $f_0(\theta)$ on Θ, $f_0(\theta) = 0$ for every $\theta \in \Theta$.

Define type set $T_1 = (1, \bar{\theta}]$ and function $f_1(\theta)$ on Θ, $f_1(\theta) = f(\theta)$ for every $\theta \in T_1$ and $f_1(\theta) = 0$ elsewhere.

Define type set T_2 as every state $x \in \Theta$ such that: $\Pr(\int_{i\in[0,1]} \mathbb{1}(s_i \in S_1^i|\theta = x)di < x) > 0$; define function $f_2(\theta)$ on Θ, for every θ, $f_2'(\theta) = f(\theta)\Pr(\int_{i\in[0,1]} \mathbb{1}(s_i \in S_1^i|\theta = x)di < x)$.

Define T_k and $f_k(\theta)$ for $k = 3, 4, ...$ similarly.

Next, we identify an upper bound of the ex ante probability that the regime persists.

Fix any information policy $\pi(\cdot|\theta)$ that induces a unique agent equilibrium which satisfies the above condition, define function $T(\theta)$ the same as in the proof of Theorem

1.

For $i \in [0,1], k = 1,2,...,$ define "discredit D_k":

$$D_k^i = \int_\Theta [f_k(\theta) - f_{k-1}(\theta)] \Pr(s_i \in S_{k-1}|\theta) d\theta.$$

For $i \in [0,1], k = 1,2,..., p = k+1, k+2,...,$ define "credit $C_{k,p}^i$":

$$C_{k,p}^i = \int_\Theta [f_k(\theta) - f_{k-1}(\theta)] \Pr(s_i \in S_p \backslash S_{p-1}|\theta) d\theta.$$

By the definition of $T_{(\cdot)}$ and $S_{(\cdot)}$, for every k, i, $c \sum_{m=1}^{k} \sum_{p=m}^{k} C_{m,p}^i \geq (1-c) \sum_{m=1}^{k+1} D_m^i$. Also note that for the status quo to persist it must send signals in S_k to a population greater or equal to $1-\theta$. Thus, similar to the proof of Theorem 1, for every i, θ

$$\begin{aligned} c \int_\Theta f_k(\theta) \Pr(s_i &\in S_k \backslash S_{T(\theta)-1}|\theta) d\theta \\ &\geq (1-c) \int_\Theta f_{k+1}(\theta) \Pr(s_i \in S_{T(\theta)-1}) d\theta \end{aligned}$$

Thus we have

$$\begin{aligned} & c \int_{[0,1]} \int_\Theta f_k(\theta) \Pr(s_i \in S_k \backslash S_{T(\theta)-1}|\theta) d\theta di \\ \geq \;& (1-c) \int_{[0,1]} \int_\Theta f_{k+1}(\theta) \Pr(s_i \in S_{T(\theta)-1}|\theta) d\theta di \\ \Rightarrow \;& c(\int_\Theta f_1(\theta) d\theta + \int_\Theta (f_k(\theta) - f_1(\theta))\theta f(\theta) d\theta) \geq (1-c) \int_\Theta (f_{k+1}(\theta) - f_1(\theta))(1-\theta) d\theta \end{aligned}$$

Note that Lemma 4 remains valid; the policy maker can still construct an information policy that weakly improves upon π by saving the high types. Thus in any optimum, $f_k(\theta) = f(\theta)$ for $\theta \in T_k$ and $f_k(\theta) = 0$ elsewhere. Following the proof of Theorem 1, θ^* satisfies

$$c(\int_1^{\bar{\theta}} f(\theta) d\theta + \int_{\theta^*}^1 \theta f(\theta) d\theta) = (1-c) \int_{\theta^*}^1 (1-\theta) f(\theta) d\theta$$

which is identical to the proof of Theorem 1. Notice that the last condition is irrelevant to agent identity i. Thus from here on, the rest of the proof follows the proof of Theorem 1. □

References

Acemoglu, D. and J. A. Robinson (2005). *Economic origins of dictatorship and democracy*. Cambridge University Press.

Alonso, R. and O. Câmara (2016). Persuading voters. *American Economic Review 106*(11), 3590–3605.

Bardhi, A. and Y. Guo (2018). Modes of persuasion toward unanimous consent. *Theoretical Economics 13*(3), 1111–1149.

Basak, D. and Z. Zhou (2018). Timely persuasion. working paper.

Bergemann, D. and S. Morris (2016). Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics 11*(2), 487–522.

Bergemann, D. and S. Morris (2019). Information design: A unified perspective. *Journal of Economic Literature 57*(1), 44–95.

Best, J. and D. Quigley (2017). Persuasion for the long run. working paper.

Chan, J., S. Gupta, F. Li, and Y. Wang (2019). Pivotal persuasion. *Journal of Economic Theory 180*, 178–202.

Edmond, C. (2013). Information manipulation, coordination, and regime change. *Review of Economic Studies 80*(4), 1422–1458.

Galperti, S. and J. Perego (2019). Belief meddling in social networks: An information-design approach. working paper.

Goldstein, I. and C. Huang (2016). Bayesian persuasion in coordination games. *American Economic Review: Papers & Proceedings 106*(5), 592–596.

Goldstein, I. and C. Huang (2018). Credit rating inflation and firms' investments. working paper.

Guo, Y. and E. Shmaya (2019). Costly miscalibration. working paper.

Guriev, S. and D. Treisman (2019). Informational autocrats. *Journal of Economic Perspectives 33*(4), 100–127.

Halac, M., E. Lipnowski, and D. Rappoport (2020). Rank uncertainty in organizations. *Available at SSRN 3553935*.

Hébert, B. and M. Woodford (2017). Rational inattention and sequential information sampling. working paper.

Hoshino, T. (2019). Multi-anent persuasion: Leveraging strategic uncertainty. working paper.

Inostroza, N. and A. Pavan (2018). Persuasion in global games with application to stress testing. working paper.

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

King, G., J. Pan, and M. E. Roberts (2017). How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review 111*(3), 484–501.

Levy, G. and R. Razin (2018). Information diffusion in networks with the bayesian peer influence heuristic. *Games and Economic Behavior 109*, 262–270.

Li, W. and X. Tan (2019). Locally bayesian learning in networks. *Theoretical Economics* (forthcoming).

Lipnowski, E. and E. Sadler (2019). Peer-confirming equilibrium. *Econometrica 87*(2), 567–591.

Mathevet, L., D. Pearce, and E. Stacchetti (2018). Reputation and information design. working paper.

Mathevet, L., J. Perego, and I. Taneva (2019). On information design in games. *Journal of Political Economy* (forthcoming).

Mathevet, L. and I. Taneva (2020). Organized information transmission. *Available at SSRN 3656555*.

Morris, S., D. Oyama, and S. Takahashi (2019). Adversarial information design in binary-action supermodular games. working paper.

Morris, S. and H. S. Shin (1998). Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review 88*(3), 587–597.

Morris, S. and H. S. Shin (2003). Global games: Theory and applications. In M. Dewatripont, L. P. Hansen, and S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Volume 1, Cambridge. Cambridge University Press.

Morris, S. and M. Yang (2019). Coordination and continuous stochastic choice. Technical report, Duke University.

Olson, M. (1965). *The logic of collective action*, Volume 124. Harvard University Press.

Ong, J. C. and J. Cabanes (2018). Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the philippines. *Newton Tech4Dev Network*.

Pomatto, L., P. Strack, and O. Tamuz (2018). The cost of information. working paper.

Taneva, I. (2019). Information design. *American Economic Journal: Microeconomics* (forthcoming).