Global Manipulation by Local Obfuscation*

Fei Li† Yangbo Song‡ Mofei Zhao§

August 8, 2021

Abstract

We study adversarial information design in a regime-change context. A continuum of agents simultaneously chooses whether to attack the current regime. The attack succeeds 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. Our optimal information structure exhibits local obfuscation: some agents receive a signal matching the regime's true strength, and others receive an elevated signal professing slightly higher strength. This policy is the unique limit of finite-signal problems. Public signals are strictly suboptimal, and in some cases where public signals become futile, local obfuscation guarantees the collapse of agents' coordination.

Keywords: Coordination, information design, regime-change game

JEL Classification: C7, D7, D8.

*We thank Yu Awaya, Arjada Bardhi, Gary Biglaiser, Daniel Bernhardt, James Best, Odilon Camara, 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, Christopher Sandmann, Mehdi Shadmehr, Xianwen Shi, Joel Sobel, Satoru Takahashi, Ina Taneva, Can Tian, Kyle Woodward, Xi Weng, Ming Yang, Jidong Zhou, Zhen Zhou, and participants at various conferences and seminars for comments. Mofei Zhao's work on this project began when he was affiliated with Capital University of Economics and Business.

†Department of Economics, University of North Carolina, Chapel Hill, NC 27599, United States. Email: lifei@email.unc.edu.

‡School of Management and Economics, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China. Email: yangbosong@cuhk.edu.cn.

§School of Economics and Management, Beihang University, Beijing 100083, China. Email: zhaomf.06@gmail.com.

1 Introduction

Many economic problems with strategic complementaries are modeled as regime-change games where a status quo is overturned if a sufficiently large number of agents attack it. Examples include speculation against a pegged currency, run against a bank, and revolution against an authoritarian government.[1] In these settings, a central element determining the agents' coordination outcome is the information structure. Therefore, a regime's defender will pursue information manipulation to collapse coordination and preserve the status quo to the largest possible extent. Depending on contexts, the regime's tool ranges from monetary policy (Angeletos et al. 2006), to stress testing (Inostroza and Pavan 2018), and to media outlets capture (Edmond 2013).

In this paper, we study an information design problem in a regime-change context. The designer can commit to any information policy of his choice. This framework unveils the fundamental trade-off of information manipulation and depicts the maximum value achievable in regime-change games. Our work makes two contributions. First, we characterize a simple optimal information policy in closed form. The characterization provides a benchmark to assess the role of numerous application-specific constraints in shaping optimal information policies.[2] Second, we show that among possibly multiple policies to attain the unconstrained optimum, this policy is the limit of the unique solution to bounded-depth problems where only *finite* levels of agents' strategic reasoning are up for manipulation. This exercise sheds light on the impact of realistic constraint, e.g., limitation of agents' cognitive abilities and complexity of signals, and proposes a selection criterion for the unconstrained problem.

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 is rewarded 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 *vincible*. The information designer commits to a state-dependent information policy that sends a signal, which can be public

[1] See Morris and Shin (2003) and Angeletos and Lian (2016) for two survey papers.

[2] In applications, information manipulation is often subject to various restrictions. For example, it may be unlawful to release differential information to different audiences. Another well-known example is that audiences may have the access to other information sources which are out of the reach of the information designer. See more discussion in section 1.1.

or private, to each agent. His objective is to maximize the probability of preserving the regime in his *least-preferred (adversarial) equilibrium*.

Adversarial information design captures the idea of robustness in the information design, but also poses a challenge: it breaks the applicability of the standard Bayes correlated equilibrium (BCE) method (Bergemann and Morris (2016) and Taneva (2019)), which implicitly selects the designer's favorite equilibrium. The key to constructing an optimal information policy is to recognize that regime-change games are supermodular. In these games, Milgrom and Roberts (1990) show that, under each information structure, the adversarial (or smallest/lowest) equilibrium can be obtained by iterative elimination of strictly dominated strategies (IESDS). Consequently, adversarial information design in supermodular games can be treated as endogenzing the process of IESDS. As far as we know, this conceptual connection is first formally pointed out by Bergemann and Morris (2019), and has inspired several recent studies in other applications (see subsection 1.1 for a detailed discussion).

We explore this conceptual connection in regime-change settings to study adversarial information design as managing an endogenous IESDS process. It enables us to take advantage of a potentially infinite chain of state obfuscation. As a natural starting point, consider public persuasion where all agents are always sent identical signals. In some vincible states, the designer sends all agents the same signal as in invincible states, so that agents 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 to manipulate information more subtly when not constrained to sending public signals: a state does not have to be truly invincible to be leveraged on, but only needs to convince agents of no sufficient coordination. In this spirit, the designer may leverage not only 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 never face sufficiently coordinated attacks, they, in turn, become a "conditionally invincible" tier and may be leveraged on to save more vincible states, and so on. The linkages between these tiers are endogenously characterized by the designer's information policy, so our 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.

Local Obfuscation. The optimal information structure we identify 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. The measure of each such proportion is deterministic; thus, although each agent receiving s_k remains uncertain about the true state, the measure of each signal sent conditional on states is fixed. Finally, the weakest tier, characterized by an endogenously determined threshold, always sends a self-identifying signal s_a. In other words, except for the invincible and the weakest states, the information designer under each state executes a *local obfuscating* policy, essentially revealing the actual tier to some agents but deceiving other agents by a slightly stronger tier. Heuristically, 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. At the optimum, the regime-change outcome is characterized by a cutoff value of the regime's strength, i.e. the strongest state sending signal s_a. While each agent's uncertainty about the regime's strength is limited, their higher-order uncertainty (regarding the regime's state) remains. In fact, the only common knowledge among all agents is whether the regime's strength is above the cutoff. The remaining higher-order uncertainty makes agents' actions perfectly coordinated: An agent attacks the regime if and only if the regime strength is below the cutoff. In this case, the status quo fails.

Our construction can be viewed as an endogenous design of a series of "state infection" which is an analogy proposed by Rubinstein (1989). We further show that to attain optimum, it suffices at every round of the process to extend the contagion to only the strongest currently uncontaminated states, up to a recursive belief updating constraint. As a corollary, limitation on the designer's degree of freedom in controlling information – such as confinement to public signals or exogenous information – is the main potential source of sub-optimality compared to the unconstrained optimum. Notably, the term "local" indicates the adjacency between tiers of states, which differs qualitatively from its description of small signal perturbation in the classical email game (Rubinstein (1989)) or standard global games (Carlsson and van Damme (1993), Morris and Shin (2003)). Indeed, as we will discuss later, this seemingly minor difference substantially distinguishes our construction from some recent information design work building on a similar "small-noise signal" infection argument.

The optimality of the *monotonic* relationship between the regime-change outcome and the regime strength is a natural consequence of state monotonicity and strategic complementarity of regime-change games (Frankel et al. 2003) and echoes the intuitive constructions in previous studies (e.g., Goldstein and Huang 2016) and the equilibrium outcome under optimal public disclosure in some circumstances (see Inostroza and Pavan 2018).

Our optimal information policy establishes a *deterministic* mapping between states and regime-change outcomes, which differs from most canonical information design papers. Consequently, what matters to an agent is to predict the regime's state given her received signal. All the higher-order uncertainty is averaged out, simplifying the application of our model for further studies (e.g., policy intervention and the effect of various informational constraints). A similar property emerges in the global regime-change games due to the assumption that private signals are identical and independent distributed (i.i.d.) across agents. However, in our model, the property is an implication under an *optimal* information structure in which signal independence is not a priori. This echoes with Inostroza and Pavan (2018), who show the optimality of deterministic pass/fail tests when the designer can only make a public disclosure.

Level-K Obfuscation. We then investigate how the *manipulable reasoning depth* of agents determines the implementable outcome of an optimal information design. A practical way to evaluate bounded depths of manipulable reasoning is to impose the assumption of a finite signal space of information structures. Notice that this "level-K" obfuscation exercise substantially differs from the level-K thinking in the behavioral economics literature (see, e.g., Alaoui and Penta (2016), De Clippel et al. (2019) and Crawford (2021)). For each information structure, we still look for Nash equilibria of the corresponding Bayesian regime-change game. That is, agents are capable of conducting infinitely higher-order strategic reasoning, but only the first K levels are up for manipulation. Hence, it captures (i) the designer's inability of flexible information control due to either signal design cost or communication obstacle, or (ii) the agents' bounded cognitive ability to comprehend/distinguish infinite signals.

We fully characterize the *unique* optimal information policy in closed form when information design is constrained by the agents' level of reasoning up for manipulation. This differs from the benchmark unconstrained model that has multiple solutions (we discuss some of these policies in the analysis; also see the discussion regarding the implementation in Morris et al. (2019)).[3] In particular, when the designer

[3]We are also aware of a global-game like optimal policy in regime-change-games using the idea of Morris et al. (2019) from a private communication with Stephen Morris, Daisuke Oyama and Satoru

is only capable of manipulating agents' higher-order reasoning up to a finite level-K, a locally obfuscating policy producing $K+1$ tiers in total is the unique optimal information structure. Thus, our result highlights the designer's advantage resulting from manipulating agents' higher-level reasoning and explicitly identifies the magnitude of this advantage as agents' depth of reasoning improves. Manipulating higher levels of reasoning benefits the information designer by creating more "conditionally invincible" states. It also indicates a one-to-one relation between the depth of manipulable reasoning and signal complexity: the level-K locally obfuscating policy achieves the designer's unique optimum when agents are fully rational, but the designer has only $K+1$ distinct signals at his disposal. The result holds for an arbitrary K, making it a natural selection criterion that uniquely identifies our local obfuscating policy among policies that may achieve the designer's optimal outcome. As a practical implication, the designer should always adopt local obfuscation when constrained by agents' manipulable reasoning depth or signal availability.

We discover that the optimal level-K obfuscation could lead to *coordination failure* among agents. To maximize the set of infected states using finite signals, it is optimal when the true state is slightly above the margin of collapse, to make attack conditionally dominated for only a fraction of agents. This is in sharp contrast to the results in our benchmark model as well as the literature of adversarial information design in supermodular games (see subsection 1.1).

Local obfuscation has a unique advantage over public information structures, which manipulate agents' reasoning only up to the first level. We demonstrate this advantage in two distinct ways. First, 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. The difference between the cost thresholds coincides with the conditionally expected strength of the persisting states below one. Second, when the measure of invincible states converges to zero, an 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, making the policy ex post optimal. It highlights the power of manipulating higher-order uncertainty: it is more likely to relieve us from the time-inconsistent commitment concern.

Implementation. Our framework allows the information designer the maximum de-

Takahashi.

gree of freedom to choose from all information policies, including ones with correlated and/or non-anonymous signals. Nevertheless, the optimal policy only requires a very simple implementation. It is essentially an anonymous and uncorrelated signal whose realization is at most binary given each possible state. In practical scenarios, the policy can be understood as either (i) i.i.d. private signals, (ii) one random signal with each realization covering a predetermined measure of randomly selected agents, or (iii) a combination between public signal and private endorsement *a la* Alonso and Câmara (2016).[4] The implementation simplicity sheds light on information manipulation practices in the digital era. For instance, recent studies (e.g., Guriev and Treisman 2019) have shown that a growing number of autocrats adopt information repression strategy, instead of terrorizing citizens in old-and-bloody style, to stabilize their governance. Our result begins a formal investigation that (i) unpacks the secret of such information repression and (ii) helps to understand this trend. To collapse citizens' hostile collective coordination, all it takes is to create minor belief uncertainty among citizens by dividing message recipients. Due to the penetration of social networks, this less bloody and more effective repression becomes increasingly appealing to autocrats.[5]

1.1 Related Literature

Information Manipulation in Global Games. This paper contributes to the large 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, 2019) through public information disclosures and Angeletos et al. (2006, 2007), Edmond (2013), Huang (2017), and Cong et al. (2019) through policy intervention context). On the front of information design context, our paper is mostly related to Inostroza and Pavan (2018) which is the first paper studying adversarial information design in global games. With attention to applications such as stress testing, they make realistic modeling choices: agents are endowed with exogenous private information, and the designer is allowed only to choose a public disclosure. They show an optimal information structure takes the form of a pass/fail test and derive conditions under which the optimal test is monotone in the regime's state. On the contrary, we assume that the information designer is the only source of information and possesses

[4] In fact, as will be clear in the subsequent analysis, exogenous complexity (such as Gaussian private signals) or mandatory correlation (such as public signals) may result in sub-optimality for the designer.

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

a maximum degree of freedom to choose from all information policies to isolate and highlight the fundamental trade-off of manipulating information in regime-change settings. 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 game, as well as higher-order uncertainty of his opponents. However, compared to familiar Gaussian information structure that encompasses the entire class of games in an agent's (first-order) belief, we show that it is sufficient to maintain 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.

Adversarial information design. In general, adversarial information design in games inevitably involves higher-order belief manipulation, which complicates the analysis. Mathevet et al. (2020) first point out the conceptual connection between adversarial information design and concavification on the space of belief hierarchy. Hoshino (2019) shows that, using the leverage of strategic uncertainty, agents can be persuaded to take an action profile that satisfies a generalization of risk dominance given any non-degenerate prior. Bergemann and Morris (2019) further discuss the connection between adversarial information design and the literature on higher-order beliefs. The tractability of our analysis results from the fact that adversarial information design in supermodular games is equivalent to manipulating the process of iterative elimination of dominated strategies. We are certainly not the first to notice this conceptual connection. The idea of deterring coordination by creating non-common knowledge dates back to the classical email game by Rubinstein (1989), and also underlies monotone equilibrium behavior in global games (Carlsson and van Damme (1993), Morris and Shin (2003)) although these clever constructions are never meant for optimal information design. In an insightful example, Bergemann and Morris (2019) point out that adversarial information design can be achieved by an email-game-like construction of information structure. Several companion recent studies, e.g., Halac et al. (2021), Moriya and Yamashita (2020), and Sandmann (2020), explore this property in various of finite supermodular games. The connection also plays a substantial role in deriving optimal public disclosure in Inostroza and Pavan (2018).

Contemporaneously with our paper, Morris et al. (2019) propose a tractable method to characterize the set of implementable adversarial equilibrium outcomes in all binary action supermodular games, including regime change games. Our paper's novelty is that, by taking advantage of some unique features of regime-change games

(such as binary regime status and continuum of agents), we can derive a simple and intuitive optimal information structure in close form. It is the unique limit of finite signal problems and establishes a *deterministic* mapping between states and regime-change outcomes.

A distinguishing feature of our construction is its robustness to the perturbation of manipulating finite-depth reasoning. With the luxury of manipulating infinite levels of agents' reasoning in the unconstrained problem, the designer can afford inefficient infection in finite steps of IESDS. Therefore, there are multiple optimal policies as we illustrate in section 3.2. Also see the implementation discussion of Morris et al. (2019) which is featured with small signal noise/belief uncertainty as in Rubinstein (1989) and Carlsson and van Damme (1993). On the contrary, our information policy possesses a greedy algorithm-like feature that maximizes the measure of infected states in each step of IESDS.This logic is crucial for our information policy being the unique solution to the bounded-depth problem. Also, the unique optimal information policy will lead to imperfect coordination among agents, which is in sharp contrast to the literature of adversarial information design in supermodular games (see, e.g., Morris et al. (2019) and Inostroza and Pavan (2018)). The difference relies on the designer's ability to precisely control the first K steps of state infection in IESDS endowed by the feature of regime-change games and the flexibility to control information asymmetry among agents.

Information design with multiple receivers. 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), Heese and Lauermann (2020)), and social network (Galperti and Perego (2019) and Candogan and Drakopoulos (2020)), etc. In this literature, a receiver often faces uncertainty about the set of opponents receiving signals identical to him/her, similar to our paper. However, in these papers, 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 we identify 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 "spreading the words" to a selected group of receivers dominates public persuasion in certain circumstances.

Organization. The rest of the paper is organized as follows. Section 2 lays out the

model. Section 3 presents main results. Section 4 concludes. Proofs are in the Appendices.

2 Model

Base game. Consider a canonical regime-change game studied by Angeletos et al. (2007). 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

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

A random variable θ represents the strength of the status quo. 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 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. 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,[6] 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. The information

[6] This assumption rules out only an uninteresting case and is not essential. When $\theta \leq 0$, it automatically collapses. There is no point in manipulating information.

designer's proposed *information structure* is a mapping from Θ to $\Delta(M(S))$, where $M(S) \subset \{S^{[0,1]}\}$ is a set of integrable functions with codomain S. That is, an information structure is a conditional probability distribution over the signal profile of agents. This configuration allows arbitrary correlation among signals and non-anonymous information structures that send different signal distributions to different agents. Without loss of generality, we focus on the class of distributions where the density is almost everywhere well-defined and integrable, and thereby restrict our attention to policies under which the regime outcome is measurable in the designer's information. We use $\pi_i(s|\theta)$ to denote the probability of agent i receiving signal $s \in S$. In an *anonymous* information structure, i.e. when $\pi_i(s|\theta) = \pi_j(s|\theta)\ \forall i, j \in [0,1]$, we omit the subscript and simply refer to the probability as $\pi(s|\theta)$.

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, each agent i receives signal $s \in S$ according to $\pi_i(s|\theta)$, 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.[7]

For a given information structure, there may be multiplicity. We solve for the information designer's *worst* Bayesian Nash equilibrium to capture the idea of adversarial/robust information design.[8] 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 an information structure to induce an adversarial Bayesian Nash equilibrium which maximizes the regime's expected probability of persistence.

3 Analysis

We begin with equilibrium characterization for an arbitrary information structure.

[7]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.

[8]The implementation based on the designer's favorite equilibrium is trivial. Since $\theta \geq 0$, the designer can disclose nothing, and there is an equilibrium where no agent attacks.

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

The regime-change game is supermodular. Given an information structure, the adversarial (also known as lowest/smallest) equilibrium can be established by the familiar argument of iterated elimination of strictly dominated strategies (IESDS) as in Milgrom and Roberts (1990) and Frankel et al. (2003). We relegate the proof to online Appendix B and provide the intuition using an anonymous information structure for expositional convenience. Beginning 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 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 complements, 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 $S \setminus S^*$. To see the uniqueness, suppose two distinct (adversarial) 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 aggressively 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 Local Obfuscation

One of our main results is the characterization of an optimal information structure, which maximizes the probability of the status quo's persistence. This information policy will be the benchmark to be compared to various sub-optimal policies in the rest of the paper. In what follows, we introduce a class of information structures.

First, notice that an information structure stochastically specifies a distribution over agents receiving signal $s \in S$ for each state.[9] An information structure is *deter-*

[9]For instance, consider an anonymous information structure that sends all agents s_1 with proba-

ministic if (i) for each θ, such a distribution is deterministically pinned down, and (ii) it is anonymous. Under a deterministic information structure, the fraction of agents who receive signal s in state θ is equal to $\pi(s|\theta)$, the probability that each agent receives signal s. It is immediate that the equilibrium regime-change outcome is fully determined by the state, the strategic uncertainty is completely averaged out, and an agent essentially faces only fundamental uncertainty. A familiar example of deterministic information structure is considered by Morris and Shin (1998): each agent independently draws a private signal from an identical and state-dependent distribution. Another example is the public pass/fail test studied by Goldstein and Huang (2016) and Inostroza and Pavan (2018), which sends a "pass" signal to every agent if the state belongs to some pre-specified set and a "fail" signal to every agent otherwise. In what follows, we show that a very simple deterministic information structure is indeed optimal.

Definition 1. *A deterministic information structure 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* $\lim_{k\to\infty} \theta_k = \theta^*$,
2. *the signal space* S *is such that* $\{s_k\}_{k=1}^{\infty} \cup \{s_a\} = S$, *and*
3. *the state-dependent signal distribution satisfies*

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

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 $\theta \geq 1$, all agents receive signal s_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, all agents receive 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

bility $p \in (0,1)$ and s_2 with probability $1-p$ under θ. The measure of agents receiving s_1 is 1 with probability p and 0 with probability $1-p$.

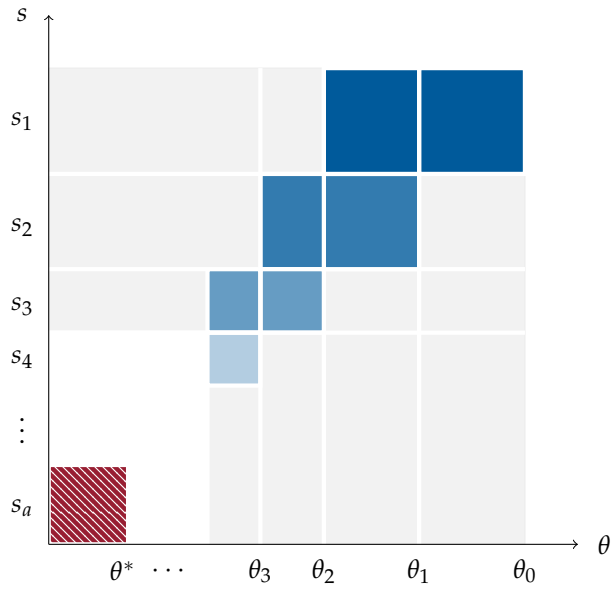

Figure 1: Illustration of local obfuscator. The horizontal axis represents states and the vertical axis represents signals and their face values. We use differential shades to distinguish information sets following different signals.

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 action profiles by manipulating the information structure.

We are now ready to present our main result.

Theorem 1. *The designer's optimum is achieved by a local obfuscator where*

1. *the state-dependent signal distribution* π^* *satisfies* $\pi^*(s_1|\theta) = 1$ *if* $\theta \in (\theta_1, \theta_0]$, *and for each* $k = 1, 2, ...$, *if* $\theta \in (\theta_{k+1}, \theta_k] \cap \Theta$,

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

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*

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

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

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

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

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 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}. They correspond to agents' binding incentive-compatibility constraints at each step of IESDS. Also, Theorem 1 proposes a simple algorithm to construct the optimal local obfuscator. Essentially, one needs only to partition the state space according to $\{\theta_k\}$ characterized by equations (1) and (2). As we will demonstrate later, summing up conditions (1) and (2) over k leads to condition (3), establishing the lowest state can persist, θ^*.

Equilibrium under optimal local obfuscator. The equilibrium analysis is similar to

the infection argument proposed by Rubinstein (1989). 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 solves equation (1), given s_1, the conditional expected benefit of attack does not exceed the cost even if *all other* agents attack. Consequently, the agent does not attack regardless of what others do. 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 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 choose not to 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 solve (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 $\{s_k\}_{k=1}^{\infty}$, and by a similar IESDS argument, no agent attacks given signal $s_k, k = 1, 2, ...$

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 the limit of sequence $\{\theta_k\}$, 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. However, the only common knowledge among agents is whether θ is above the cutoff θ^*, making the local obfuscator have a *global* impact. First, when $\theta > \theta^*$, it prevents all agents from attacking by sending disinformation to a proportion of agents only. Second, it suppresses agents' attacks 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. For better exposition, we develop a "credit-discredit" system to describe the hierarchy of endogenously induced beliefs among the agents. It depicts information manipulation as a process of alternating *states obfuscation* (Bergemann and Morris 2016) and *infection* (Rubinstein 1989). The mix differentiates our construction from Rubinstein (1989) whose construction is featured with small i.i.d.

"noise" at each step of infection.

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 obedient state obfuscation.

However, credit consumption does not stop here: for some states $\theta < 1$ 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. This is the classic states infection argument. The designer can then send the signal s_2 to these agents in state $\theta \in (\theta_2, \theta_1]$. In doing so, additional credit is produced since signal s_2 matches its face value. 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.[10]

Despite our illustration using deterministic information structure, this system applies to every (non-deterministic or even non-anonymous) 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

[10]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 rounds of IESDS, a state that will survive even under the currently most adversarial coordination possible creates credit, while another state that survives only by mimicking the former's signal creates discredit. In the next round, however, the latter state becomes one to create credit with weakened coordination among agents.

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 $(1-c)/c$. 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^*. Again, we will explain the optimality of π^* first among deterministic information structures heuristically and briefly discuss why its optimality is preserved in the most general environment. Proofs are relegated to Appendix A.

The above iterated reasoning process reveals two basic principles that the designer must abide by when seeking the optimal information structure. First, every information structure induces IESDS, possibly of infinite rounds, in the adversarial equilibrium among agents. Second, each round of IESDS features an incentive compatibility constraint, which asserts that discredit created (or equivalently, credit consumed) in the current round cannot exceed $\frac{c}{1-c}$ of net credit accumulated up to the previous round. Combining these constraints provides a *necessary* condition on what states may persist under the particular information structure: the total credit created by the persisting states must be no less than $\frac{1-c}{c}$ of the total discredit.

To further characterize an optimum, we now observe a third principle: in each round of IESDS, it always weakly benefits the designer to infect the highest not-yet-infected states, i.e., letting such states consume the existing credit for the current round and create new credit for the next round. A formal argument can be found in the proof of Lemma 2 in Appendix A. Intuitively, a state $\theta < 1$ 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); therefore a higher state always requires less credit consumption to persist while it contributes more to credit creation. Hence at least one optimal information structure must induce, through IESDS, a series of *connected* intervals of states in descending order of the states' magnitude. The set of persisting states under this information structure is then $(\theta^*, \bar{\theta}]$, which represents the union of the intervals.

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 the

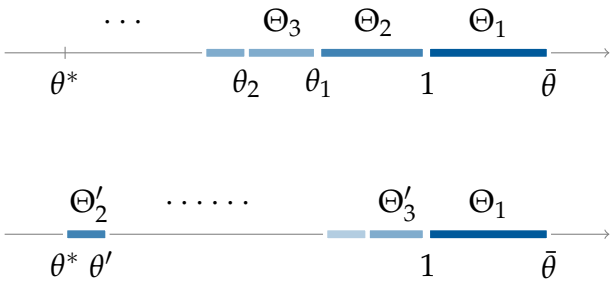

Figure 2: The upper panel corresponds to IESDS for the optimal local obfuscator. Denote $\Theta_1 = (1, \bar{\theta}]$ and Θ_k as the set of states being leveraged in the $k+1$th round of IESDS. As $k \to \infty$, every $> \theta^*$ state is leveraged. In the lower panel, the procedure is similar except 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. $\int_{\Theta_k'} \theta df(\theta) < \int_{\Theta_k} \theta df(\theta)$, making the measure of Θ_{k+1}' less than Θ_{k+1} for $k = 3, 4, 5...$

above-mentioned necessary condition:

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

This inequality, which coincides with (3) when binding, implies that the ratio between the total measures of credit and discredit created by $(\theta^*, \bar{\theta}]$ must be at least $\frac{1-c}{c}$.

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 result is in sharp contrast to resulting optimal information structures in the literature, which often generate a stochastic mapping between states and players' action profiles, but turns out to be robust in regime-change games (see Inostroza and Pavan (2018) where agents receive private information, and the designer makes public disclosure). The difference is driven by the assumptions of continuum

agents and the supermodularity of the base game. It significantly simplifies agents' strategic reasoning as in global regime-change games: although higher-order uncertainty among agents remains, what essentially matters for an agent is his belief about the fundamental state only. Note that, unlike in the global game of regime-change models (see, e.g., Morris and Shin (2003)), this property is derived as an optimal information structure rather than the assumption that signals are i.i.d. across agents.

We briefly discuss here why π^* remains optimal when general information policies are feasible and leave the formal argument to the proof. On the one hand, when the measure of realized signals given some state θ is uncertain, we can without loss of generality re-label θ as multiple replicas 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 our previous argument for the optimality of π^*. On the other hand, as agents coordinate on the information designer's least preferred equilibrium, it is only reasonable that keeping signals anonymous, i.e., introducing no additional correlation among signals, will only reduce the threat to the regime's persistence. Therefore, compared to deterministic information structures, the ability to further complicate the signals yields no extra leverage for the information designer.

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 standard revelation principle/BCE approach (Bergemann and Morris (2016) and Taneva (2019)) searches for the designer's optimal BCE, implicitly selecting his favorite equilibrium in the corresponding Bayes game. We, on the other hand, focus on the designer's worst equilibrium. Second, consider the optimal local obfuscation's outcome-equivalent binary recommendation signal (its BCE) $\pi^\dagger$ which recommends $a = 0$ to every agent if $\theta \geq \theta^*$ and $a = 1$ to every agent otherwise. While there is a Bayesian Nash equilibrium where agents follow recommendations, there is another equilibrium where agents attack regardless of recommendations. In this equilibrium, invincible states fail to infect any vincible states, and the designer's payoff is lower. Therefore, multiple (and possibly infinite) signals are necessary to preserve belief uncertainty which maximizes the status quo's survival in the designer's worst equilibrium.

3.2 Level-K Obfuscation

This section studies a natural way to extend our analysis to an environment where the designer faces a constraint in his capacity of manipulating information. In particular, suppose that the signal space is now finite and contains only $K \in \mathbb{N}^+$ distinct elements.

There are at least two practical interpretations of this setting. First, K indicates the level of agents' higher-order reasoning that can be manipulated. 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.

Second, K can also take the literal meaning of the number of available signals, which partially reflects the information structure's complexity. In practice, information design is restricted by agents' cognitive abilities to comprehend/distinguish signals, the communication capacity, and the designing cost. For a concrete example where agents have bounded cognitive ability to distinguish signals, imagine $K = 1$ corresponds to the case where agents cannot distinguish between any number of different signals, and thus the designer is essentially incapable of manipulating information; $K = 2$ implies that agents can tell, upon receiving a signal, whether the signal is some s_1 or not; and so on. The communication capacity matters because in reality, it is costly for the designer to communicate with agents about the information structure that he commits to. It is natural to assume the communication cost goes to infinity as the signal space expands.

When K is finite, we show that an optimal information structure must exhibit local obfuscation. The uniqueness result holds for every arbitrary K, making it a natural selection criterion that uniquely identifies our local obfuscating policy among policies that may achieve the designer's optimal outcome.

Theorem 2. *For $K = 2, 3, \cdots$, let π_K denote the following state-dependent signal distribution:*

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

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

1. π_K *is the unique optimal information policy, under which agents attack if and only if receiving* s_a *and the status quo persists if and only if* $\theta > \theta_K$, *or*

2. *under an optimal information policy, no agent ever attacks and the status quo always persists.*

Generically, these two cases are mutually exclusive.[11]

Theorem 2 says that the finite-signal problem either has a unique optimal policy, or is trivial since the regime can always persist. The first case is more interesting where the regime-change outcome is determined by a cutoff state θ_K. Also, in the first case, if $\theta_K > 0$, the perfect coordination property fails (see Figure 3 for illustration). These results substantially differ from the unconstrained ($K = \infty$) case, highlighting the significance of agents' reasoning depth up for manipulation. Finally, Theorem 2 immediately implies that θ_K decreases in the number of signals, which is intuitive since a higher maximum level of manipulation can only benefit the designer.

Unique optimum. The argument underlying Theorem 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 go through only 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. Moreover, since credit creation in each round of IESDS is independent of the number of signals used for the particular round, a designer constrained by finite signals should use one signal for each round to maximize the number of rounds. By induction, the optimal information structure must seek to use

[11] The non-generic case corresponds to parameter combination which leads to the minimum value of θ_K such that, upon receiving signal s_a, no attack remains dominated.

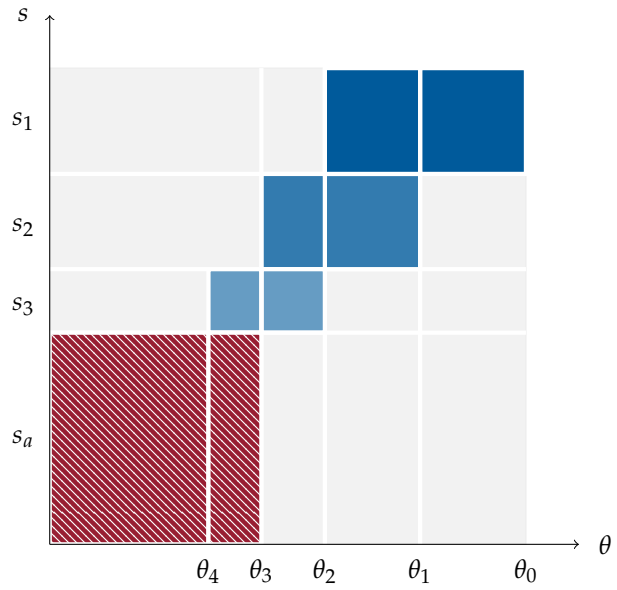

Figure 3: Illustration of level-4 local obfuscation. The horizontal axis represents states and the vertical axis represents signals and their face values. We use differential shades to distinguish information sets following different signals. The regime persists if and only if $\theta \geq \theta_4$, but when $\theta \in [\theta_4, \theta_3)$, perfect coordination property fails.

one signal per round to create maximum possible credit sequentially, which uniquely corresponds to π_K.

Coordination failure under π_K. In stark contrast to Theorem 1, perfect coordination fails when π_K is indeed the unique optimal information structure, or equivalently when the designer is unable to preserve the status quo regardless of θ (the first alternative in Theorem 2). In this case, although agents still coordinate perfectly when $\theta > \theta_{K-1}$ (no agent ever attacks) or when $\theta \leq \theta_K$ (all agents attack), they choose different actions when $\theta \in (\theta_K, \theta_{K-1}]$. Specifically, fraction $1-\theta$ of agents receives s_{K-1} and refrains from attacking, while fraction θ receives s_a and attacks. This is a distinct feature introduced by the constraint in signal space: should the designer have one additional available signal, he would have used it to create another round of IESDS and save more states from attack, but the constraint deprives him of this option. As a result, the last available signal is used as s_a, and although the states in $(\theta_K, \theta_{K-1}]$ still persist because only fraction θ of agents attacks, they no longer facilitate credit creation. The optimality of making coordination imperfect is rather easy to see in Figure 3. Suppose, instead, that coordination is perfect (s_3 is sent to all agents) when $\theta \in [\theta_4, \theta_3)$. Upon receiving signal s_3, an agent now believes that the state is sufficiently likely to be in $[\theta_4, \theta_3)$, and no attack is no longer dominated. To sustain perfect

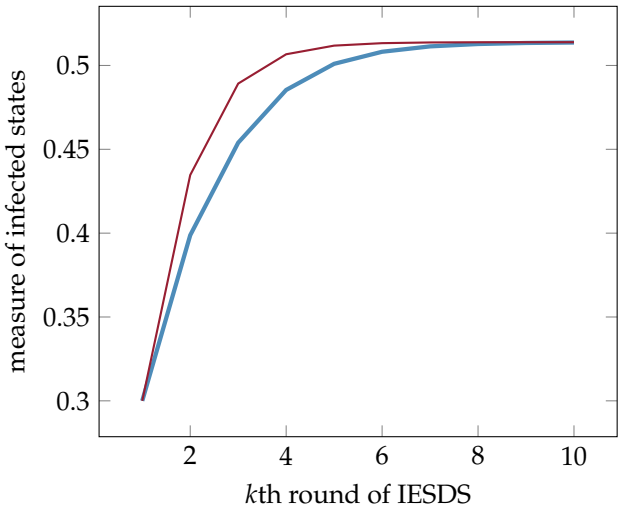

Figure 4: The horizontal axis represents the round of IESDS, and the vertical axis represents the cumulative measure of states being leveraged until each round. The thin red curve corresponds to the optimal local obfuscator π^*, while the thick blue curve corresponds to the alternative information structure 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.

coordination, the designer has to increase the value of θ_4, which is undesirable. This is unnecessary if the entire state space has been infected in finite steps of IESDS or the designer is not constrained by level-K obfuscation.

Multiplicity of optimum when $K = \infty$. 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 maximizes the credit production in every round of IESDS and uses stocking credit most economically, i.e., saves 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 dis-

tributed 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 4: 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 disclosure; after all, a public information policy produces at best the outcome from level-1 manipulation. In section 3.3, we will highlight this advantage of local obfuscation via comparative static analysis.

3.3 Public Disclosure

This section compares the unconstrained benchmark case studied in subsection 3.1 with the most restrictive and perhaps also the most prominent non-trivial bounded-depth obfuscation in subsection 3.2: public disclosure.[12] We study the difference between these two cases by varying the cost of attack, c and the likelihood of attack being dominated, $F(1)$. This exercise allows us to understand the advantage of using private signals to manipulate the higher-order reasoning of agents systematically.

Public signals. First, we derive the optimal public information structure, i.e., for every state θ, signals received by any two agents i, j must be identical. In this case, the higher-order uncertainty among agents is missing. 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.

[12] It is worth noting that although public disclosure essentially represents one scenario of bounded-depth obfuscation, it still imposes an additional constraint on available policies and the optimal public information structure may not correspond to the optimal local obfuscation with $K = 2$ signals. In the latter scenario and following Theorem 2, the designer may benefit from sending different (thus non-public) signals when $\theta \in (\theta_2, \theta_1]$.

Public vs private signals. We are now ready to discuss the advantage of the local obfuscation compared to the public signal (or manipulating higher-order vs first-order reasoning of agents). 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 2. *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. *Consider $\{F_n\}_{n\in\mathbb{N}^+}$ (with f_n, $\theta_n^\dagger$ 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 $\liminf_{n\to\infty} F_n(\theta_n^\dagger) - F_n(\theta_n^*) > 0$.*

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.

3.4 Exogenous Information

In subsections 3.2-3.3, we focused on information constraints such as the limitation on the complexity of information structure and the impossibility of differentiating information across agents. Now we turn to another class of constraint: agents' exogenous private information. Notably, the form of private information will significantly shape the optimal information policy, making it difficult to draw any general conclusion.[13] Therefore we decide to reexamine some classic information manipulation applications in global-game environments, and compare these exercises with our results. This comparison unveils how different economic relevant factors shape optimal information policies in different ways.

In many applications discussed in the related literature section, the agents have access to rich private signals, and the designer is restricted to issuing only public announcements. In these cases, the endogenous contagion logic in agents' iterated reasoning still prevails. The designer's interest, therefore, is still to maximize the "infection" of states in each step of IESDS, which is essentially Inostroza and Pavan (2018)'s logic to prove their Theorem 1. However, the designer's state contagion management is inefficient without full control of information structures. First, the presence of private signals makes it inevitable for each round of IESDS to involve a significant measure of low states, i.e., a high signal implies only a high probability instead of certainty of a strong state. Hence our optimum can never be reached in an environment with such exogenous information due to weakened credit creation. Second, having only public signals at her disposal restrains the designer's ability to distort the influence of agents' private information. Should the designer be free to choose

[13] For instance, Inostroza and Pavan (2018) shows that the optimal public disclosure is a pass/fail test but it can be either monotone or non-monotone in states, depending on the shape of agents' private signals. We have also constructed some special private signals under which a local obfuscator remains optimal, for instance when some agents observe the true state while others remain privately uninformed. The construction is available upon request.

any policy, she could have used a local obfuscator to exclude certain low states from credit consumption in IESDS, thus improving upon even the optimal public signal.

As a concrete example, consider the problem studied by Goldstein and Huang (2016). Each agent independently draws a private signal from Gaussian distribution $\mathcal{N}(\theta, \alpha^{-1})$, while the designer chooses a public disclosure characterized by a cutoff θ_* such that all agents receive a public signal "pass" if $\theta \geq \theta_*$ and "fail" otherwise. The designer looks for the smallest cutoff θ_* s.t. upon receiving the pass signal, agents do not attack regardless of their private signals. Given θ_*, the adversarial equilibrium can still be established by IESDS: there is a sequence of decreasing $\{\hat{\theta}_k\}$ and a sequence of $\{\hat{S}_k\}$ s.t. after k rounds of IESDS, (i) upon receiving signals from $\hat{S}_k$, an agent finds it suboptimal to attack, and (ii) states in $[\hat{\theta}_{k+1}, \hat{\theta}_k)$ have been infected. However, due to the Gaussian nature of agents' private information, each signal in $\hat{S}_k$ may also be sent in every (relatively low) state above the pass-fail cutoff. Therefore, to sustain attack being a dominated strategy, we must have $\hat{\theta}_k$ to be strictly greater than θ_k established by the optimal obfuscator. The procedure continues for infinitely many rounds to produce the final cutoff θ_*. The optimal cutoff θ_* is the smallest value such that $\lim_{k\to\infty} \hat{\theta}_k = \theta_*$, which is higher than θ^*, the cutoff established by the optimal local obfuscator in Theorem 1. The infection inefficiency reflects the economic relevancy of agents' rich private information in information manipulation.

The above Gaussian example can be viewed as an intermediate case between two benchmarks we analyzed in subsections 3.1 and 3.3 in the sense that (i) the designer's ability to effectively manage infection in each step of IESDS is compromised, but (ii) higher-order belief uncertainty among agents remains and can be partially controlled. In the presence of exogenous information, the optimal policy deviates from the benchmark case due to the inefficient infection in each round of IESDS. This substantially differs from another intermediate case studied in subsection 3.2 for $K > 2$ where the inefficiency is due to the constraint of finite rounds of infection. This comparison reflects a conceptual difference between the impact of agents' rich private information and the complexity of signals on information policy.

4 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, there is a optimal persuasion scheme that 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.

References

Alaoui, L. and A. Penta (2016). Endogenous depth of reasoning. *The Review of Economic Studies 83*(4), 1297–1333.

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

Angeletos, G.-M., C. Hellwig, and A. Pavan (2006). Signaling in a global game: Coordination and policy traps. *Journal of Political economy 114*(3), 452–484.

Angeletos, G.-M., C. Hellwig, and A. Pavan (2007). Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks. *Econometrica 75*(3), 711–756.

Angeletos, G.-M. and C. Lian (2016). Incomplete information in macroeconomics: Accommodating frictions in coordination. In *Handbook of macroeconomics*, Volume 2, pp. 1065–1240. Elsevier.

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.

Basak, D. and Z. Zhou (2019). Diffusing coordination risk. *American Economic Review* (forthcoming).

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.

Candogan, O. and K. Drakopoulos (2020). Optimal signaling of content accuracy: Engagement vs. misinformation. *Operations Research 68*(2), 497–515.

Carlsson, H. and E. van Damme (1993). Global games and equilibrium selection. *Econometrica 61*(5), 989–1018.

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

Cong, L. W., S. R. Grenadier, and Y. Hu (2019). Dynamic interventions and informational linkages. *Journal of Financial Economics* (forthcoming).

Crawford, V. P. (2021). Efficient mechanisms for level-k bilateral trading. *Games and Economic Behavior 127*, 80–101.

De Clippel, G., R. Saran, and R. Serrano (2019). Level-mechanism design. *The Review of Economic Studies 86*(3), 1207–1227.

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

Frankel, D. M., S. Morris, and A. Pauzner (2003). Equilibrium selection in global games with strategic complementarities. *Journal of Economic Theory 108*(1), 1–44.

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.

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

Halac, M., E. Lipnowski, and D. Rappoport (2021). Rank uncertainty in organizations. *American Economic Review 111*(3), 757–86.

Heese, C. and S. Lauermann (2020). Persuasion and information aggregation in elections. Technical report, Technical Report, Tech. rept. Working Paper.

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

Huang, C. (2017). Defending against speculative attacks: The policy maker's reputation. *Journal of Economic Theory 171*, 1–34.

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.

Mathevet, L., J. Perego, and I. Taneva (2020). On information design in games. *Journal of Political Economy 128*(4), 1370–1404.

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

Milgrom, P. and J. Roberts (1990). Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica: Journal of the Econometric Society*, 1255–1277.

Moriya, F. and T. Yamashita (2020). Asymmetric-information allocation to avoid coordination failure. *Journal of Economics & Management Strategy 29*(1), 173–186.

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.

Rubinstein, A. (1989). The electronic mail game: Strategic behavior under" almost common knowledge". *The American Economic Review*, 385–391.

Sandmann, C. (2020). Recursive information design. working paper.

Taneva, I. (2019). Information design. *American Economic Journal: Microeconomics 11*(4), 151–85.

A Proofs of Main Results

A.1 Proof of Theorem 1

We first focus on information structures that are deterministic across agents, and show that our proposed optimal local obfuscator, π^*, achieves the designer's optimum among all such information structures. Then we extend our argument to the most general environment allowing for arbitrary information structures.

Consider an arbitrary information structure that is deterministic across agents. We begin by defining two useful series through agents' iterated reasoning, assuming that agents are adversarial against the regime. Series $\{S_k\}_{k=1}^{\infty}$ 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.

Definition of series $\{S_k\}_{k=1}^{\infty}$. Consider an initial strategy profile where everyone attacks regardless of their signal, denoted by a_i^0 such that $a_i^0(s) \equiv 1$ for every i, and $s \in S$. Define $S_1 \subseteq S$ as the set of signal s such that

$$\int_{\Theta} \frac{f(\theta)\pi(s|\theta)}{\int_{\Theta} f(\theta')\pi(s|\theta')d\theta'} \mathbb{1}\{\theta < \int_{[0,1]} \int_{S} a_j^0(v)\pi(v|\theta)dvdj\}d\theta \leq c. \tag{5}$$

The left-hand side of (5) 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 induces a weakly smaller measure of attacking agents than a_i^0. By Lemma 5, 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

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

Since the regime change game is supermodular, $S_k \supseteq S_{k-1}$. Therefore, at the limit as $k \to \infty$, the set $S^* = \lim_{k\to\infty} S_k$ exists, and $S^* \subseteq S$. Also, define

$$a_i^*(s) = \begin{cases} 0 & \text{if } s \in S^* \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

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

Definition of series $\{T_k\}_{k=1}^{\infty}$. 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 in addition to $\cup_{n=1}^{k-1} T_n$ where more than $1 - \theta$ measure of agents receive signals in S_{k-1} , i.e.

$$T_k \equiv \{\theta \in \Theta, \theta \notin \cup_{n=1}^{k-1} T_n : \int_{s \in S_{k-1}} \pi(s|\theta) ds > 1 - \theta\}$$

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.

We may now use the above terminology to characterize the regime's persistence.

Lemma 1. *A necessary and sufficient condition for the regime to persist in equilibrium is $\theta \in T^*$.*

Proof. 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 above construction, 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. □

Next, we identify an upper bound of the ex ante probability that the regime persists under any deterministic information structure.

Lemma 2. *An upper bound of the ex ante probability that the regime persists under a deterministic information structure is given by* $1 - F(\theta^{*\prime})$ *where* $\theta^{*\prime}$ *either uniquely solves*

$$c \int_1^{\bar{\theta}} f(\theta) d\theta + \int_{\theta^{*\prime}}^1 (\theta + c - 1) f(\theta) d\theta = 0, \tag{7}$$

or equals 0 *when (7) has no solution.*

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

For $k = 1, 2, ...$, define

$$D_k = \int_{T_k} f(\theta) \int_{S_{k-1}} \pi(s|\theta) ds d\theta,$$

which is the measure of signals in S_{k-1} being sent for all $\theta \in T_k$. Similarly, for $k = 1, 2, \cdots$, $p = k+1, k+2, \cdots$, define

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

which is the measure of signals in $S_p \backslash S_{p-1}$ being sent when $\theta \in T_k$.

Consider round k of IESDS. 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 not to attack when receiving any signal in S_k is

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

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

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

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

Expanding (IC^k) yields

$$\begin{aligned}
& c \sum_{m=1}^{k} \int_{T_m} f(\theta) \int_{S_k \setminus S_{m-1}} \pi(s|\theta) ds d\theta \\
\geq \;& (1-c) \sum_{m=1}^{k+1} \int_{T_m} f(\theta) \int_{S_{m-1}} \pi(s|\theta) ds d\theta \\
\Leftrightarrow \;& c \int_{\cup_{n=1}^{k} T_n} f(\theta) \int_{S_k \setminus S_{K(\theta)-1}} \pi(s|\theta) ds d\theta \\
\geq \;& (1-c) \int_{\cup_{n=1}^{k+1} T_n} f(\theta) \int_{S_{K(\theta)-1}} \pi(s|\theta) ds d\theta \\
\Rightarrow \;& c\left(\int_{T_1} f(\theta) d\theta + \int_{\cup_{n=2}^{k} T_n} \theta f(\theta) d\theta \right) \geq (1-c) \int_{\cup_{n=2}^{k} T_n} (1-\theta) f(\theta) d\theta.
\end{aligned}$$

Now we are in the position to identify an upper bound of the ex ante probability that the regime persists, $\int_{T^*} f(\theta) d\theta$. Note that $T^* = \cup_{n=1}^{+\infty} T_n$ and $T_1 = (1, \bar{\theta}]$. Therefore, one way to identify a certain superset of the designer's optimum, in terms of states that survive, is to identify a T^* – with a slight abuse of notation – to maximize $\int_{T^* \setminus T_1} f(\theta) d\theta$ under the following constraint

$$c\left(\int_{T_1} f(\theta) d\theta + \int_{T^* \setminus T_1} \theta f(\theta) d\theta \right) \geq (1-c) \int_{T^* \setminus T_1} (1-\theta) f(\theta) d\theta$$

$$
\begin{aligned}
&\Leftrightarrow \quad c\int_{T_1} f(\theta)d\theta \geq \int_{T^*\setminus T_1}(1-c-\theta)f(\theta)d\theta \\
&\Leftrightarrow \quad \int_{T^*\setminus T_1} f(\theta)d\theta \leq \frac{1}{1-c}\left(\int_{T^*\setminus T_1}\theta f(\theta)d\theta + c\int_{T_1} f(\theta)d\theta\right) \\
&\Leftrightarrow \quad \int_{T^*\setminus T_1}(1-\theta)f(\theta)d\theta \leq \frac{c}{1-c}\left(\int_{T^*\setminus T_1}\theta f(\theta)d\theta + \int_{T_1} f(\theta)d\theta\right). \qquad (8)
\end{aligned}
$$

Note that $\int_{T_1} f(\theta)d\theta$ is a constant. We will call this constrained maximization the "relaxed problem."

We assert that the desired T^* which solves the relaxed problem must take the form of $(\theta', \bar{\theta}]$ for some θ', i.e. to solve the relaxed problem, it is always optimal to include in T^* only the strongest states. To see this, first suppose that the right-hand side of (8) were a constant, and suppose that some subset of states outside T^* with a positive probability measure is stronger than another subset of states in T^* (again with a positive probability measure). The term $(1-\theta)f(\theta)$ on the left-hand side of (8) then implies a strict improvement to increase the maximand $\int_{T^*\setminus T_1} f(\theta)d\theta$ without violating (8): switch an identical probability measure of states between the above two subsets, then include more states in T^*. Next, note that this operation will only increase $\int_{T^*\setminus T_1}\theta f(\theta)d\theta$ on the right-hand side, hence preserving the constraint. Therefore the solution to the relaxed problem will include only the strongest states.

Next, let $\tilde{\theta} = F^{-1}(1-\int_{T^*} f(\theta)d\theta)$, as $k \to \infty$ we have

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

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 strictly positive, we use $\theta^{*\prime}$ to denote this lower bound, and $\theta^{*\prime}$ solves

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

If the lower bound is non-positive, there exists π such that every state persists, in which case we simply let $\theta^{*\prime} = 0$.

It's straightforward that $\theta^{*\prime}$ is unique, and then the upper bound of the ex ante

probability that the regime persists is $1 - F(\theta^{*\prime})$. □

We then prove that the optimal local obfuscator π^* achieves exactly this upper bound, and thus is an optimal information structure among those that are deterministic across agents.

Lemma 3. *The probability of the status quo's persistence under π^* is equal to $1 - F(\theta^{*\prime})$.*

Proof. 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 they receive 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 (7); as the solution is unique, we have $\theta^* = \theta^{*\prime}$, i.e. the measure of $\int_{T^*} f(\theta)d\theta$ exactly equals the upper bound we proposed in Lemma 2.

Lastly, all the steps above assume that not every state persists under π^*. If otherwise, for some k we have $\theta_k < 0$; then the regime will always persist under π^*, which is consistent with $\theta^{*\prime} = 0$. □

Lemmas 2 and 3 establish the optimality of π^* among information structures that are deterministic across agents, and we now proceed to show that π^* remains optimal when correlated or non-anonymous policies are allowed. The difference between this specification and an i.i.d information structure 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.

Lemma 4. *π^* remains optimal when the designer may commit to any arbitrary information structure.*

Proof. We start by making slight changes to relevant notations. In the following proofs we use s^i to denote the signal received by agent i. Fix any arbitrary information policy, which we denote here by ψ. We define for every i a series $S^i_{(\cdot)}$ on the signal space.

For every i, let $S_0^i \equiv \emptyset$ and $a_i^0(s) \equiv 1$. Define $S_1^i \subseteq S$ as the set of states satisfying the following condition:

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

and define

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

The term $\Pr(\theta < \int_{[0,1]} a_j^0(s^j)dj|\theta, s^i = s)$ captures the possibilities that (1) ψ may not be anonymous, since the measure of attacking agents results from integrating their individual actions along $[0,1]$ instead of being represented by a single distribution/measure π; (2) ψ may correlate agents' signals and make the aggregate signal distribution random even for a fixed θ, which makes the event $\theta < \int_{[0,1]} a_j^0(s^j)dj$ bearing a probability instead of either occurring or not with certainty.

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

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

and define

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

We also define $S^{i*} = \lim_{k\to\infty} S_k^i$ and $a_i^*(\cdot) = \lim_{k\to\infty} a_k^i(\cdot)$ accordingly. Following the proof of Proposition 1[14], $\{S^{i*}\}_{i\in[0,1]}$ characterizes the unique agent equilibrium.

To simplify our discussion, we may without loss of generality restrict our attention to policies under which every agent's action upon receiving the same signal is identical. That is, we can always construct such a policy ψ' which always yields the same outcome as ψ does. Note that, for every bijective mapping A from S to S and every $s \in S$, if we let ψ' send $A(s)$ to agent i whenever ψ sends s, then agent i's action when receiving $A(s)$ under policy ψ' will be the same as his action when receiving s under policy ψ. Thus through selecting for each agent an appropriate bijective mapping, we can always find ψ' such that $a_i^k(s) \equiv a_j^k(s)$ for every k and every $i, j \in [0,1]$, and the status quo's ex ante probability of persistence under ψ' is identical to its ex ante probability of persistence under ψ. As a result, we may now omit the superscript

[14] In the proof in Appendix B, we show that the unique equilibrium strategy of agent i is to attack if and only if $s^i \in S^{i*}$.

in S_k^i for every k and i, and use notation S_k henceforward.

Next, we identify, by explicit construction, an upper bound of the regime's ex ante probability of persistence. First define a series of function $h_0(\theta), h_1(\theta), \cdots$ as follows:

$$\begin{cases} h_0(\theta) = 0 \forall \theta \in \Theta \\ h_1(\theta) = f(\theta) \forall \theta \in T_1 \text{ and } = 0 \text{ elsewhere} \\ h_2(\theta) = f(\theta) \Pr(\int_{i \in [0,1]} \mathbb{1}(s^i \in S_1 | \theta) di > 1 - \theta) \\ \cdots \end{cases}$$

where $h_k(\theta)$ is the probability density function that state θ survives after k rounds of IESDS. By construction, the distance between $h_k(\theta)$ and $h_{k-1}(\theta)$ converges to 0 as $k \to +\infty$.

For $i \in [0,1]$, $k = 1, 2, ...$, define

$$D_k^i = \int_\Theta (h_k(\theta) - h_{k-1}(\theta)) \Pr(s^i \in S_{k-1} | \theta) d\theta.$$

For $i \in [0,1]$, $k = 1, 2, ...$, $p = k+1, k+2, ...$, define

$$C_{k,p}^i = \int_\Theta (h_k(\theta) - h_{k-1}(\theta)) \Pr(s^i \in S_p \backslash S_{p-1} | \theta) d\theta.$$

By the definition of $T_{(\cdot)}$ and $S_{(\cdot)}$, for every k, i, $c \sum_{m=1}^k \sum_{p=m}^k C_{m,p}^i \geq (1-c) \sum_{m=1}^{k+1} D_m^i$. Thus, similar to the previous proof, for every i, θ

$$\begin{aligned} & c \sum_{m=1}^k \sum_{p=m}^k C_{m,p}^i \geq (1-c) \sum_{m=1}^{k+1} D_m \\ \Leftrightarrow \quad & c \sum_{m=1}^k \sum_{p=m}^k \int_\Theta (h_m(\theta) - h_{m-1}(\theta)) \Pr(s^i \in S_p \backslash S_{p-1} | \theta) d\theta \\ \geq \quad & (1-c) \sum_{m=1}^{k+1} \int_\Theta (h_m(\theta) - h_{m-1}(\theta)) \Pr(s^i \in S_{m-1} | \theta) d\theta \\ \Leftrightarrow \quad & c \int_{[0,1]} \sum_{m=1}^k \sum_{p=m}^k \int_\Theta (h_m(\theta) - h_{m-1}(\theta)) \Pr(s^i \in S_p \backslash S_{p-1} | \theta) d\theta di \\ \geq \quad & (1-c) \int_{[0,1]} \sum_{m=1}^{k+1} \int_\Theta (h_m(\theta) - h_{m-1}(\theta)) \Pr(s^i \in S_{m-1} | \theta) d\theta di. \end{aligned}$$

The last step is an integration across all agents.

Note that, except for the states in $[1, \bar{\theta}]$, for the status quo to persist after the kth round of IESDS it must send signals in S_k to a population greater than or equal to $1-\theta$. Also note that $\int_\Theta (h_m(\theta) - h_{m-1}(\theta))d\theta$ is the probability measure of the states that persists after and only after the kth round of IESDS. Thus, for every θ, m, and p, $\int_{[0,1]} (h_m(\theta) - h_{m-1}(\theta)) \Pr(s^i \in S_{m-1}|\theta)di$ must be greater than or equal to $(1-\theta)(h_m(\theta) - h_{m-1}(\theta))$; $\int_{[0,1]} \sum_{p=m}^{k} (h_m(\theta) - h_{m-1}(\theta)) \Pr(s^i \in S_p \backslash S_{p-1}|\theta)di$ must be less than or equal to $\theta(h_m(\theta) - h_{m-1}(\theta))$. Then, the last inequality above leads to the following necessary condition:

$$
\begin{aligned}
& c[\int_\Theta h_1(\theta)d\theta + \sum_{m=2}^{k} \int_\Theta (h_m(\theta) - h_{m-1}(\theta))\theta d\theta] && (9) \\
\geq \quad & (1-c)[\sum_{m=2}^{k+1} \int_\Theta (h_m(\theta) - h_{m-1}(\theta))(1-\theta)d\theta + \int_\Theta h_1(\theta)0d\theta] \\
\Leftrightarrow \quad & c(\int_\Theta h_1(\theta)d\theta + \int_\Theta (h_k(\theta) - h_1(\theta))\theta d\theta) \\
\geq \quad & (1-c) \int_\Theta (h_{k+1}(\theta) - h_1(\theta))(1-\theta)d\theta. && (10)
\end{aligned}
$$

Note that the last condition is irrelevant to agent identity i. A necessary condition that bound from above the measure of persisting states, is (10) with $k = \infty$, which turns out to be identical to (8). Therefore we can adopt the previous proof to claim the optimality of π^*. □

A.2 Proof of Theorem 2

We prove that if the regime persists with probability less than 1 under any optimal policy, then π_K is the unique optimal policy.

If the regime persists with probability less than 1 under π_K, then $\theta_k \geq 0$ for every $k \leq K$. Suppose that π'_K is an optimal policy and π'_K is different from π_K.

It is straightforward that a information policy using K signals can manipulate the agents' higher-order reasoning up to level $K-1$ and can induce at most $K-1$ rounds of IESDS. The later is because that, after K rounds of IESDS, an individual agent shall not attack upon receiving at least K different signals, that is, she shall never attack, and that contradicts our assumption that the regime persists with probability less than 1. Suppose that π'_K induces $m \leq K-1$ rounds of IESDS. By the definition of π_K,

we recursively define a sequence $\{C_k, k|\pi_K, D_k|\pi_K\}_{k=1}^K$ as follows

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

For π'_K, we specify the necessary conditions for each round of IESDS:

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

The rest of our proof proceeds in the following steps.

Step 1. *We prove that* $C_{1,1}|\pi'_K = C_{1,1}|\pi_K$.

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

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

Step 2. *We prove that in every optimal information structure, the amount of credit and discredit created in each round of IESDS must be identical to that of π, i.e., for $p = 1, 2, \cdots, m$, $C_{p,p}|\pi'_K = C_{p,p}|\pi_K$, $D_{p+1}|\pi'_K = D_{p+1}|\pi_K$.*

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

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

Step 3. *We prove that, for $p = 1, 2, \cdots, m+1$, $T_p|\pi'_K$=$T_p|\pi_K$. Also, if $\theta \in T_1|\pi'_K$, the regime sends signals in $S_1|\pi'_K$ with probability 1. For $p = 2, 3, \cdots, m$, for every $\theta \in T_{p+1}|\pi'_K$, the regime sends signals in $S_{p+1}|\pi'_K \backslash S_p|\pi'_K$ with probability θ. and signals in $S_p|\pi'_K \backslash S_{p-1}|\pi'_K$ with probability $1-\theta$.*

First, note that $T_1|\pi'_K$=$T_1|\pi_K = T_1 = (1, \bar{\theta}]$ by definition. Then, as $C_{1,1}|\pi'_K = C_{1,1}|\pi_K = \bar{\theta} - 1$, the regime must send signals in S_1 with probability 1 when its true state is in $(1, \bar{\theta}]$, otherwise, $C_{1,1}|\pi'_K < C_{1,1}|\pi_K$.

Next, by definition, for $p = 1, 2, \cdots, m+1$, for every $\theta \in T_p|\pi'_K$, $\int_{S_{p-1}} \pi'_K(s_i|\theta) ds_i \geq \min\{0, 1-\theta\}$. Therefore, to have $D_2|\pi'_K = D_2|\pi_K = \frac{c}{1-c} C_{1,1}|\pi_K$ and $C_{2,2}|\pi'_K = C_{2,2}|\pi_K$ simultaneously, the regime must send signals in $S_1|\pi'_K$ with probability exactly equal to $1-\theta$ and signals in $S_2|\pi'_K \backslash S_1|\pi'_K$ with probability exactly equal to θ when θ is in $T_2|\pi'_K$, and $T_2|\pi'_K$ must be exactly equal to $T_2|\pi_K$, otherwise, $C_{2,2}|\pi'_K < C_{2,2}|\pi_K$. By mathematical induction, $T_p|\pi'_K = T_p|\pi_K$, and for every $\theta \in T_{p+1}|\pi'_K$, the regime sends signals in $S_{p+1}|\pi'_K \backslash S_p|\pi'_K$ with probability θ. and signals in $S_p|\pi'_K \backslash S_{p-1}|\pi'_K$ with probability $1-\theta$.

Step 4. *We prove that, for $p = 1, 2, \cdots, m$, $S_p|\pi'_K$ contains only 1 signal. Also, there is only one signal upon receiving which an agent will attack.*

Step 3 has established that, since $T_p|\pi'_K$=$T_p|\pi_K$ $\forall p = 1, 2, \cdots, m+1$, m must be equal to $K-1$ for π'_K to be optimal.

Now suppose that, for some p, $S_p|\pi'_K$ consists of more than 1 signal. By the proof of step 3, these signals are sent with positive probability by and only by the states in $T_p|\pi'_K$ and $T_{p+1}|\pi'_K$. Therefore, after the pth round of IESDS, only less than $K-p$ signals remain available for inducing additional rounds of IESDS, which can only induce less than $K-p-1$ rounds. Thus the total number of rounds of IESDS induced by π'_K, m, is strictly less than $p+K-p-1=K-1$. Therefore π'_K is not optimal, a contradiction.

Lastly, suppose that the agents attack upon receiving signals in S_a and S_a contains more than 1 signal. Then π'_K uses at most $K-2$ signals to induce IESDS. $K-2$ signals can induce, at most, $K-2$ rounds of IESDS. Again, m is strictly less than $K-1$ and π'_K is not optimal, a contradiction.

Combining Steps 3 and 4, π'_K and π_K must be identical. This completes the proof.

A.3 Proofs in Section 3.3

Proof of Proposition 2. 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 Online Appendix

B.1 Proof of Proposition 1

This proof is more general than the discussions in the main text in the sense that correlated or non-anonymous policies are allowed. We begin by defining 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 5. *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

$$\begin{aligned} c &< \int_\Theta \left(\frac{f(\theta)\pi_i(s|\theta)}{\int_\Theta f(\theta')\pi_i(s|\theta')d\theta'} \Pr(\theta < \int_{[0,1]\setminus\{i\}} a'_j(s^j)dj|\theta) \right) d\theta \\ &\leq \int_\Theta \left(\frac{f(\theta)\pi_i(s|\theta)}{\int_\Theta f(\theta')\pi_i(s|\theta')d\theta'} \Pr(\theta < \int_{[0,1]\setminus\{i\}} a_j(s^j)dj|\theta) \right) d\theta. \end{aligned}$$

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 $j \neq i$, which implies $a_j(s^j = s) \geq a'_j(s^j = s)$ for every $j \neq i$ and every $s \in S$. Thus, agent i finds it optimal to attack given signal s and a_{-i}. □

Now we are ready to address the equilibrium existence.

Lemma 6. *For any information structure, there exists an equilibrium.*

Proof. We show that it is indeed an equilibrium for agent i to attack if and only if $s^i \in S^{i*}$, with S^{i*} defined in the proof of Lemma 4. First, by the construction of S^{i*}, agent i prefers attacking when receiving every signal in $S \backslash S^{i*}$ given other agents follow the strategy specified in (6). Second, we show that for any non-empty S^{i*}, given that the other agents follow the strategy in (6), an individual agent i strictly prefers not attacking for every signal in S^{i*}. The proof is straightforward. Pick any $s \in S^{i*}$, there exists a unique k such that $s \in S^i_k \backslash S^i_{k-1}$. By the definition of S^i_k, given that every other agent $j \neq i$ follows a_j^{k-1} and does not attack if and only if receiving signals in S^j_{k-1}, agent i prefers not attacking when receiving signals in $S^i_k \backslash S^i_{k-1}$. Then by Lemma 1, if every other agent $j \neq i$ follows a less aggressive strategy $a_j^* < a_j^{k-1}$ and does not attack if and only if receiving signals in $S^{j*} \supseteq S^j_{k-1}$, agent i must prefer not attacking when receiving signals in $S^i_k \backslash S^i_{k-1}$. Thus, for every $s \in S^{i*}$, $a_i^*(s) = 0$. Hence, we have the desired result. □

The definitions above guarantee a unique series of $\{S^i_k\}$ and a unique S^{i*} for every $i \in [0, 1]$. In what follows, we show that $a_i^*(s)$ is the unique equilibrium as well.

Lemma 7. *For any information structure, there is a unique (adversarial) equilibrium.*

Proof. For the sake of contradiction, suppose that for some information structure, there are two distinct equilibria a, a'. Let $\{s|a_i(s) = 0\}$ denote the set of signals agent i does not attack in equilibrium a and $\{s|a'_i(s) = 0\}$ denote the set of signals agent i do not attack in equilibrium a'. By the hypothesis that a and a' are distinct equilibria, there exists i such that $\{s|a(s) = 0\} \neq \{s|a'(s) = 0\}$. Consider strategy a'' defined as follows:

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

For every i, a'' is weakly more aggressive than a and a' and strictly more aggressive than at least one of them. By Lemma 1, agent i receiving a signal in $S \backslash (\{s|a_i(s) = 0\} \cap \{s|a'_i(s) = 0\})$ 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 5-7 yields Proposition 1.

B.2 Omitted Results on Comparative Statics

Comparative statics for public disclosure. To ease the discussion of comparative statics, we rewrite equation (4) as

$$1 - c = \frac{1 - F(1)}{1 - F(\theta^\dagger)}. \tag{13}$$

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 $c \geq F(1)$.*
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.

Comparative statics for local obfuscation. Now we turn to optimal local obfuscation in the unconstrained case ($K = \infty$). Rewrite equation (3) as

$$1 - c = \frac{1 - F(1)}{1 - F(\theta^*)} + \frac{\int_{\theta^*}^1 \theta f(\theta) d\theta}{1 - F(\theta^*)}. \tag{14}$$

Compared to equation (13), equation (14) 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 $(1, \bar{\theta}]$.

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

$$c \geq F(1) - \int_0^1 \theta f(\theta) d\theta, \tag{15}$$

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 the optimal local obfuscator, 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 $\liminf_{n \to \infty} 1 - F_n(\theta_n^*) > 0$.*

Proof. 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 satisfied, 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. □