Global Manipulation by Local Obfuscation*

Fei Li† Yangbo Song‡ Mofei Zhao§

June 6, 2023

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, making the designer's information disclosure time consistent and relieving the usual commitment concern.

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. Yangbo Song gratefully acknowledges funding from Project 72192805 supported by NSFC. Mofei Zhao gratefully acknowledges funding from Project 72103016 supported by NSFC.

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

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

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

1 Introduction

Many economic problems with strategic complementaries are modeled as regime-change games where a status quo is overturned if enough agents attack it. Examples include speculation against a pegged currency, a 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 2022), 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 her 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 constraints, 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 or private, to each agent. Her objective is to maximize the probability of preserving

[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 access to other information sources out of the reach of the information designer. See more discussion in section 1.1.

the regime in her *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. 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 needs only to convince agents of no sufficient coordination. In this spirit, the designer may leverage not only on the invincible states but 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.

One practical context where this information policy can be effectively implemented is digital authoritarianism: governments worldwide are increasingly using advanced technology to consolidate their control over information and suppress dissent. The state in this application is a government's tolerance threshold for implementing an institution or imposing a regulation, i.e. the least fraction of opposing citizens that can force it to forfeit the scheme. Through censorship and the spread of disinformation, a government with "cyber sovereignty" may present each citizen with private and marginally different messages about the state.

The optimal local obfuscation collapses *global coordination* among agents by creating both fundamental and belief uncertainty. The former refers to uncertainty about the regime's strength; while the latter refers to uncertainty about other agents' beliefs on the regime's strength. 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 belief uncertainty (which corresponds to the regime's state after each round of IESDS, i.e. whether enough agents can still be coordinated to overthrow the regime) 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," an analogy proposed by Rubinstein (1989). We further show that to at-

tain 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 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 2022).

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. (2020)).[3] In particular, when the designer

[3]Also see Morris et al. (2022a) for a global-game-like optimal policy in regime-change games using

is capable only 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 her 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 in 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 ways. First, given a target set of persisting states, optimal local obfuscation allows 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.

the idea of Morris et al. (2020).

Implementation. Our framework allows the information designer the maximum degree of freedom to choose from all information policies, including ones with correlated and/or non-anonymous signals. Nevertheless, the optimal policy requires only 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). 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.[4] We apply our theory to explain the recent transition of dominant model autocrats. We argue that employing a divide-and-conquer type of information policy is less costly to make information disclosure time-consistent.

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 (2022) 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 contexts). On the front of information design context, our paper is mostly related to Inostroza and Pavan (2022), 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

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

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 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 the 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, while he knows that all his peers receive one of at most two adjacent signals. We also explore the case where finite levels of higher-order agents' strategic reasoning are up for manipulation. We establish coordination failure under the unique optimal policy. This is in stark contrast to the coordination outcome when the designer can manipulate infinite levels of higher-order reasoning as in our baseline model, Inostroza and Pavan (2022) and Morris et al. (2020).

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. 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. They further discuss the connection between adversarial information design and the literature on information robustness (Kajii and Morris, 1997). Several companion recent studies, e.g., Halac et al. (2021), Moriya and Yamashita (2020), and Sandmann (2020), explore this property in various kinds of finite supermodular games and derive optimal information structures conceptually simi-

lar to ours. The connection also plays a substantial role in deriving optimal public disclosure in Inostroza and Pavan (2022). Hoshino (2022), instead of focusing on supermodular games, considers information design in finite games with a **p**-dominant equilibrium. Building on a similar connection between persuasion and information robustness, he shows that using the leverage of strategic uncertainty, agents can be persuaded to take such an action profile as a unique rationalizable outcome given any non-degenerate prior.

Among other companion papers, the most closely related one is Morris et al. (2020), who propose a tractable method to characterize the set of implementable adversarial equilibrium outcomes in all binary-action supermodular games, including regime-change games. They provide conditions for verifying *whether* an outcome is smallest-equilibrium implementable and fully implementable. In comparison we characterize, in the regime-change context, exactly *what* the optimal implementable outcome for the designer is and *how* to explicitly implement it with an information structure. 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 closed form, which is convenient for comparative statics exercises. 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 3.4. Also see the implementation discussion of Morris et al. (2020), which features 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, in sharp contrast to the literature of adversarial information design in supermodular games (see, e.g., Morris et al. (2020) and Inostroza and Pavan (2022)). 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 theirs, 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 (he), 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 mass of population that attacks by A. 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 assumed to be differentiable for every θ for concise exposition, and let $f(\theta)$ denote its density function.

If an agent does not attack, his payoff is zero. If he attacks, his payoff depends on the regime status: he incurs cost $c \in (0,1)$ regardless of the regime status, and if the regime is overthrown, he receives a benefit, which is normalized to be 1. An agent's

utility function is therefore

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

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

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

For consistency with the agents' utility, we assume that the regime does not fail if no agent (or a zero measure of agents) attacks.

Information structure. An information designer (she) commits to disclosing information to the agents about the state θ. For any complete separable metric space X, let $\Delta(X)$ denote the set of probability measures on X with respect to its Borel sigma-algebra $\mathcal{B}(X)$, and endow $\Delta(X)$ with the weak topology, so that it is also a complete separable metric space. An *experiment* is a pair $(S, (\pi(\cdot|\theta))_{\theta\in\Theta})$ such that S is a complete separable metric space of signals, and $\pi(\cdot|\theta) \in \Delta(S)$ is a probability distribution generating signal in state θ, where $\pi(S'|\theta)$ is measurable in θ for each $S' \in \mathcal{B}(S)$. Let Π denote the set of all experiments.

Given an experiment $(S, (\pi(\cdot|\theta))_{\theta\in\Theta})$, an *information structure* can be generated as follows. For each state $\theta \in \Theta$, each agent independently and privately receives a random draw experiment $\pi(\cdot|\theta)$. Heuristically applying the logic of "law of large numbers," $\pi(\cdot|\theta)$ also represents the empirical distribution of signal realizations.[5] For each experiment and θ, the corresponding information structure is *deterministic* (since the empirical distribution is unique) and *anonymous* (since the agents' signals are drawn from identical distributions). We thus obtain a joint probability measure $\rho \in \Delta(S \times \Theta)$ by $\rho(S' \times \Theta') = \int_{\Theta'} \pi(S'|\theta) dF(\theta)$. Furthermore, a regular conditional probability $\pi(\cdot|s)$ given signal s is well defined and unique up to ρ-a.e. equivalence; fix any member of the equivalent class. In the rest of the paper, we abuse notations to use π to denote both an experiment and its resulting information structure unless otherwise told.

A few remarks on our specification of information structure are in order. First, alternatively to the aforementioned identically and independently distributed private signals approach, one may think of the information designer as conducting the following two-step process to obtain an information structure: given an arbitrary θ, first

[5] See Sun (2006) for a formal measure-theoretic modelling with a continuum of players with independent randomness.

she takes an empirical signal distribution $\pi(\cdot|\theta) \in \Delta(S)$; then she implements $\pi(\cdot|\theta)$ among the agents in a uniformly random way as in, e.g., Prescott and Townsend (1984). For instance, suppose that $S = [0,1]$. One such implementation is to align the agents on a unit circle, randomly select an agent and label him as $i = 0$, and label the remaining agents in a clockwise order. Then agent $k \in [0,1]$ receives signal s if and only if $\pi([0,s]|\theta) \geq k$ and $\nexists\, s' < s$ such that $\pi([0,s']|\theta) \geq k$. It is easy to see that the information structure is still deterministic and anonymous.

Second, in our current specification, the designer is restricted to induce a deterministic information structure. This restriction is made for expositional convenience. Our characterization of an information structure and our main results extend readily to non-deterministic but still anonymous information structures. The idea is to randomly draw an experiment according to a committed distribution from $\Delta(\Pi)$ and then follow the aforementioned procedure, resulting in randomness on the empirical signal distribution. In this case, agents observe the distribution over experiments rather than the realized experiment. We relegate the formalization and analysis of non-deterministic information design to Appendix B.

Bayesian game and adversarial design. 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 a signal according to $\pi(s|\theta)$, and all agents simultaneously choose their actions. An agent *strategy* is a measurable function $a : S \to \{0,1\}$, and we focus on symmetric pure strategy profiles. The aggregate mass of attacking agents at state θ is then well defined and measurable in θ: $A(a|\theta) = \pi(\{s \in S|a(s) = 1\}|\theta)$.

In a Bayesian Nash equilibrium, given a and his own signal s, agent i attacks only if he weakly prefers to attack. Precisely, denote the belief about state θ formed by an agent who observes signal s as $\pi(\cdot|s)$, for each information structure, then $\int_\Theta u(a(s), A(a|\theta), \theta) d\pi(\theta|s) \geq \int_\Theta u(b, A(a|\theta), \theta) d\pi(\theta|s)$ for all $b \in \{0,1\}$. For a given information structure, there may be multiplicity of equilibrium. We solve for the information designer's *worst* Bayesian Nash equilibrium to capture the idea of adversarial/robust information design.[6] That is, agents coordinate on a strategy profile such that the largest measure of agents attacks and each agent attacks when indifferent. In the remainder of the article, we refer to the information designer's worst Bayesian Nash equilibrium as (adversarial) *equilibrium*.

[6]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.

The information designer's problem is to choose an information structure $\pi \in \Pi$ to induce an (adversarial) equilibrium which maximizes the regime's expected probability of persistence. [7]

3 Analysis

We begin with equilibrium characterization for an arbitrary information structure.

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 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).[8]

Here we provide the intuition for the proof in our model. 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 he 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:

[7] The principal's optimum may not be exactly attained, in which case we focus on the approximation by the supremum of the regime's expected probability of survival.

[8] Specifically, given an information policy where the state-dependent signal distribution is well-defined and measurable on S, let θ be the state and the signal realizations be the players' types in Van Zandt (2010) respectively. Then the unique equilibrium corresponds to the greatest Bayesian Nash equilibrium in Van Zandt (2010)'s framework.

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 to characterize an optimal information structure, which maximizes the probability of the status quo's persistence. In what follows, we introduce a class of information structures.

Definition 1. *An experiment π is a **local obfuscator** if*

1. *there is a cutoff state $\theta^* \in [0, \bar{\theta}]$ that partitions the state space into a sequence of intervals $[\theta_1, \theta_0] \cup \{[\theta_{k+1}, \theta_k)\}_{k=1}^{\infty} \cup [0, \theta^*)$ 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 the same interval. Second, when an agent is misinformed about the true interval, the signal he 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 his unresolved uncertainty about the fundamental state is local. Moreover, the obfuscation creates belief uncertainty among agents, making

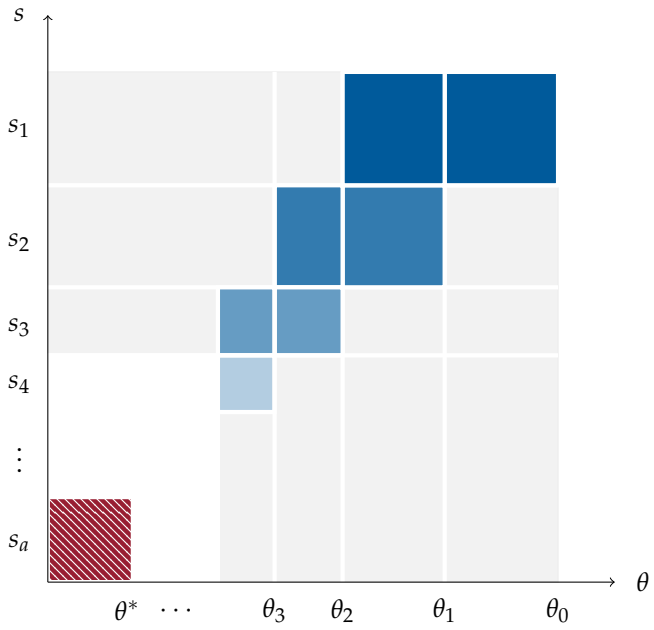

Figure 1: Illustration of local obfuscator. The horizontal axis represents states and the vertical axis represents signals and their face values. We use differential shades to distinguish information sets following different signals.

the 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 could 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 he 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^*.

3.2 Discussion of Theorem 1

Below we explain how agents conduct their iterated reasoning under π^* and elaborate our argument for its optimality to the designer. The discussion will be heuristic, and the proof of Theorem 1 can be found in the Appendix.

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 his knowledge about π^*, he infers that the true state θ is in $[\theta_1, \theta_0] \cup [\theta_2, \theta_1)$. 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 $\theta \in [\theta_3, \theta_2) \cup [\theta_2, \theta_1)$; on the other hand, he 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.

Endogenized iterated reasoning. 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 features small i.i.d. "noise" at each step of infection.

Consider the Bayesian incentive compatibility constraint for an arbitrary round of

IESDS. Upon seeing some signal designated for this round, an agent refrains from attacking because he believes that given the current maximal coordination, the state is sufficiently likely to be "conditionally invincible" after the previous IESDS. We thus use the term "credit" as a proxy for the likelihood of the state being conditionally invincible, and the term "discredit" or "consumption of credit" for the likelihood of the state being conditionally vincible. Other things constant, if a signal implies a higher probability on a conditionally invincible state, it creates more credit for the current round of IESDS; otherwise if the signal implies a higher probability on a vincible state, it consumes more existing credit or creates more discredit. Bayesian incentive compatibility is then equivalent to a balance between credit and discredit, i.e. the amount of discredit cannot exceed the budget of credit.

We illustrate how the system works with 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 he is facing an invincible state ($\theta \geq 1$). Hence the invincible states provide the initial endowment of "credit," with amount equal to the probability measure of the invincible states, $\int_1^{\bar{\theta}} f(\theta)d\theta$; 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, pooling strong and weak states together under the same signal.

The state obfuscation and infection does not stop in one round: 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 (Rubinstein, 1989). 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 (in the amount of $\int_{\theta_2}^1 \theta f(\theta)d\theta$ under π^*), since signal s_2 matches its face value. More $\theta < 1$ states can then consume this credit, avoid being attacked and further create credit themselves, and the process moves on as IESDS (the states infection) 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 con-

straint for signal s_{k+1} thanks to the coordination friction.[9]

The optimality of π^*. The above iterated reasoning process reveals a basic principle that the designer must abide by when seeking her optimum. Given any signal that induces a certain round of IESDS among agents, the likelihood of conditionally vincible states (discredit in the current round) cannot exceed $\frac{c}{1-c}$ of the likelihood of conditionally invincible states (credit in the current round). Combining these constraints for all rounds of IESDS provides a *necessary* condition on what states may persist under the particular information structure: the likelihood of all states *ever* being conditionally invincible in some round (total credit) must be no less than $\frac{1-c}{c}$ of the likelihood of them *ever* being conditionally vincible in some round (total discredit).

We further notice that state infection under π^* is *monotone*. That is, letting $\Theta_0^* = [1, \bar{\theta}]$, and for every $j = 1, 2, ..., \Theta_j^* = [\theta_{j+1}, \theta_j)$ denotes the set of states being infected in the jth round of IESDS under π^*, $\sup \Theta_j^* = \inf \Theta_{j-1}^*$. To see why this is optimal, take an arbitrary $k \geq 1$ and consider an alternative information policy π which coincides with π^* in the first $k-1$ rounds of state infection. Therefore, regarding the set of infected states via IESDS, we have $\Theta_j = \Theta_j^*, j < k$. Now suppose that in the kth round, an arbitrary measurable set $\Theta_k \subseteq \Theta \setminus \left(\cup_{j=0}^{k-1} \Theta_j \right)$ is infected. This implies that when $\theta \in \Theta_k$, the designer sends one signal s_k when she obfuscates Θ_k with Θ_{k-1}. The credit-discredit budget constraint is then

$$\underbrace{\int_{\Theta_k} \pi(s_k|\theta) dF(\theta)}_{D_k} \leq \frac{c}{1-c} \underbrace{\int_{\Theta_{k-1}} \pi(s_k|\theta) dF(\theta)}_{C_{k-1}}, \tag{4}$$

where for $\theta \in \Theta_{k-1}$,

$$\pi(s_k|\theta) = \pi^* = \begin{cases} 1 & \text{if } k = 1 \\ \theta & \text{if } k > 1 \end{cases}, \tag{5}$$

and C_k (or D_k) denotes the probability that an agent receives signal s_k and $\theta \in \Theta_{k-1}$ (or $\theta \in \Theta_k$). Constraint (4) states that the discredit for round k, D_k, must not exceed the adjusted credit for round k, $\frac{c}{1-c} C_{k-1}$. Upon receiving signal s_k, an agent knows

[9] The coordination feature of the base game plays a key role here. When $\theta \in [0, 1)$, neither attacking nor refraining is dominant – attacking is optimal if and only if enough others also attack. Hence, in some 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.

that $\theta \in \Theta_k \cup \Theta_{k-1}$ and states in Θ_{k-1} are conditionally invincible due to previous IESDS. Evidently, when constraint (4) is satisfied (slack), the agent's attack payoff

$$(1-c)D_k - cC_{k-1}$$

is (strictly) less than 0, making $a(s_k) = 1$ (strictly) dominated. Furthermore, for $\theta \in \Theta_k$ to serve as the leverage of infecting more states in round $k+1$, it must be conditionally invincible after round k, i.e. $\pi(s_k|\theta) \geq 1-\theta$. Meanwhile note that the remaining probability $1-\pi(s_k|\theta)$ creates credit for the next round $k+1$: $1-\pi(s_k|\theta) \equiv \pi(s_{k+1}|\theta)$ (where, without loss of generality, s_{k+1} is also the signal sent under π when $\theta \in \Theta_k$ but is not obfuscated with Θ_{k-1}), implying $1-D_k \equiv C_{k+1}$. Hence given Θ_k, it is optimal to set $\pi(s_k|\theta) = 1-\theta \; \forall \theta \in \Theta_k$, so that (4) is satisfied if it can ever be, while the maximum amount of credit is created for round $k+1$.

Now we consider the choice of Θ_k. Rewrite constraint (4), given $\pi(s_k|\theta) = 1-\theta$, as

$$\left(1 - \mathbb{E}[\theta|\theta \in \Theta_k]\right) \int_{\theta \in \Theta_k} dF(\theta) \leq \frac{c}{1-c} C_{k-1}, \tag{6}$$

where the left-hand side is the measure of infected states $\int_{\theta \in \Theta_k} dF(\theta)$ multiplied by $1 - \mathbb{E}[\theta|\theta \in \Theta_k]$. Clearly, the mass of state infection is maximized by setting $\Theta_k = \Theta_k^*$ at which (i) the conditional expectation of infected states in the kth round is maximized, and (ii) the credit-discredit constraint (6) is binding. This is intuitive. It is sufficient to make state θ conditionally invincible if $1-\theta$ mass of agents are persuaded not to attack, so infecting the highest-not-yet-infected states minimizes the necessary mass of agents to be persuaded to prevent regime-change in each infected state, giving rise to the largest mass of state infection. Moreover, (6) can also be written as

$$\int_{\theta \in \Theta_k} dF(\theta) - \frac{c}{1-c} C_{k-1} \leq \int_{\theta \in \Theta_k} \theta dF(\theta),$$

the right-hand side of which represents the credit for round $k+1$, given that every $\theta \in \Theta_k$ has optimally sent s_k to mass $1-\theta$ of agents. Note that this credit amount coincides with C_k when formula (5) extends to round $k+1$. That is, setting $\Theta_k = \Theta_k^*$ not only maximizes the mass of state infection in the kth round but also relaxes the credit-discredit budget constraint to the largest extent in round $k+1$. Because k is arbitrary, the optimality of monotone infection follows by induction. See Figure 2 for a numerical example demonstrating suboptimality of non-monotone infection.

This argument has three implications. First, sending one signal for each round of obfuscation, or state infection via IESDS, is indeed optimal. If some round of IESDS

were induced by multiple signals, (4) must be satisfied given every such signal. It then follows, by summing up the constraints, that the designer must be able to implement the same state infection by sending only one signal instead. Second and most importantly, for every information structure infecting finite sets of states sequentially, there always exists an information structure which infects the same number of sets of states in descending order of magnitude but results in a weakly larger probability measure of the total infected states. We may then conclude that at least one optimal information structure must induce, through IESDS, a series of *connected* intervals of states in descending order of magnitude. The set of persisting states under this information structure is then $[\theta^*, \bar{\theta}]$, which represents the union of the intervals. Finally and as a side remark, it is without loss of generality to assume no atom in $F(\theta)$, because what essentially matters for constraint (4) is two expected values D_k and C_{k-1}.

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 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.

The optimality of π^*: an alternative angle. Finally, an alternative way to understand the optimality of π^* is to view the regime-change game as a binary-action, supermodular potential game (Monderer and Shapley, 1996; Sandholm, 2001) with a convex potential. This approach is contemporaneously and independently developed by Morris et al. (2020). Specific to our framework, the potential function $\Phi(A, \theta)$ of the mass of attacking players A and the state θ takes the value of

$$\Phi(A, \theta) = \begin{cases} 1 - c - \theta & \text{if } \theta < 1 \\ -c & \text{if } \theta \geq 1 \end{cases},$$

when all agents attack, i.e. $A = 1$. The designer's optimal outcome is to preserve the status quo for states beyond θ^*, characterized as the solution to

$$\int_{\theta^*}^{\bar{\theta}} \Phi(1, \theta) dF(\theta) = 0,$$

if a solution exists, which is exactly condition (3). This is the continuum-of-players counterpart of the optimal threshold derived in Morris et al. (2020) in a finite-player binary-action supermodular setting which covers the case of finite-player regime change games. Building on this insight, Morris et al. (2022a) consider a continuum-player regime change game and show that a global-game-like information policy minimizes the risk of regime change in the adversarial equilibrium.

3.3 On optimal local obfuscator

We offer a few remarks on the optimal information structure π^*.

Perfect coordination. Agents coordinate perfectly under π^*. Despite the fundamental and belief 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 (2022), 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. As it will soon be clear in Section 3.4, the perfect coordination implication is not robust when the designer is allowed to manipulate finitely many levels of agents' strategic reasoning.

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 her 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.

Possible multiplicity of optimum. 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 goes forward. Indeed, one can gain some intuition from condition (6). Imagine that the infection was not monotone in the first $k-1$ round, and some positive mass of states above $\inf \cup_{j=0}^{k-1} \Theta_j$ has not been infected. Then it is optimal to infect the highest not-yet-infected states as it maximizes the conditional expectation $\mathbb{E}[\theta|\theta \in \Theta_k]$ and therefore the mass of kth round infection. This will allow it to make up for "mistakes" in previous rounds. It turns out that in some cases, the total mass of infection can catch up in (infinitely many) subsequent rounds of IESDS, but in some other cases (e.g., infecting a positive mass of states below θ^* in the first k rounds of IESDS), it cannot.

Figure 2 presents a numerical example. The top panel represents the monotone state infection under π^*, whereas the middle and the bottom panels represent some

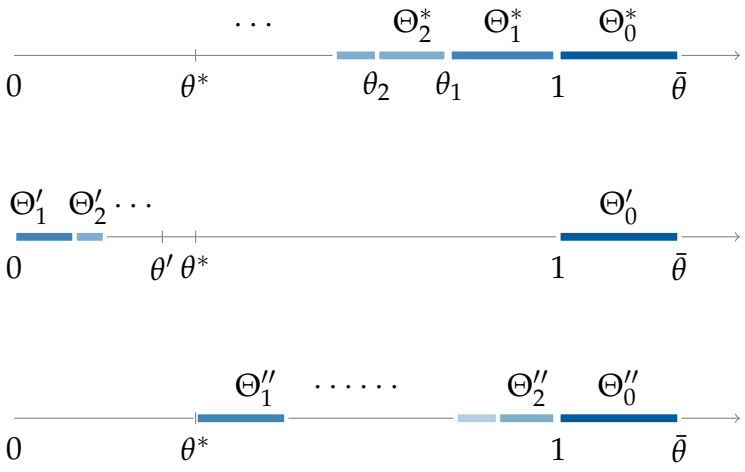

Figure 2: In this example, $\bar{\theta} = 1.1$, $c = 1/6$ and F is the uniform distribution. Each panel corresponds to an order of state infection through its process of IESDS, and the sets of states being infected in each round are labeled in order. For example, in the top panel, states in Θ_k^* are infected in the kth round of IESDS. In the top and bottom panels, the set of surviving states is $(\theta^*, \bar{\theta}]$ where $\theta^* \approx 0.586$. The probability of the regime's survival is then approximately $(1.1 - 0.586)/1.1 \approx 0.467$. In the middle panel, the largest possible set of surviving states is $[1, \bar{\theta}] \cup [0, \theta']$ where $\theta' \approx 0.020$, characterized by $[\int_1^{\bar{\theta}} f(\theta)d\theta + \int_0^{\theta'} \theta f(\theta)d\theta] / \int_0^{\theta'} (1-\theta)f(\theta)d\theta = (1-c)/c$. The probability of the regime's survival is then bounded above by approximately $(0.1 + 0.020)/1.1 \approx 0.109$.

non-monotone state infection. The bottom panel corresponds to a "one-shot small deviation" from the optimal local obfuscator: in the first round: $\Theta_1'' = (\theta^*, \theta'']$ where θ'' is chosen to balance the Bayesian constraint, $c \int_1^{\bar{\theta}} f(\theta)d\theta = (1-c) \int_{\theta_*}^{\theta''} (1-\theta)f(\theta)d\theta$. The resulting measure of Θ_k'' is strictly less than Θ_k for all $k \geq 1$, but the union up to $k = \infty$ remains the same: $\cup_{k=1}^{\infty} \Theta_k'' = \cup_{k=0}^{\infty} \Theta_k = [\theta^*, \bar{\theta}]$. The middle panel corresponds to a non-monotone infection IESDS, which is strictly suboptimal.

Non-deterministic information structure. We briefly discuss here why π^* remains optimal when more general information policies are feasible and leave the formal argument to the proof. Consider a random or non-deterministic information structure, where the measure of realized signals given some state θ is uncertain. Whether the information structure is random or not, the measure of newly created credit for each round of IESDS is computed in a way analogous to taking expectation: determine $f(\theta)$ times the probability measure of signals not used for the previous rounds, and then integrate this term over the set of infected states in the current round. The calculation is *linear* in signal distribution, and thus including aggregate uncertainty will not, compared to our optimal deterministic information structure π^*, cre-

ate more credit in any round. Alternatively and more heuristically, one may imagine re-labeling 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^*. Therefore, compared to deterministic information structures, the ability to further complicate the signals yields no extra leverage for the information designer.

More general settings. The main argument, and hence our qualitative results, generalize to a richer set of designer's payoff functions such that her payoff is increasing in θ and decreasing in A, conditional on the regime surviving, while constant and always below the former payoff conditional on regime change. Suppose that, if the regime falls, the designer has a constant payoff r; if the regime survives, the designer has payoff function $g(A, \theta)$ where $\theta \in [0, \bar{\theta}]$ is the regime's strength and $A \in [0, \theta]$ is the measure of attacking agents (which is insufficient to overthrow the regime). Suppose that

Assumption 1. *$g(A, \theta)$ is decreasing in A, increasing in θ, and that $g(A, \theta)$ is always weakly larger than r.*

A simple example is $g(\theta, A) = \theta - A$ and $r = 0$. The implicit assumption is that the regime has limited liability. This is natural in many classic settings, such as bank run, currency attack, and democratization after regime-change, where coordination is relevant.[10] Then our results extend to this environment as the optimal local obfuscator specifies a set of states that both maximizes the probability measure of the regime's survival and maximizes the expectation of $g(A, \theta)$ on this set. Formally,

Proposition 2. *Suppose that the designer's payoff is*

$$\begin{cases} g(\theta, A) & \text{if } A \leq \theta \\ r & \text{if } A > \theta \end{cases},$$

and $g(\cdot), r$ satisfy Assumption 1. Then the designer's optimum is achieved by the local obfuscator π^* specified in Theorem 1.

[10] However, other configurations of the designer's payoff, as well as different settings on agents' payoff from the regime-change model, e.g. a linear payoff of attacking $A - \theta - c$, suggest that our local obfuscator may not be optimal. Examples are available upon request.

The proof is simple, and goes as follows. Consider an arbitrary information structure $\pi' \in \Pi$. The designer's payoff can be written as

$$v(\pi') \equiv \int_{\Theta'} g(A,\theta) dF(\theta) + \int_{\Theta \setminus \Theta'} r dF(\theta),$$

where Θ' is the set of surviving states under π'. Take a variation of $v(\pi')$

$$\hat{v}(\pi') \equiv \int_{\Theta'} g(0,\theta) dF(\theta) + \int_{\Theta \setminus \Theta'} r dF(\theta) \geq v(\pi')$$

and consider an alternative optimization problem for the designer, i.e. to maximize $\hat{v}(\pi')$ by choosing π'. Since $g(0,\theta)$ is increasing in θ, the designer's optimum would be achieved by π' which both maximizes the probability measure of Θ' and includes in Θ' the highest states in Θ. According to Theorem 1, this means that $\pi' = \pi^*$ solves the alternative problem. Since $\hat{v}(\pi') \geq v(\pi')$ for all π' and equality is obtained when $\pi' = \pi^*$, we know that $\pi' = \pi^*$ also solves the designer's original problem of maximizing $v(\pi')$.

To concentrate on the main economic insights in a regime-change context, we will maintain the current stylized payoff structure for the rest of the paper.

3.4 Level-K Obfuscation

This section studies a natural way to extend our analysis to an environment where the designer faces a constraint in her capacity of manipulating information. In particular, suppose that the signal space is now finite and contains only $K \in \mathbb{N}^+$ distinct elements. We show that (i) there is a *unique* optimal information structure which exhibits local obfuscation, and (ii), perhaps surprisingly, it is optimal to induce *imperfect coordination* among agents.

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 is 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.*

These two cases are mutually exclusive.

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

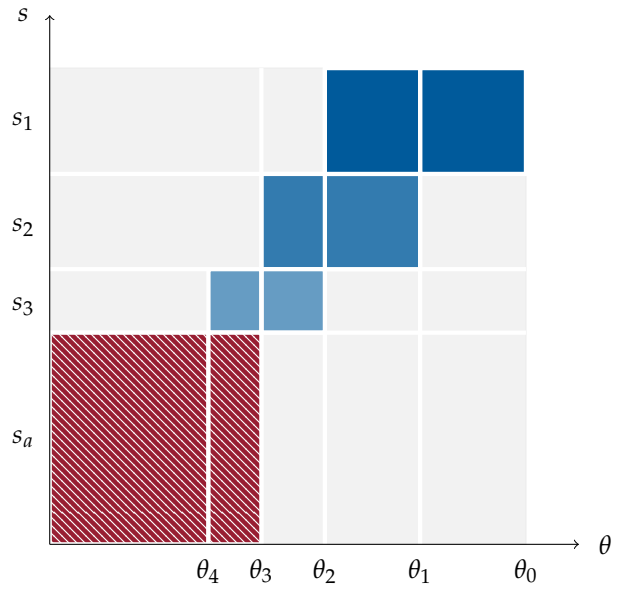

Figure 3: Illustration of level-4 local obfuscation. The horizontal axis represents states, and the vertical axis represents signals and their face values. We use differential shades to distinguish information sets following different signals. The regime persists if and only if $\theta \geq \theta_4$,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

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 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, she would have used it to create another round of IESDS and save more states from attack, but the constraint deprives her 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. This is perhaps surprising because recent studies, e.g., Inostroza and Pavan (2022) and Morris et al. (2020), show that perfect coordination typically appears as an equilibrium outcome under optimal adversarial information design in binary-action supermodular games.

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 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.

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 3.5, we will highlight this advantage of local obfuscation via comparative static analysis.

3.5 Public vs Private 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 3.4: public disclosure.[11] We study the difference

[11] 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-

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. We apply the result to understand the transition of the dominant model of informational autocrats in the 21st century.

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

$$c = \frac{F(1) - F(\theta^\dagger)}{1 - F(\theta^\dagger)}. \tag{7}$$

The right-hand side of equation (7) is an agent's expected benefit if he 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.

Public vs private signals. We now 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 3. *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 $\liminf_{n\to\infty} F_n(\theta_n^\dagger) - F_n(\theta_n^*) > 0$.*

public) signals when $\theta \in [\theta_2, \theta_1)$.

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.

3.5.1 On Commitment and Informational Autocrats

Comparing local obfuscation and public disclosure helps explain the transition of dominant autocratic models in the digital era. In the twentieth century, the most popular dictatorship model was based on fear, and rulers combined public propaganda with violent repression, terrorizing their citizens. Recently, a growing number of au-

tocrats have abandoned this formula and switched to a less-bloody way based on spreading misinformation and creating a gap in political perception among citizens (Guriev and Treisman 2019). A simple extension of our model can rationalize this trend.

The central element of our explanation is that the regime needs to combine information manipulation with physical repression to maintain its committed policy time-consistent. Recall that for some parameter values, the threshold $\theta^* > 0$ in the optimal local obfuscator. If the realized state is below θ^*, the regime is supposed to send signal s_a, resulting in all agents' attacks and a regime change. However, in this case, the regime, which *designs* and *implements* the information structure, has a strong incentive to deviate from the committed information structure to avoid a regime change. This time-inconsistency issue is common in the literature of information design (see, e.g, Kamenica and Gentzkow (2011)), but in the setting of informational autocrats, it is hard to believe that anything in practice can stop the regime from an ex post profitable deviation. The regime suffers from a lack of commitment power because agents do not believe that the regime's ex ante optimal information policy π^* will ever be implemented.

One solution is to allow the regime to endogenously choose agents' attack cost such that the ex ante optimal information structure remains optimal for every possible θ. In other words, the regime can use costly investment in physical repression as a commitment device to make its ex ante optimal information structure time-consistent. According to this view, (i) the regime's investment in physical repression serves not only as a threat to attacking citizens but as an instrument to build its commitment in information disclosure,[12] and (ii) a regime collapses when it does not have enough financial resources to maintain a sufficiently high level of attack cost to make its information policy credible.

To fix the idea, suppose that the designer chooses an information structure as well as the agents' attack cost, through recruiting more police officers, building more prisons, etc., so that the information policy is *credible*, i.e., the information disclosure must be time-consistent for the designer in every state $\theta \in \Theta$. For simplicity, suppose that the spending on physical repression is proportional to the target attack cost; i.e., κc where $\kappa > 0$, and the regime aims to minimize the expenditure on physical repression investment subject to its information policy being credible.[13]

[12] Also see Chen and Xu (2017), who emphasize that the credibility of an authoritarian regime's information disclosure results from its flexibility of making policy concessions.

[13] Therefore, the regime's problem is essentially a combination between a contract design, which specifies a punishment contingent on the agent's action and the regime status, and an information

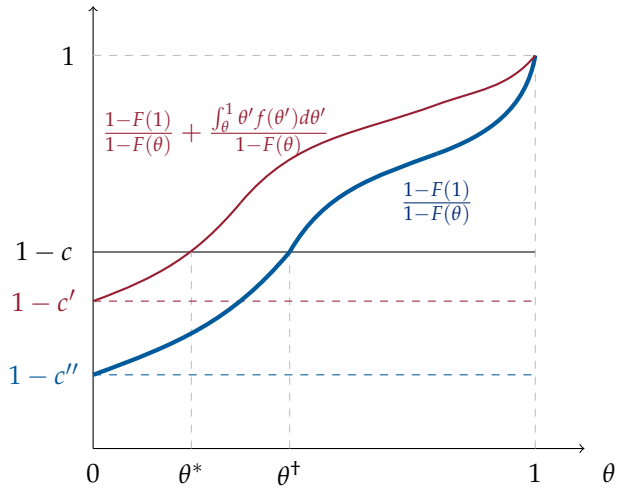

Figure 4: The thick blue curve represents the right-hand side of equation (22), and the thin red curve represents the right-hand side of equation (23). When c increases, the black curve $1-c$ is shifted down for every θ, and so both θ^* and $\theta^\dagger$ decrease. When the cost is greater than c' (or c''), the regime persists under optimal local obfuscation (public disclosure) whenever $\theta > 0$.

As demonstrated in Figure 4, when public propaganda is the only way of information manipulation, and agents' attack cost is c, the regime fails if $\theta < \theta^\dagger$ under the optimal public information design. The information structure is not credible because, in this case, the regime will renege to avoid the attack. The regime needs to invest heavily in physical repression to increase agents' attack cost to $c'' = F(1)$ to make its public information policy time-consistent. In this case, the public information structure is credible but trivial: it is *completely uninformative*. The regime needs to maintain absolute censorship, which requires placing barriers between its citizens and the rest of the world. The regime's stability is solely built on the threat of iron-hand response to agents' anti-regime actions, which requires a sufficiently high expenditure. This prediction is roughly consistent with the policies adopted by most twentieth-century totalitarians.

The revolution of information and communication technology, especially the emergence of social media, makes divide-and-conquer information manipulation available in the 21st century.[14] Suppose that the regime can take advantage of the social network penetration and employ the optimal local obfuscator by building an army of "disinformation architects" (Ong and Cabanes 2018) or "50-cent gang" (King et al.

design. See, e.g., Halac et al. (2021), Halac et al. (2022), and Morris et al. (2022b) for the combination of contract and information design in other supermodular settings.

[14] Also see Edmond and Lu (2021), who demonstrate that the emergence of social media helps politicians to manipulate information.

2017), and save some expense of physical repression investment, as the required level of attack cost drops to $c' = c'' - \int_0^1 \theta' f(\theta') d\theta'$ as shown in Figure 4. In this case, the regime provides divided but *partially informative signals* about the state and *strictly benefits* from communication. Also, the minimum expenditure to maintain its information policy time-consistent decreases by $\kappa \int_0^1 \theta' f(\theta') d\theta' > 0$. Remarkably, expanding the designer's feasible set of information policies eases the attack cost requirement making information disclosure time-consistency, relieving the usual concern of commitment assumption. This naturally reflects that more sophisticated information manipulation substitutes high-powered stick-and-carrot incentives in defending the regime.

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, an optimal persuasion scheme takes a simple and intuitive form. The information designer randomizes between honesty and deceit, which takes the particular form of local obfuscation. We believe that our stylized framework can be enriched to build a research agenda on many related topics, including competitive information designers, dynamic persuasion and communication among agents.

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.

Chen, J. and Y. Xu (2017). Information manipulation and reform in authoritarian regimes. *Political Science Research and Methods 5*(1), 163–178.

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-k 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.

Edmond, C. and Y. K. Lu (2021). Creating confusion. *Journal of Economic Theory 191*, 105145.

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.

Halac, M., E. Lipnowski, and D. Rappoport (2022). Addressing strategic uncertainty with incentives and information. In *AEA Papers and Proceedings*, Volume 112, pp. 431–37.

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

Hoshino, T. (2022). Multi-agent persuasion: Leveraging strategic uncertainty. *International Economic Review 63*(2), 755–776.

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

Inostroza, N. and A. Pavan (2022). Adversarial coordination and public information design. working paper.

Kajii, A. and S. Morris (1997). The robustness of equilibria to incomplete information. *Econometrica: Journal of the Econometric Society*, 1283–1309.

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

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

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.

Monderer, D. and L. S. Shapley (1996). *Games and economic behavior 14*(1), 124–143.

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 (2020). Implementation via information design in binary-action supermodular games. working paper.

Morris, S., D. Oyama, and S. Takahashi (2022a). Implementation via information design using global games. *Available at SSRN*.

Morris, S., D. Oyama, and S. Takahashi (2022b). On the joint design of information and transfers. *Available at SSRN*.

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.

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

Prescott, E. C. and R. M. Townsend (1984). Pareto optima and competitive equilibria with adverse selection and moral hazard. *Econometrica: journal of the econometric society*, 21–45.

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

Sandholm, W. H. (2001). Potential games with continuous player sets. *Journal of Economic theory 97*(1), 81–108.

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

Sun, Y. (2006). The exact law of large numbers via fubini extension and characterization of insurable risks. *Journal of Economic Theory 126*, 31–69.

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

Van Zandt, T. (2010). Interim Bayesian Nash equilibrium on universal type spaces for supermodular games. *Journal of Economic Theory 145*(1), 249–263.

A Proofs of Main Results

A.1 Proof of Theorem 1

We first show that π^* achieves the designer's optimum among all anonymous and deterministic information structures. We extend our argument to allow for anonymous and possibly non-deterministic information structures in Appendix B.3.

Consider an arbitrary information structure which is anonymous and deterministic. We begin by defining two useful series through agents' iterated reasoning, assuming that agents are adversarial against the regime. Series $\{S_k\}_{k=0}^{\infty}$ contains the signal sets which the agents refrain from attacking after the kth round of *iterated elimination of strictly dominated strategies (IESDS)*. Series $\{\Theta_k\}_{k=0}^{\infty}$ satisfies the following condition: $\cup_{n=0}^{k}\Theta_n$ contains the states in which the status quo persists after the kth round of IESDS.

Definition of series $\{S_k\}_{k=0}^{\infty}$ and $\{\Theta_k\}_{k=0}^{\infty}$. Define $S_0 = \varnothing$, $\Theta_0 = [1, \bar{\theta}]$. For $k = 1, 2, ... + \infty$, define recursively

$$\begin{aligned} S_k &= \{s \in S | \pi(\Theta \backslash \cup_{n=0}^{k-1} \Theta_n | s) \leq c\} \\ &= \{s \in S | \pi(\cup_{n=0}^{k-1} \Theta_n | s) \geq 1 - c\}, \\ \Theta_k &= \{\theta \in \Theta | \pi(S \backslash S_k | \theta) \leq \theta\} \backslash \cup_{n=0}^{k-1} \Theta_n \end{aligned}$$

$$= \quad \{\theta \in \Theta | \pi(S_k|\theta) \geq 1-\theta\} \setminus \cup_{n=0}^{k-1} \Theta_n.$$

It is straightforward that $\pi(\cup_{n=0}^{k-1}\Theta_n|s)$ is weakly increasing in k for any s; therefore, $S_k \supseteq S_{k-1}$ for every k, $\lim_{k\to\infty} S_k$ exists, denote this limit by $S^* \subseteq S$; also denote $\Theta^* = \cup_{k=0}^{\infty}\Theta_k$.

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

For ease of exposition, for $i \in [0,1]$, $k = 0, 1, \cdots$, define a strategy

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

When all agents use strategy $a^k(s)$, it means that every agent $i \in [0,1]$ attacks if and only if his signal is not in S_{k-1}.

The definition of $\{S_k\}_{k=0}^{\infty}$ and $\{\Theta_k\}_{k=0}^{\infty}$ guarantees that if every agent $j \neq i$ plays strategy $a^k(s)$, then agent i weakly prefers not to attack when receiving a signal $s \in S_{k+1}$.[15] To see this, consider a symmetric strategy profile where every agent attacks if his signal is not in S_0 (i.e. he attacks regardless of his signal). By definition, Θ_0 is the set of the states which satisfy $\theta \geq \pi(S\setminus S_0|\theta)$. The right-hand side of the inequality is the mass of attack, assuming every agent plays $a^0(s)$. Therefore, Θ_0 is the set of states which persist when every agent plays $a^0(s)$. By the definition of S_1, for every $s \in S_1$, s induces the following posterior: the probability that the true state is not in Θ_0 is smaller than or equal to c, i.e. $\pi(\Theta\setminus\Theta_1|s) \leq c$. The left-hand side of the inequality is the expected payoff of attacking, before cost, conditional on the signal being s and every other agent playing $a^0(s)$. Therefore, $\pi(\Theta\setminus\Theta_0|s) \leq c$ is exactly the condition such that an agent receiving $s \in S_1$ is not going to attack when every other agent plays $a^0(s)$.

For $k = 1, 2, \cdots$, as $S_k \supseteq S_{k-1}$, it is straightforward that a strategy profile with all agents playing $a^k(s)$ induces a weakly smaller mass of attack than a strategy profile with all agents playing $a^{k-1}(s)$. By Lemma 4, given that an agent i with a signal realization in S_k weakly prefers not to attack if all other agents play $a^{k-1}(s)$, he also

[15] Recall that we approximate the regime's expected probability of survival, under any optimal information policy of the designer, by its supremum. Equivalently, one may regard the supremum as the maximized probability of regime's survival in any Bayesian Nash equilibrium where (1) the agents coordinate on a strategy profile such that the largest measure of agents attacks, but (2) each agent does *not* attack when indifferent.

weakly prefers not to attack if all other agents play $a^k(s)$.

By definition, $\cup_{n=0}^{k}\Theta_n$ is the set of the states which satisfy $\theta \geq \pi(S\backslash S_k|\theta)$. The right-hand side of the inequality is the mass of attack if every agent plays $a^k(s)$. Therefore, $\cup_{n=0}^{k}\Theta_n$ is the set of states which persist when every agent plays $a^k(s)$. By the definition of S_{k+1}, for every $s \in S_{k+1}$, s induces the following posterior: the probability that the true state is not in $\cup_{n=0}^{k}\Theta_n$ is smaller than or equal to c, i.e. $\pi(\Theta\backslash \cup_{n=0}^{k}\Theta_n|s) \leq c$. The left-hand side of the inequality is the payoff of attacking, before cost, conditional on the signal being s and every other agent playing $a^k(s)$. Therefore, $\pi(\Theta\backslash \cup_{n=1}^{k}\Theta_n|s) \leq c$ is exactly the condition such that an agent receiving $s \in S_{k+1}$ is not going to attack when every other agent plays $a^k(s)$.

The definition of $\{S_k\}_{k=0}^{\infty}$ and $\{\Theta_k\}_{k=0}^{\infty}$ also guarantees that if, for every $i \in [0,1]$, agent i plays $a^k(s)$, then states in $\cup_{n=0}^{k}\Theta_n$ persist. This is straightforward as Θ_k contains the states which are stronger than the measure of attack, i.e. the measure of agents receiving signals not in S_k.

Now we are ready to characterize the regime's persistence.

Lemma 1. *Under π^*, a necessary and sufficient condition for the regime to persist in equilibrium is $\theta \in \Theta^*$.*

Proof. We first show the sufficiency. If $\theta \in \Theta^*$, there exists k such that $\theta \in \Theta_k$ and $\theta \notin \Theta_l$ for $l = 0, 1, ..., k-1$. We show that the regime persists for any $k = 0, 1, ...$ First suppose that every agent attacks if and only if his signal is in S. By the definition of Θ_0 and S_1, an individual agent whose signal is in S_1 prefers to deviate and not attack. As all agents choose to not to attack when receiving a signal in S_1, every $\theta \in \Theta_0 \cup \Theta_1$ persists under information policy $\pi(\cdot|\theta)$. By a similar argument, suppose every agent attacks if and only if his signal is in $S\backslash S_1$; then an individual agent whose signal is in S_2 prefers to deviate and not attack. As all agents choose not to attack when receiving a signal in S_2, every $\theta \in \Theta_0 \cup \Theta_1 \cup \Theta_2$ persists. The rest of the proof follows by mathematical induction.

We prove the necessity by contrapositive. First, by the above construction, every agent attacks if and only if he receives a signal not in S^*. Then by the definition of Θ^*, for every state θ not in Θ^*, the designer sends a signal in S^* with probability less than $1-\theta$; otherwise θ is in Θ^*. Thus, every state θ not in Θ^* is attacked by a mass greater than θ and eventually fails, the agents get $1-c$ for sure and therefore have no incentive to deviate. 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*

$$c\int_{1}^{\bar{\theta}} f(\theta)d\theta + \int_{\theta^{*\prime}}^{1}(\theta + c - 1)f(\theta)d\theta = 0, \tag{8}$$

or equals 0 *when (8) has no solution.*

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

Consider round k of IESDS. For every $s \in S_k$, an individual agent i receiving s prefers not to attack even if every other agent j plays a^{k-1} (i.e. attacks if j's signal is not in S_{k-1}). That is to say, i expects that the probability of regime change is smaller than or equal to c when he receives $s \in S_k$. Note that given the above strategy profile, by the definition of $\{\Theta_k\}_{k=0}^{\infty}$, the regime persists if and only if the true state is in $\cup_{n=0}^{k-1}\Theta_n$. Thus a necessary condition for agent i not to attack when receiving any signal in $S_k \backslash S_{k-1}$ is

$$\begin{aligned}
& c \geq \frac{\int_{\Theta\backslash(\cup_{n=0}^{k-1}\Theta_n)} f(\theta)\pi(S_k\backslash S_{k-1}|\theta)d\theta}{\int_{\cup_{n=0}^{k-1}\Theta_n} f(\theta)\pi(S_k\backslash S_{k-1}|\theta)d\theta + \int_{\Theta\backslash(\cup_{n=0}^{k-1}\Theta_n)} f(\theta)\pi(S_k\backslash S_{k-1}|\theta)d\theta} \\
\Rightarrow \quad & c \geq \frac{\int_{(\cup_{n=0}^{k}\Theta_n)\backslash(\cup_{n=0}^{k-1}\Theta_n)} f(\theta)\pi(S_k\backslash S_{k-1}|\theta)d\theta}{\int_{\cup_{n=0}^{k-1}\Theta_n} f(\theta)\pi(S_k\backslash S_{k-1}|\theta)d\theta + \int_{(\cup_{n=0}^{k}\Theta_n)\backslash(\cup_{n=0}^{k-1}\Theta_n)} f(\theta)\pi(S_k\backslash S_{k-1}|\theta)d\theta} \\
\Leftrightarrow \quad & c\sum_{n=0}^{k-1}\int_{\Theta_n} f(\theta)\pi(S_k\backslash S_{k-1}|\theta)d\theta \geq (1-c)\sum_{n=k}^{k}\int_{\Theta_n} f(\theta)\pi(S_k\backslash S_{k-1}|\theta)d\theta.
\end{aligned} \tag{9}$$

The right-hand side of the first inequality measures the probability with which agent i believes the regime will fall, conditional on i receiving a signal in $S_k\backslash S_{k-1}$ and the other agents attacking if and only if their signal is not in S_{k-1}. To see this, observe that the numerator (and the second term of the denominator) measure the probability that agent i receives a signal in $S_k\backslash S_{k-1}$ and the regime is going to be overthrown (i.e. $\theta \in \Theta\backslash \cup_{n=0}^{k-1}\Theta_n$), while the first term of the denominator measures the probability that agent i receives such a signal but the regime will persist (i.e. $\theta \in \cup_{n=0}^{k-1}\Theta_n$). Since i's benefit from regime change is 1, this conditional probability equals i's expected benefit from attacking. Therefore, the first inequality states that an agent receiving a signal in $S_k\backslash S_{k-1}$ finds the cost of attack exceeding the expected benefit and there-

fore does not attack, assuming every other agent attacks if and only if receiving a signal not in S_{k-1}. This is exactly the agent's incentive compatibility constraint in the kth round of IESDS, according to the definition of S_k. The second inequality results from the fact that $(\Theta \setminus \cup_{n=0}^{k-1} \Theta_n) \supseteq (\cup_{n=0}^{k} \Theta_n \setminus \cup_{n=0}^{k-1} \Theta_n)$. Finally, we obtain the third inequality since all the elements in $\{\Theta_k\}_{k=0}^{\infty}$ are mutually exclusive.

A clear correspondence exists between constraints (4) and (9) , our "credit-discredit budget constraint" in the main text. Specifically, $\sum_{n=k}^{k} \int_{\Theta_n} f(\theta)\pi(S_k \setminus S_{k-1}|\theta)d\theta$ corresponds to D_k while $\sum_{n=0}^{k-1} \int_{\Theta_n} f(\theta)\pi(S_k \setminus S_{k-1}|\theta)d\theta$ corresponds to C_{k-1}. The difference is that (4) only holds for an information policy which coincides with π^* in the first $k-1$ rounds of infection, while (9) is a general version for an arbitrary information policy.

Next, consider all the previous rounds of the IESDS; each round corresponds to a necessary condition analogous to (9). For instance, consider round p of IESDS, $p < k$, and the corresponding condition is

$$c \sum_{n=0}^{p-1} \int_{\Theta_n} f(\theta)\pi(S_p \setminus S_{p-1}|\theta)d\theta \geq (1-c) \sum_{n=p}^{k} \int_{\Theta_n} f(\theta)\pi(S_p \setminus S_{p-1}|\theta)d\theta.$$

This is a necessary condition for an agent receiving a signal in $S_p \setminus S_{p-1}$ to find the cost of attack exceeding the expected payoff and therefore not attack, assuming every other agent attacks if and only if receiving a signal not in S_{p-1}.

Sum these conditions up for $p = 1, 2, \cdots, k$, then a necessary condition for the regime to persist in states $\cup_{n=0}^{k} \Theta_n$ is

$$\begin{aligned}
& c \sum_{p=1}^{k} \sum_{n=0}^{p-1} \int_{\Theta_n} f(\theta)\pi(S_p \setminus S_{p-1}|\theta)d\theta \geq (1-c) \sum_{p=1}^{k} \sum_{n=p}^{k} \int_{\Theta_n} f(\theta)\pi(S_p \setminus S_{p-1}|\theta)d\theta \\
\Leftrightarrow \quad & c \sum_{n=0}^{k-1} \int_{\Theta_n} f(\theta)\pi(S_k \setminus S_n|\theta)d\theta \geq (1-c) \sum_{n=1}^{k} \int_{\Theta_n} f(\theta)\pi(S_n|\theta)d\theta \\
\Leftrightarrow \quad & c \int_{\cup_{n=0}^{k-1}\Theta_n} f(\theta)\pi(S_k \setminus S_{K(\theta)}|\theta)d\theta \geq (1-c) \int_{\cup_{n=1}^{k}\Theta_n} f(\theta)\pi(S_{K(\theta)}|\theta)d\theta. \qquad (10)
\end{aligned}$$

The second inequality is obtained by changing the order of summation, while the third inequality results from the definition of $K(\theta)$. Note that by the definition of $\{S_k\}_{k=0}^{\infty}$ and $\{\Theta_k\}_{k=0}^{\infty}$, for $n = 0, 1, \cdots, k$ and for every $\theta \in \Theta_n$, $\pi(S_n|\theta) \geq \min\{0, 1-\theta\}$ and $\pi(S_k \setminus S_n|\theta) \leq \max\{1, \theta\}$. Then the above equation yields

$$c(\int_{\Theta_0} f(\theta)d\theta + \int_{\cup_{n=1}^{k-1}\Theta_n} \theta f(\theta)d\theta) \geq (1-c)\int_{\cup_{n=1}^{k}\Theta_n} (1-\theta)f(\theta)d\theta.$$

Now we are in the position to identify an upper bound of the ex ante probability that the regime persists, $\int_{\Theta^*} f(\theta)d\theta$. Note that $\Theta^* = \cup_{n=0}^{+\infty}\Theta_n$ and $\Theta_0 = [1, \bar{\theta}]$. Therefore, one way to identify a certain superset of the designer's optimum, in terms of states that survive, is to identify a Θ^* – with a slight abuse of notation – to maximize $\int_{\Theta^*\setminus\Theta_0} f(\theta)d\theta$ under the following constraint

$$\begin{aligned} & c\left(\int_{\Theta_0} f(\theta)d\theta + \int_{\Theta^*\setminus\Theta_0} \theta f(\theta)d\theta\right) \geq (1-c)\int_{\Theta^*\setminus\Theta_0} (1-\theta)f(\theta)d\theta \\ \Leftrightarrow \quad & \int_{\Theta^*\setminus\Theta_0} (1-\theta)f(\theta)d\theta \leq \frac{c}{1-c}\left(\int_{\Theta^*\setminus\Theta_0} \theta f(\theta)d\theta + \int_{\Theta_0} f(\theta)d\theta\right). \end{aligned} \tag{11}$$

We will call this constrained maximization the "relaxed problem."

We assert that one desired $\Theta^* \supset \Theta_0$ which solves the relaxed problem takes the form of $\Theta^* = [\theta', \bar{\theta}]$ for some θ'. That is, to solve the relaxed problem, it is always optimal to include in Θ^* only the strongest states. To see this, rewrite (11) as

$$\int_{\Theta^*\setminus\Theta_0} (1 - \frac{1}{1-c}\theta)f(\theta)d\theta \leq \frac{c}{1-c}\int_{\Theta_0} f(\theta)d\theta \tag{12}$$

and note that the right-hand side is a constant. Therefore (1) it only weakly relaxes (12) to include in Θ^* all states θ such that $\frac{1}{1-c}\theta \geq 1$; (2) For all θ such that $\frac{1}{1-c}\theta < 1$, to increase the maximand $\int_{\Theta^*\setminus\Theta_0} f(\theta)d\theta$ by an infinitesimal d by including θ into Θ^*, the left-hand side of (12) increases by $\frac{d}{f(\theta)}(1 - \frac{1}{1-c}\theta)f(\theta) = d(1 - \frac{1}{1-c}\theta)$, which is strictly decreasing in θ. (1) and (2) imply that one solution to the relaxed problem includes only the strongest states. Furthermore, every solution only differs from this particular Θ^* by at most a zero-measure set of states.

Focusing on the specific form $\Theta^* = [\tilde{\theta}, \bar{\theta}]$ for some $\tilde{\theta}$, we can write (11) as

$$c\int_1^{\bar{\theta}} f(\theta)d\theta + \int_{\tilde{\theta}}^1 (\theta + c - 1)f(\theta)d\theta \geq 0. \tag{13}$$

Note that when $\tilde{\theta} = 1$, the above inequality is always satisfied; moreover, as $\tilde{\theta}$ decreases continuously from 1, the left-hand side of the inequality above increases continuously at first, then decreases continuously. Thus there exists a lowest $\tilde{\theta} \in \Theta = [0, \bar{\theta}]$ which satisfies the above inequality, denoted by $\theta^{*\prime}$. If $\theta^{*\prime}$ is strictly positive, it

solves

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

i.e. equation (8); otherwise $\theta^{*\prime} = 0$, meaning that every state persists at optimum of the designer's relaxed problem. This completes the proof. □

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

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 Θ_k, under π^* we have $\Theta_0 = [\theta_1, \theta_0]$ and $\Theta_k = [\theta_{k+1}, \theta_k)$ for $k = 1, 2, \cdots$.

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

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

Notably, θ^* indeed solves (8); as the solution is unique, we have $\theta^* = \theta^{*\prime}$, i.e. the measure of $\int_{\Theta^*} f(\theta)d\theta$ 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 $\theta^{*\prime} = 0$. □

A.2 Proof of Theorem 2

We prove that π_K is the unique optimal policy if the regime persists with probability less than 1, i.e. $\theta_k > 0$ for every $k \leq K$.

Consider an optimal policy π'_K which may differ from π_K, and note that an information policy using K signals can induce at most $K - 1$ (different) rounds of IESDS. Otherwise, after K rounds of IESDS, S_k already contains all the K different signals, which implies that an individual agent shall always not attack. That contradicts our assumption that the regime persists with probability less than 1. Hence suppose that π'_K induces $m \leq K - 1$ rounds of IESDS.

For arbitrary π and $k = 0, 1, ...$, define

$$D_k|\pi = \int_{\Theta_k} f(\theta)\pi(S_k|\theta)d\theta,$$

which is the (probability) measure of signals in S_k being sent by $\theta \in \Theta_k$. This measure of signal corresponds to the measure of agents that are not going to attack a regime in Θ_k after the kth round of IESDS. Similarly, for $k = 0, 1, \cdots$, $p = k+1, k+2, \cdots$, define

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

which is the (probability) measure of signals in $S_p \backslash S_{p-1}$ being sent by all $\theta \in \Theta_k$. This measure of signal corresponds to the measure of agents that, when the state is in Θ_k, are not going to attack after and only after the pth round of IESDS.

By the definition of π_K, we recursively specify a sequence of conditions regarding $\{C_{k,k+1}|\pi_K\}_{k=0}^{K-2}$, $\{D_k|\pi_K\}_{k=0}^{K-1}$:

$$\begin{aligned}
D_0|\pi_K &= 0 \\
D_1|\pi_K &= \frac{c}{1-c}C_{0,1}|\pi_K = \frac{c}{1-c}(1 - F(1)) \\
D_1|\pi_K + D_2|\pi_K &= \frac{c}{1-c}(C_{1,2}|\pi_K + C_{0,1}|\pi_K) \\
&\cdots \\
\sum_{p=1}^{K-1} D_p|\pi_K &= \frac{c}{1-c}\sum_{p=0}^{K-2} C_{p,p+1}|\pi_K. \qquad (14)
\end{aligned}$$

For π'_K, we specify a necessary condition for each round of IESDS using (10). Take the second round of IESDS as an example, i.e. let $k = 2$, according to (10), we have $c\sum_{n=0}^{1}\int_{\Theta_n} f(\theta)\pi(S_2 \backslash S_n|\theta)d\theta \geq (1-c)\sum_{n=1}^{2}\int_{\Theta_n} f(\theta)\pi(S_n|\theta)d\theta$, the left-hand side is $c(C_{1,2}|\pi'_K + C_{0,1}|\pi'_K + C_{0,2}|\pi'_K)$, the right-hand side is $(1-c)(D_1|\pi'_K + D_2|\pi'_K)$. Similarly, let $k = 1, 2, \cdots, m$, according to (10), we have

$$\begin{aligned}
D_0|\pi'_K &= 0 \\
D_1|\pi'_K &\leq \frac{c}{1-c}C_{0,1}|\pi'_K \\
D_1|\pi'_K + D_2|\pi'_K &\leq \frac{c}{1-c}(C_{1,2}|\pi'_K + C_{0,1}|\pi'_K + C_{0,2}|\pi'_K) \\
\sum_{p=1}^{3} D_p|\pi'_K &\leq \frac{c}{1-c}(C_{2,3}|\pi'_K + \sum_{p=2}^{3} C_{1,p}|\pi'_K + \sum_{p=1}^{3} C_{0,p}|\pi'_K)
\end{aligned}$$

$$\dots$$

$$\sum_{p=1}^{m} D_p|\pi'_K \leq \frac{c}{1-c}(C_{m-1,m}|\pi'_K + \sum_{p=m-1}^{m} C_{m-2,p}|\pi'_K + \dots + \sum_{p=1}^{m} C_{0,p}|\pi'_K). \quad (15)$$

The rest of our proof proceeds in the following steps.

Step 1. *We prove that* $C_{0,1}|\pi'_K = C_{0,1}|\pi_K$.

It is obvious that $C_{0,1}|\pi'_K$ cannot be greater than $C_{0,1}|\pi_K$. If $C_{0,1}|\pi'_K < C_{0,1}|\pi_K$, then by the second line of (14) and (15), $D_1|\pi'_K < D_1|\pi_K$. From the proof of Lemma 2, for all π and every state θ in $\Theta_1|\pi$, we have $\pi(S_1|\theta) \geq 1-\theta$ and $\pi(S_2|\theta) < \theta$. As $\frac{1-\theta}{\theta}$ is decreasing in θ and by the definition of π_K, $\pi \equiv \pi_K$ uniquely maximizes $C_{1,2}|\pi$, which implies $C_{1,2}|\pi'_K < C_{1,2}|\pi_K$. From the definition of π_K we also know $C_{0,2}|\pi'_K + C_{0,1}|\pi'_K \leq C_{0,1}|\pi_K$, then by the third line of (14) and (15), $D_1|\pi'_K + D_2|\pi'_K < D_1|\pi_K + D_2|\pi_K$, which, by an argument similar to the above, implies $C_{2,3}|\pi'_K + C_{1,3}|\pi'_K + C_{1,2}|\pi'_K < C_{2,3}|\pi_K + C_{1,2}|\pi_K$. From the definition of π_K we also know $C_{0,3}|\pi'_K + C_{0,2}|\pi'_K + C_{0,1}|\pi'_K \leq C_{0,1}|\pi_K$, then by the fourth line of (14) and (15), $D_1|\pi'_K + D_2|\pi'_K + D_3|\pi'_K < D_1|\pi_K + D_2|\pi_K + D_3|\pi_K, \cdots$. By mathematical induction, we have $\sum_{p=1}^{m} D_p|\pi'_K < \sum_{p=1}^{m} D_p|\pi_K$. Furthermore as $m \leq K-1$, $\sum_{p=1}^{m} D_p|\pi'_K < \sum_{p=1}^{K-1} D_p|\pi_K$.

From the proof of Lemma 2, for any information policy π and every state in $\Theta_n|\pi$, $n > 0$, we have $\pi(S_n|\theta) \geq 1-\theta$. That is to say, $\int_{\cup_{p=1}^{m}\Theta_p|\pi'_K}(1-\theta)f(\theta)d\theta \leq \sum_{p=1}^{m} D_p|\pi'_K < \sum_{p=1}^{m} D_p|\pi_K \leq \sum_{p=1}^{K-1} D_p|\pi_K$. Also from the proof of Lemma 2, to maximize $\int_{\hat{\Theta}\setminus\Theta_0} f(\theta)d\theta$ over (arbitrary) $\hat{\Theta} \subseteq \Theta$ while holding $\int_{\hat{\Theta}\setminus\Theta_0}(1-\theta)f(\theta)d\theta < \sum_{p=1}^{K-1} D_p|\pi_K$, it is optimal to include in $\hat{\Theta}$ only the strongest states, precisely, the optimal $\hat{\Theta}$ for the above problem either takes the form $[\theta', \tilde{\theta}]$ for some θ' or differs from it by a set of measure zero. As we have assumed that π'_K is an optimal policy, we must have $\cup_{p=0}^{m}\Theta_p|\pi'_K = [\theta', \bar{\theta}]$ for some θ', then $\int_{\theta'}^{1}(1-\theta)f(\theta)d\theta = \int_{\cup_{p=1}^{m}\Theta_p|\pi'_K}(1-\theta)f(\theta)d\theta$. By the definition of π_K and $\{\theta_p|\pi_K\}_{p=0}^{K}$

$$\begin{aligned}
&\int_{\cup_{p=1}^{m}\Theta_p|\pi'_K}(1-\theta)f(\theta)d\theta < \sum_{p=1}^{m} D_p|\pi_K \\
\Leftrightarrow\quad &\int_{\theta'}^{1}(1-\theta)f(\theta)d\theta < \int_{\theta_{m+1}|\pi_K}^{1}(1-\theta)f(\theta)d\theta \\
\Leftrightarrow\quad &\int_{\theta'}^{1} f(\theta)d\theta < \int_{\theta_{m+1}|\pi_K}^{1} f(\theta)d\theta \leq \int_{\theta_K|\pi_K}^{1} f(\theta)d\theta = \int_{\cup_{p=0}^{K-1}\Theta_p|\pi_K} f(\theta)d\theta.
\end{aligned}$$

That is to say, the information designer's ex ante probability of persistence under

π'_K is strictly smaller than under π_K, which contradicts our presumption that π'_K is an optimal policy. Thus, in every optimal design π'_K, $C_{0,1}|\pi'_K = C_{0,1}|\pi_K$, $D_1|\pi'_K = D_1|\pi_K$, $\Theta_1|\pi'_K = \Theta_1|\pi_K$; then we also have $C_{0,p}|\pi'_K = 0$ for $p = 2, 3, \cdots, m$.

Step 2. *We prove that in every optimal information structure π'_K, for $p = 1, 2, \cdots, m-1$, $C_{p,p+1}|\pi'_K = C_{p,p+1}|\pi_K$, $D_{p+1}|\pi'_K = D_{p+1}|\pi_K$.*

Similar to step 1, given $C_{0,1}|\pi'_K = C_{0,1}|\pi_K$, $C_{0,2}|\pi'_K = 0$, $D_1|\pi'_K = D_1|\pi_K$, and $\Theta_1|\pi'_K = \Theta_1|\pi_K$, suppose that $C_{1,2}|\pi'_K < C_{1,2}|\pi_K$; then by the third line of (14) and (15), $D_2|\pi'_K < D_2|\pi_K$, which (by an argument similar to Step 1) implies $C_{2,3}|\pi'_K < C_{2,3}|\pi_K$. From the definition of π_K we also know $C_{1,3}|\pi'_K + C_{1,2}|\pi'_K \leq C_{1,2}|\pi_K$, then by the fourth line of (14) and (15), $D_2|\pi'_K + D_3|\pi'_K < D_2|\pi_K + D_3|\pi_K$, which implies $C_{3,4}|\pi'_K + C_{2,4}|\pi'_K + C_{2,3}|\pi'_K < C_{3,4}|\pi_K + C_{2,3}|\pi_K$. From the definition of π_K we know $C_{1,4}|\pi'_K + C_{1,3}|\pi'_K + C_{1,2}|\pi'_K \leq C_{1,2}|\pi_K$; then by the fifth line of (14) and (15), $D_2|\pi'_K + D_3|\pi'_K + D_4|\pi'_K < D_2|\pi_K + D_3|\pi_K + D_4|\pi_K, \cdots$ Following a mathematical induction we have $\sum_{p=2}^{m} D_p|\pi'_K < \sum_{p=2}^{m} D_p|\pi_K \leq \sum_{p=2}^{K-1} D_p|\pi_K$. Note that by Step 1, $D_1|\pi'_K = D_1|\pi_K$; thus we have $\sum_{p=1}^{m} D_p|\pi'_K < \sum_{p=1}^{K-1} D_p|\pi_K$, 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, which contradicts our presumption that π'_K is an optimal policy. Thus, in every optimal design, $C_{1,2}|\pi'_K = C_{1,2}|\pi_K$, $D_2|\pi'_K = D_2|\pi_K$; then we also have $C_{1,p}|\pi'_K = 0$ for $p = 3, 4, \cdots, m$.

Iterate the above process for $C_{p,p+1}$, $p = 2, 3, \cdots, m-1$. By mathematical induction, we conclude that in every optimal design, for $p = 1, 2, \cdots, m-1$, $C_{p,p+1}|\pi'_K = C_{p,p+1}|\pi_K$, $D_{p+1}|\pi'_K = D_{p+1}|\pi_K$, $\Theta_{p+1}|\pi'_K = \Theta_{p+1}|\pi_K$, and $C_{p,q}|\pi'_K = C_{p,q}|\pi_K = 0$ for $q = p+2, p+3, \cdots, m$.

Step 3. *We prove that, if $\theta \in \Theta_0|\pi'_K$, the regime sends signals in $S_1|\pi'_K$ with probability 1; for $p = 1, 2, \cdots, m-1$, if every $\theta \in \Theta_p|\pi'_K$, 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 $1-\theta$; if $\theta \in \Theta_m|\pi'_K$, the regime sends some attack signals (i.e. signals given which an agent attacks in equilibrium) with probability θ and signals in $S_m|\pi'_K \backslash S_{m-1}|\pi'_K$ with probability $1-\theta$.*

This step is straightforward. First, by the definition of $C_{0,1}|\cdot$, as $C_{0,1}|\pi'_K = C_{0,1}|\pi_K = 1 - F(1)$, under information policy π'_K, the regime sends signals in $S_1|\pi'_K$ with probability 1 when the state is in $[1, \bar{\theta}]$.

Next, as we proved in Step 1 and 2, for $p = 1, 2, \cdots, m-1$, $C_{p,p+1}|\pi'_K = C_{p,p+1}|\pi_K$, $D_{p+1}|\pi'_K = D_{p+1}|\pi_K$, $\Theta_{p+1}|\pi'_K = \Theta_{p+1}|\pi_K$, and $C_{p,q}|\pi'_K = C_{p,q}|\pi_K = 0$ for $q = p+2, p+3, \cdots, m$. Then, by the definition of $C_{p,p+1}|\cdot$, $D_{p+1}|\cdot$, and $C_{p,q}|\cdot$, for $p = 1, 2, \cdots, m-1$ and for every $\theta \in \Theta_p|\pi'_K$, the regime sends signals in $S_{p+1}|\pi'_K \backslash S_p|\pi'_K$

with probability θ and signals in $S_p|\pi'_K$ with probability $1-\theta$; for every $\theta \in \Theta_m|\pi'_K$, the regime sends some attack signals with probability θ and signals in $S_m|\pi'_K$ with probability $1-\theta$.

Lastly, as states in $\Theta_1|\pi'_K$ send signals in $S_1|\pi'_K$ with probability $1-\theta$, and $D_1|\pi'_K = D_1|\pi_K = \frac{c}{1-c}C_{0,1}|\pi'_K = \frac{c}{1-c}C_{0,1}|\pi_K$, all the states not in $\Theta_0|\pi'_K \cup \Theta_1|\pi'_K$ send signals in $S_1|\pi'_K$ with probability 0; otherwise an individual agent receiving a signal in $S_1|\pi'_K$ shall attack. Then, as states in $\Theta_2|\pi'_K$ send signals in $S_2|\pi'_K$ with probability $1-\theta$ and signals in $S_1|\pi'_K$ with probability 0, they must send signals in $S_2|\pi'_K \backslash S_1|\pi'_K$ with probability $1-\theta$. By mathematical induction, for $p = 2, 3, \cdots, m-1$, as states in $\Theta_p|\pi'_K$ send signals in $S_p|\pi'_K \backslash S_{p-1}|\pi'_K$ with probability $1-\theta$, and $D_p|\pi'_K = D_p|\pi_K = \frac{c}{1-c}C_{p-1,p}|\pi'_K = \frac{c}{1-c}C_{p-1,p}|\pi_K$, all the states not in $\cup_{k=0}^{p}\Theta_k|\pi'_K$ send signals in $S_p|\pi'_K \backslash S_{p-1}|\pi'_K$ with probability 0; otherwise an individual agent receiving a signal in $S_p|\pi'_K \backslash S_{p-1}|\pi'_K$ shall attack. Note that, in previous rounds of the induction, we already proved that all the states not in $\cup_{k=0}^{p}\Theta_k|\pi'_K$ send signals in $S_{q-1}|\pi'_K \backslash S_{q-2}|\pi'_K$ with probability 0 for $q = 2, 3, \cdots, p$. Then, as states in $\Theta_{p+1}|\pi'_K$ send signals in $S_{p+1}|\pi'_K$ with probability $1-\theta$ and signals in $S_p|\pi'_K$ with probability 0, they must send signals in $S_{p+1}|\pi'_K \backslash S_p|\pi'_K$ with probability $1-\theta$. In conclusion, for $p = 1, 2, \cdots, m$ and for every $\theta \in \Theta_p|\pi'_K$, the regime sends 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 attack signal.*

Step 3 has established that, since $\Theta_p|\pi'_K$=$\Theta_p|\pi_K \; \forall p = 0, 1, \cdots, m$, 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 $\Theta_{p-1}|\pi'_K$ and $\Theta_p|\pi'_K$. 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 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 (along with signals in S_a) to induce IESDS. $K-2$ signals (along with signals in S_a) 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, we conclude that π'_K and π_K must be identical. This completes the proof.

A.3 Proofs in Section 3.5

Proof of Proposition 3. We first consider increasing c. Note that $\theta^\dagger$ and θ^* are characterized by

$$c(F(\bar{\theta}) - F(\theta^\dagger)) - (F(1) - F(\theta^\dagger)) = 0 \tag{16}$$

$$c(F(\bar{\theta}) - F(\theta^*)) - (F(1) - F(\theta^*)) + \int_{\theta^*}^{1} \theta f(\theta) d\theta = 0, \tag{17}$$

where (17) is a representation of (3). (16)-(17) 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 Appendix: Non-Deterministic Information Structure

Lemmas 2 and 3 establish the optimality of π^* among information structures that are deterministic across agents, and we now show that π^* remains optimal when non-deterministic and anonymous information structures are allowed.

B.1 Definition of Information Structure

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. If the information structure specifies a particular pair of signal space and signal generating probability measure $(S, (\pi(\cdot|\theta))_{\theta\in\Theta})$ defined as above, then the information structure is said to be *deterministic*. Meanwhile, a *non-deterministic* and anonymous information structure is defined by $(S, (\psi(\cdot|\theta))_{\theta\in\Theta})$. S is a complete separable metric space of signals, and for each $\theta \in \Theta$, $\psi(\cdot|\theta) \in$

$\Delta(\Delta(S))$, i.e. a signal generating probability measure $\pi \in \Delta(S)$ as above is drawn according to $\psi(\cdot|\theta)$, such that $\psi(U|\theta)$ is measurable in θ for each $U \in \mathcal{B}(\Delta(S))$. The joint probability measure $\phi \in \Delta(S \times \Delta(S) \times \Theta)$ is defined by $\phi(S' \times U \times \Theta') = \int_{\Theta'} \int_U \pi(S') d\psi(\pi|\theta) dF(\theta)$, and a regular conditional probability $\psi(\cdot|s) \in \Delta(\Delta(S) \times \Theta)$ is well defined. Then the interim expected payoff of player i observing signal s when other players follow strategy a is given by

$$\int_{\Delta(S)\times\Theta} u(a_i, \pi(\{s \in S | a(s) = 1\}|\theta), \theta) d\psi(\pi, \theta|s).$$

The definition of the Bayesian game and the solution concept remain unchanged. That is to say, we still focus on symmetric pure strategy profiles; and solve for the information designer's worst Bayesian Nash equilibrium to capture the idea of adversarial/robust information design.

B.2 Proof of Proposition 1

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

Definition 2. *For player 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 agent's strategy becomes more aggressive, an agent's best response is either unchanged or more aggressive.

Lemma 4. *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 i to attack given a'_{-i} and signal s, $a_j \geq a_j'$ for every $j \neq i$, and $a_j > a_j'$

for all j in a subset of $[0,1] \setminus \{i\}$ with positive measure. We must have

$$\begin{aligned} 0 &< \int_{\Delta(S)\times\Theta} u(1, \pi(\{s \in S | a'_j(s) = 1\}|\theta), \theta) d\psi(\pi, \theta|s) \\ &\leq \int_{\Delta(S)\times\Theta} u(1, \pi(\{s \in S | a_j(s) = 1\}|\theta), \theta) d\psi(\pi, \theta|s). \end{aligned}$$

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

Now we are ready to address the equilibrium existence.

Lemma 5. *For any (S, ψ), there exists an equilibrium.*

Proof. Fix (S, ψ), we construct an equilibrium through IESDS. We start by defining a series $\{S_k\}_{k=1}^{\infty}$ on the signal space.

Let $S_0 = \emptyset$ and $a^0(s) \equiv 1$ for every $s \notin S_0$. Given π, θ and an arbitrary $i \in [0,1]$, if every agent $j \neq i$ plays strategy $a^0(s)$, then the mass of agents attack (except agent i) is $\pi(\{s \in S | a^0(s) = 1\}|\theta)$. Define $S_1 \subseteq S$ as the set of signal s such that

$$\int_{\Delta(S)\times\Theta} u(1, \pi(\{s \in S | a^0(s) = 1\}|\theta), \theta) d\psi(\pi, \theta|s) \leq 0.$$

The left-hand side of the above inequality is agent i's expected net payoff from attacking when he receives signal s and given that all other agents play $a^0(s)$. Hence the condition means that if agent i receives signal $s \in S_1$, he weakly prefers not to attack even if every other agent attacks if and only if receiving a signal not in S_0 (i.e. always attacks).

Next, for $i \in [0,1]$, $k = 1, 2, 3, \cdots$, define

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

Given π, θ and an arbitrary $i \in [0,1]$, if every agent $j \neq i$ plays strategy $a^k(s)$, the mass of agents attack (except agent i) is $\pi(\{s \in S | a^k(s) = 1\}|\theta)$. Define $S_{k+1} \subseteq S$ as the set of signal s such that

$$\int_{\Delta(S)\times\Theta} u(1, \pi(\{s \in S | a^k(s) = 1\}|\theta), \theta) d\psi(\pi, \theta|s) \leq 0.$$

Note that $S_1 \supseteq S_0$ and the strategy profile where all agents play $a^0(s)$ is weakly more aggressive than the strategy profile where all agents play $a^1(s)$. By mathematical induction, if $S_k \supseteq S_{k-1}$, then $a^{k-1}(s)$ is weakly more aggressive than $a^k(s)$, then $\pi(\{s \in S | a^k(s) = 1\}|\theta) \leq \pi(\{s \in S | a^{k-1}(s) = 1\}|\theta)$, which implies $u(1, \pi(\{s \in S | a^k(s) = 1\}|\theta), \theta) \leq u(1, \pi(\{s \in S | a^{k-1}(s) = 1\}|\theta), \theta)$; thus, the above condition also holds for every $s \in S_k$. Therefore, $S \supseteq S_{k+1} \supseteq S_k$, and the strategy profile where all agents play $a^k(s)$ becomes weakly less aggressive as k increases. At the limit, as $k \to \infty$, the set $S^* = \lim_{k\to\infty} S_k$ exists, and $S^* \subseteq S$. Also, define

$$a^*(s) \equiv \begin{cases} 0 & \text{if } s \in S^* \\ 1 & \text{otherwise.} \end{cases} \tag{18}$$

Next we show that the strategy profile where all agents play $a^*(s)$ is indeed an equilibrium. First, by the construction of S^*, an agent prefers to attack when receiving any signal in $S \backslash S^*$ given other agents follow the strategy specified in (18). Second, we show that for any non-empty S^*, given that the other agents follow the strategy in (18), an individual agent i strictly prefers not to attack for every signal in S^*. The proof is straightforward. Pick any $s \in S^*$, and there exists a unique k such that $s \in S_k \backslash S_{k-1}$. By the definition of S_k, given that all other agents follow $a^{k-1}(s)$ and do not attack if and only if receiving a signal in S_{k-1}, an individual agent prefers not to attack when receiving a signal in $S_k \backslash S_{k-1}$. Then by Lemma 4, if all other agents follow a less aggressive strategy $a^*(s) < a^{k-1}(s)$ and do not attack if and only if receiving signals in $S^* \supseteq S_{k-1}$, an agent must prefer not to attack when receiving signals in $S_k \backslash S_{k-1}$. Thus, for every $s \in S^*$, it is optimal for agent i to choose $a_i(s) = a^*(s) = 0$. This completes the proof. □

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

Lemma 6. *For any (S, π), there is a unique equilibrium.*

Proof. For the sake of contradiction, suppose that for (S, ψ), there are two distinct symmetric equilibria a, a'. Let $\{s|a(s) = 0\}$ denote the set of no-attack signals (i.e. signals given which an agent does not attack in equilibrium) in equilibrium a and $\{s|a'(s) = 0\}$ denote the set of no-attack signals in equilibrium a'. As a and a' are distinct equilibria, $\{s|a(s) = 0\} \neq \{s|a'(s) = 0\}$; by the definition of adversarial equilibrium, a is not more aggressive than a', and vice versa. Consider the following

strategy a'' defined as follows:

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

which is strictly more aggressive than a and a'. By Lemma 4, an individual agent i receiving a signal in $S\backslash(\{s|a(s) = 0\} \cap \{s|a'(s) = 0\})$ prefers to attack if every other agent is adopting strategy a''. Note that by Lemma 5, an equilibrium always exists. According to the procedure of IESDS, 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 in a and in a', which is a contradiction. □

The combination of Lemmas 4-6 yields Proposition 1.

B.3 Optimality of π^* among Possibly Non-Deterministic Policies

Theorem 1′. *π^* remains optimal when the designer may commit to a non-deterministic but still anonymous information structure.*

Proof. Fix (S, ψ), we follow the definition of series $\{S_k\}_{k=0}^{\infty}$ and S^* in the proof of Lemma 5. For ease of exposition, we also define $S_{-1} = \emptyset$. Following the proof of Proposition 1, S^* characterizes the unique agent equilibrium.

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

$$\begin{cases} h_0(\theta) = f(\theta)\ \forall \theta \in [1, \bar{\theta}] \text{ and } = 0 \text{ elsewhere} \\ h_1(\theta) = f(\theta)\psi(\{\pi \in \Delta(S)|\pi(S\backslash S_1|\theta) \leq \theta\}|\theta) \\ h_2(\theta) = f(\theta)\psi(\{\pi \in \Delta(S)|\pi(S\backslash S_2|\theta) \leq \theta\}|\theta) \\ \cdots \end{cases}$$

Heuristically speaking, $h_k(\theta)$, $k \geq 0$, is the ex-ante probability density function that state θ realizes and survives after k rounds of IESDS. For every θ, the distance between $h_k(\theta)$ and $h_{k-1}(\theta)$ converges to 0 as $k \to +\infty$, and we denote by $h^*(\theta)$ the limit of $h_k(\theta)$ as $k \to +\infty$.

Consider round k of IESDS. For every $s \in S_k$, an individual agent i receiving s shall not attack even if every other agent j plays a^{k-1}, i.e. attacks if and only if j's signal is not in S_{k-1}. That is to say, if he attacks, the probability of regime change is smaller

than or equal to c. A necessary condition for agent i not to attack when receiving any signal in S_k (note that we have defined $S_{-1} = \emptyset$) is

$$
\begin{aligned}
& c \geq \frac{\int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_{k-1}|\theta) > \theta\}} f(\theta) \pi(S_k \backslash S_{k-1}|\theta) d\psi(\pi|\theta) dF(\theta)}{\int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_{k-1}|\theta) \leq \theta\}} f(\theta) \pi(S_k \backslash S_{k-1}|\theta) d\psi(\pi|\theta) dF(\theta) + \int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_{k-1}|\theta) > \theta\}} f(\theta) \pi(S_k \backslash S_{k-1}|\theta) d\psi(\pi|\theta) dF(\theta)} \\
\Rightarrow \quad & c \geq \frac{\int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_{k-1}|\theta) > \theta, \pi(S \backslash S_k|\theta) \leq \theta\}} f(\theta) \pi(S_k \backslash S_{k-1}|\theta) d\psi(\pi|\theta) dF(\theta)}{\int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_{k-1}|\theta) \leq \theta\}} f(\theta) \pi(S_k \backslash S_{k-1}|\theta) d\psi(\pi|\theta) dF(\theta) + \int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_{k-1}|\theta) > \theta, \pi(S \backslash S_k|\theta) \leq \theta\}} f(\theta) \pi(S_k \backslash S_{k-1}|\theta) d\psi(\pi|\theta) dF(\theta)} \\
\Leftrightarrow \quad & c \sum_{n=0}^{k-1} \int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_n|\theta) \leq \theta, \pi(S \backslash S_{n-1}|\theta) > \theta\}} f(\theta) \pi(S_k \backslash S_{k-1}|\theta) d\psi(\pi|\theta) dF(\theta) \\
\geq \quad & (1-c) \sum_{n=k}^{k} \int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_{n-1}|\theta) > \theta, \pi(S \backslash S_n|\theta) \leq \theta\}} f(\theta) \pi(S_k \backslash S_{k-1}|\theta) d\psi(\pi|\theta) dF(\theta).
\end{aligned}
$$

Similar to (9), the right-hand side of the first inequality measures the probability with which agent i believes the regime will fall, conditional on i receiving a signal in $S_k \backslash S_{k-1}$ and the other agents attacking if and only if their signal is not in S_{k-1}. The second inequality results from the fact that $\{\pi \in \Delta(S) | \pi(S \backslash S_{k-1}|\theta) > \theta\} \supseteq \{\pi \in \Delta(S) | \pi(S \backslash S_{k-1}|\theta) > \theta, \pi(S \backslash S_k|\theta) \leq \theta\}$. Finally, we obtain the third inequality since all the elements in $\{\{\pi \in \Delta(S) | \pi(S \backslash S_{k-1}|\theta) > \theta, \pi(S \backslash S_k|\theta) \leq \theta\}\}_{k=0}^{\infty}$ are mutually exclusive.

Then consider all the previous rounds of IESDS; each round corresponds to a necessary condition analogous to the above. For instance, consider round p of IESDS, $p < k$, and the corresponding condition is

$$
\begin{aligned}
& c \sum_{n=0}^{p-1} \int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_n|\theta) \leq \theta, \pi(S \backslash S_{n-1}|\theta) > \theta\}} f(\theta) \pi(S_p \backslash S_{p-1}|\theta) d\psi(\pi|\theta) dF(\theta) \\
\geq \quad & (1-c) \sum_{n=p}^{k} \int_\Theta \int_{\{\pi \in \Delta(S) | \pi(S \backslash S_{n-1}|\theta) > \theta, \pi(S \backslash S_n|\theta) \leq \theta\}} f(\theta) \pi(S_p \backslash S_{p-1}|\theta) d\psi(\pi|\theta) dF(\theta).
\end{aligned}
$$

Sum these conditions up for $p = 1, 2, \cdots, k$; then a necessary condition for the regime to persist with probability (function) $h_k(\theta)$ is

$$
\begin{aligned}
& c\sum_{p=1}^{k}\sum_{n=0}^{p-1}\int_{\Theta}\int_{\{\pi\in\Delta(S)|\pi(S\backslash S_n|\theta)\leq\theta,\pi(S\backslash S_{n-1}|\theta)>\theta\}} f(\theta)\pi(S_p\backslash S_{p-1}|\theta)d\psi(\pi|\theta)dF(\theta) \\
\geq\ & (1-c)\sum_{p=1}^{k}\sum_{n=p}^{k}\int_{\Theta}\int_{\{\pi\in\Delta(S)|\pi(S\backslash S_{n-1}|\theta)>\theta,\pi(S\backslash S_n|\theta)\leq\theta\}} f(\theta)\pi(S_p\backslash S_{p-1}|\theta)d\psi(\pi|\theta)dF(\theta) \\
\Leftrightarrow\quad & c\sum_{n=0}^{k-1}\int_{\Theta}\int_{\{\pi\in\Delta(S)|\pi(S\backslash S_n|\theta)\leq\theta,\pi(S\backslash S_{n-1}|\theta)>\theta\}} f(\theta)\pi(S_k\backslash S_n|\theta)d\psi(\pi|\theta)dF(\theta) \\
\geq\ & (1-c)\sum_{n=1}^{k}\int_{\Theta}\int_{\{\pi\in\Delta(S)|\pi(S\backslash S_{n-1}|\theta)>\theta,\pi(S\backslash S_n|\theta)\leq\theta\}} f(\theta)\pi(S_n|\theta)d\psi(\pi|\theta)dF(\theta). \qquad (19)
\end{aligned}
$$

Similar to (10), the second inequality is obtained by changing the order of summation. Note that except for the states in $[1,\bar{\theta}]$, if, under some π, the status quo persists after the nth round of IESDS, the designer must send signals in S_n to a population greater than or equal to $1-\theta$, and hence send signals in $S_k\backslash S_n$ with probability less than or equal to θ. Also note that $\int_{\Theta}(h_n(\theta)-h_{n-1}(\theta))d\theta$ is the probability measure of the states that persists after and only after the nth round of IESDS. Thus, for $n\geq 1$, we have

$$
\begin{aligned}
& \int_{\{\pi\in\Delta(S)|\pi(S\backslash S_{n-1}|\theta)>\theta,\pi(S\backslash S_n|\theta)\leq\theta\}} \pi(S_n|\theta)d\psi(\pi|\theta) \\
\geq\ & (1-\theta)\psi(\{\pi\in\Delta(S)|\pi(S\backslash S_n|\theta)\leq\theta,\pi(S\backslash S_{n-1}|\theta)>\theta\}|\theta) \\
\Leftrightarrow\quad & \int_{\{\pi\in\Delta(S)|\pi(S\backslash S_{n-1}|\theta)>\theta,\pi(S\backslash S_n|\theta)\leq\theta\}} f(\theta)\pi(S_n|\theta)d\psi(\pi|\theta) \\
\geq\ & (1-\theta)(h_n(\theta)-h_{n-1}(\theta)),
\end{aligned}
$$

and

$$
\begin{aligned}
& \int_{\{\pi\in\Delta(S)|\pi(S\backslash S_n|\theta)\leq\theta,\pi(S\backslash S_{n-1}|\theta)>\theta\}} \pi(S_k\backslash S_n|\theta)d\psi(\pi|\theta) \\
\leq\ & \theta\psi(\{\pi\in\Delta(S)|\pi(S\backslash S_n|\theta)\leq\theta,\pi(S\backslash S_{n-1}|\theta)>\theta\}|\theta) \\
\Leftrightarrow\quad & \int_{\{\pi\in\Delta(S)|\pi(S\backslash S_n|\theta)\leq\theta,\pi(S\backslash S_{n-1}|\theta)>\theta\}} f(\theta)\pi(S_k\backslash S_n|\theta)d\psi(\pi|\theta) \\
\leq\ & \theta(h_n(\theta)-h_{n-1}(\theta)).
\end{aligned}
$$

Apply the above two conditions, then (19) leads to the following necessary condition:

$$\begin{aligned} & c\left(\int_{\Theta} h_0(\theta)d\theta + \sum_{n=1}^{k-1}\int_{\Theta}(h_n(\theta)-h_{n-1}(\theta))\theta d\theta\right) \\ \geq\ & (1-c)\sum_{n=1}^{k}\int_{\Theta}(h_n(\theta)-h_{n-1}(\theta))(1-\theta)d\theta \\ \Leftrightarrow\quad & c\left(\int_{\Theta} h_0(\theta)d\theta + \int_{\Theta}(h_{k-1}(\theta)-h_0(\theta))\theta d\theta\right) \\ \geq\ & (1-c)\int_{\Theta}(h_k(\theta)-h_0(\theta))(1-\theta)d\theta. \end{aligned} \tag{20}$$

Letting $k = \infty$ in (20) produces an upper bound of the probability measure that the regime persists, which turns out to be identical to (11).

$$\begin{aligned} c\left(\int_{\Theta} h_0(\theta)d\theta + \int_{\Theta}(h^*(\theta)-h_0(\theta))\theta d\theta\right) &\geq (1-c)\int_{\Theta}(h^*(\theta)-h_0(\theta))(1-\theta)d\theta \\ \int_{\Theta}(1-\frac{1}{1-c}\theta)(h^*(\theta)-h_0(\theta))d\theta &\leq \frac{c}{1-c}\int_{\Theta} h_0(\theta)d\theta. \end{aligned} \tag{21}$$

Therefore, one way to identify a certain superset of the designer's optimum is to identify a function $h^*(\theta)$ to maximize $\int_{\Theta}(h^*(\theta)-h_0(\theta))d\theta$ under constraint (21). We will call this constrained maximization the "new relaxed problem." We assert that one desired $h^*(\theta)$ which solves the new relaxed problem takes the following form: $h^*(\theta) = 1$ for $\theta \in [\theta', \bar{\theta}]$ for some θ', and $h^*(\theta) = 0$ otherwise. That is, similar to the proof of Theorem 1, it is always optimal to "save" only the strongest states (with probability 1). To see this, following the same argument of Theorem 1, (1) it only weakly relaxes (21) to let $h^*(\theta) = 1$ for all states θ such that $\frac{1}{1-c}\theta \geq 1$; (2) For all θ such that $\frac{1}{1-c}\theta < 1$, to increase the maximand $\int_{\Theta}(h^*(\theta)-h_0(\theta))d\theta$ by an infinitesimal d by increasing $h^*(\theta)$, the left-hand side of (21) increases by $d(1-\frac{1}{1-c}\theta)$, which is strictly decreasing in θ. (1) and (2) imply that one solution to the new relaxed problem saves only the strongest states with probability 100%. Furthermore, every solution only differs from this particular $h^*(\theta)$ by a distance of zero. Following this specific form, (21) turns out to be identical to (13). Hereinafter we can adopt the previous proof of Theorem 1 to claim the optimality of π^*. □

C Appendix: Other Results

C.1 Omitted Results on Comparative Statics

Comparative statics for public disclosure. To ease the discussion of comparative statics, we rewrite equation (7) as

$$1 - c = \frac{1 - F(1)}{1 - F(\theta^\dagger)}. \tag{22}$$

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 4.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

$$1 - c = \frac{1 - F(1)}{1 - F(\theta^*)} + \frac{\int_{\theta^*}^1 \theta f(\theta) d\theta}{1 - F(\theta^*)}. \tag{23}$$

Compared to equation (22), equation (23) 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

$$c \geq F(1) - \int_0^1 \theta f(\theta) d\theta, \tag{24}$$

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 in 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 4.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

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

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. □