Incentive Compatibility and Its Betrayal: A Mathematical History of Political Campaigns in Communist China

Heng-Fu Zou

May 22, 2025

Abstract

This paper develops a unified theoretical and historical framework to explain the cyclical structure of political purging and institutional collapse in totalitarian regimes, with a focus on the Chinese Communist Party (CCP) from 1942 to the present. We introduce three formal models: a dynamic signaling game with time-inconsistent punishment, a modified epidemic model of political participation and removal, and a quota-driven reinterpretation game that captures the regime's shifting definition of guilt. These models reveal a structural logic in which early-stage participation is incentive-compatible but later becomes retroactively incriminating as political needs evolve. Simulations demonstrate how participation spreads rapidly, trust deteriorates, and long-run strategy space collapses under coercive reinterpretation. By mapping these models onto major Chinese political campaigns—from the Yan'an Rectification and the Anti-Rightist purge to the Cultural Revolution and Xi Jinping's anti-corruption drive—we show that totalitarian governance operates through sequential betrayal of compliance. The regime manufactures loyalty, rewards it in the short term, then redefines it as disloyalty to fulfill ideological or factional quotas. This dynamic creates a system where no strategy ensures safety and all signals become vulnerable to reinterpretation. We conclude that this self-consuming logic of incentive collapse is not an aberration but a mathematically demonstrable feature of totalitarianism itself.

Keywords: Chinese Communist Party; totalitarianism; political purges; game theory; incentive compatibility; signaling games; reinterpretation of guilt; quota systems; Cultural Revolution; authoritarian collapse

1 Introduction

Totalitarian regimes often possess a paradoxical vitality. At their inception, they mobilize mass enthusiasm, trigger ideological fervor, and rapidly transform society. Yet, over time, the very institutions they create become mechanisms of repression, mistrust, and systemic collapse. Nowhere has this pattern manifested more clearly—or more repeatedly—than in the campaigns of the Chinese

Communist Party (CCP). From the Yan'an Rectification Movement of the 1940s to the Anti-Rightist Campaign, the Cultural Revolution, and the ongoing anti-corruption purges under Xi Jinping, the CCP has employed a system of rule in which short-run incentives to participate are strong, but long-run outcomes are uniformly destructive to those who participate (MacFarquhar & Schoenhals, 2006; Dikötter, 2010; Lü, 2000).

This paper proposes a formal explanation for this dynamic using game-theoretic models that capture the CCP's core institutional logic: short-run incentive compatibility combined with long-run incentive incompatibility. In the early phases of political campaigns, individuals are rewarded for participation—through public praise, promotion, or redistribution of land and power. However, as the campaign escalates and definitions of loyalty are reinterpreted, the same individuals are punished for the very behaviors that were once rewarded. No strategy—cooperation, silence, or denunciation—ensures long-run safety. As Timur Kuran (1995) observes in the context of preference falsification under authoritarian rule, strategic behavior under coercion is ultimately self-defeating when institutional rules change arbitrarily.

Our central theoretical claim aligns with the historical thesis developed by Chenggang Xu in Institutional Genes: Origins of China's Institutions and Totalitarianism (2025). Xu argues that the CCP's ability to construct totalitarianism derives from its capacity to decompose fundamentally incompatible long-run goals into sequences of short-term, incentive-compatible revolutionary stages. These stages mobilize different social groups—peasants promised land, intellectuals promised democracy—against one another. Once their political utility is exhausted, each group becomes the target of the next campaign. The system is path-dependent, stepwise, and self-consuming. As Xu notes, this allows the CCP to violate the basic principle of incentive compatibility while still mobilizing mass participation and eliminating resistance. This pattern mirrors the logic of what Wintrobe (1998) calls "the dictator's dilemma": the more repression a ruler uses, the less reliable information and loyalty signals become, requiring further repression.

To explain these dynamics rigorously, we develop three formal models:

1. A Dynamic Signaling Game with Time-Inconsistent Punishment, in which agents who participate early to gain favor are later punished retroactively as definitions of guilt shift.

2. A Modified Epidemic Model, where political participation spreads contagiously (as in the Cultural Revolution), but results in widespread removal or purging once saturation is reached.

3. A Quota-Driven Reinterpretation Game, in which the regime continually expands the definition of guilt to meet internal political quotas, undermining the safety of all strategies and eroding trust across the system.

These models synthesize insights from political game theory (McCarty & Meirowitz, 2007), selectorate theory (Bueno de Mesquita et al., 2003), and the literature on authoritarian purges (Edmond, 2013; Bueno de Mesquita, 2005), while grounding them in detailed empirical histories of Chinese political campaigns (MacFarquhar, 1997; Dikötter, 2016; Pei, 2016). We show that the

collapse of strategic equilibrium under totalitarianism is not accidental—it is structurally embedded in a regime that cannot credibly commit to fixed rules of guilt, loyalty, or protection.

This dynamic is especially persistent in China due to its historical institutional inheritance. As Xu (2025), Ebrey (1996), and Hucker (1985) document, the Chinese imperial state—beginning with the Qin dynasty (221–206 BCE)—developed core mechanisms of totalitarianism long before the modern era: bureaucratic surveillance, quota-driven governance, collective punishment, and ideological reinterpretation of law. While Marxism-Leninism provided the ideological rationale, and Stalinism provided the administrative template (Lü, 2000; Pei, 2016), the CCP inherited a deeply rooted state tradition in which loyalty was always contingent, retroactive punishment routine, and institutional trust nonexistent.

The structure of this paper is as follows. Section 2 develops the dynamic signaling game and formalizes the logic of short-term cooperation followed by long-term punishment. Section 3 presents the epidemic model of participation and removal. Section 4 constructs the quota-driven reinterpretation game, focusing on retroactive guilt expansion and strategic collapse. Section 5 simulates these models, revealing dynamic feedback loops, the erosion of trust, and eventual paralysis. Section 6 maps our models onto key moments in Chinese political history. Section 7 synthesizes the findings in terms of equilibrium breakdown under totalitarian rule.

By combining formal modeling with deep historical structure, this paper seeks to explain why totalitarian regimes appear rational and effective in the short run, yet ultimately destroy the very basis for stable governance. In doing so, we provide both a general theory of political incentive collapse and a historically grounded explanation of one of the most enduring totalitarian systems in modern history.

2 The Dynamic Signaling Game with Time-Inconsistent Punishment

The dynamics of political participation under totalitarian regimes reveal a central paradox: individuals who initially comply with state incentives often find themselves punished for that very compliance in later stages of the regime's evolution. This paradox is especially acute in the Chinese Communist Party's political campaigns, where participation is heavily incentivized in the short run, but retroactively criminalized in the long run. To formalize this phenomenon, we construct a dynamic signaling model in which the regime's punishment strategy is time-inconsistent and strategically opaque. The purpose of this model is to capture both the rational decision-making of agents in the early stages of a political campaign and the endogenous collapse of incentive compatibility as the campaign intensifies and the regime reinterprets prior behavior.

Let time be discrete and indexed by $t = 0, 1, 2, \ldots$. The regime R presides

over a continuum of agents indexed by $i \in [0, 1]$. Each agent is characterized by a private type $\theta_i \in \{L, U\}$, where L denotes "loyal" (either ideologically committed or conformist), and U denotes "unfaithful" (dissident, independent-minded, or ideologically suspect). The proportion of loyal agents in the population is given by a prior $p \in (0, 1)$, and the regime cannot observe θ_i directly. Each period, the agent chooses an action $a_i(t) \in \{0, 1\}$, where $a_i(t) = 1$ denotes signaling loyalty (through confession, denunciation, or visible participation), and $a_i(t) = 0$ denotes remaining silent.

The agent receives a reward $r_t \geq 0$ from the regime for signaling loyalty at time t, and faces a punishment cost $\phi > 0$ if the regime later determines that the agent is politically guilty. The regime constructs a punishment rule $\pi_t : \mathcal{H}_i(t) \to \{0, 1\}$, mapping the agent's observed action history $\mathcal{H}i(t) = \{a_i(s)\}s \leq t$ into a binary punishment decision. The key feature of the model is that π_t is not stationary: it evolves over time as the regime's political objectives change. This models the Party's repeated practice of reinterpreting prior behavior to fulfill evolving ideological goals or purge quotas. The agent's objective is to maximize expected discounted utility:

$U_i = \sum_{t=0}^{\infty} \beta^t \left[r_t \cdot a_i(t) - \pi_t(\mathcal{H}_i(t)) \cdot \phi \right]$,

where $\beta \in (0, 1)$ is the agent's discount factor.

At the beginning of a campaign ($t = 0$), the regime commits to a low-punishment strategy—i.e., $\pi_0(\cdot) = 0$—to induce mass participation. This is consistent with Mao Zedong's behavior during the Hundred Flowers Movement (1956), when intellectuals were encouraged to speak freely and criticize the Party. The effective reward r_0 for voicing criticism was high, and the expected probability of punishment $\mathbb{E}[\pi_T \mid a_i(0) = 1]$ was perceived as negligible. Under these conditions, it is rational for most agents—both loyal and unfaithful types—to participate and signal loyalty, as the expected utility from participation outweighs that from silence.

However, at a later time $t = T$, the regime changes its punishment rule. The regime's internal logic now requires the identification and elimination of hidden enemies, and agents who participated earlier (i.e., $a_i(0) = 1$) are reevaluated. If an agent's earlier speech or denunciation is deemed ideologically impure or strategically deviant, the regime imposes punishment. Formally, the regime now applies a new punishment rule π_T such that:

$$\pi_T(a_i(0)) = \begin{cases} 1 & \text{if } a_i(0) = 1 \text{ and } \theta_i = U, \\ 0 & \text{otherwise.} \end{cases}$$

But since θ_i is unobservable, the regime uses $a_i(0)$ as a noisy proxy for θ_i, with the likelihood of punishment increasing as the regime's need for targets grows. This captures the logic of retrospective reinterpretation of guilt, evident during the Anti-Rightist Campaign (1957–1959), when individuals who spoke out under the Hundred Flowers directive were labeled "rightists" and subjected to imprisonment, forced labor, or execution (MacFarquhar, 1997; Dikötter, 2016). What had been incentive-compatible behavior in the short run was redefined as betrayal in the long run.

We can now state a fundamental incentive condition. At $t = 0$, an agent will

choose $a_i(0) = 1$ if the expected reward exceeds the discounted cost of future punishment:

$r_0 > \beta^T \cdot \phi \cdot \Pr[\pi_T(a_i(0) = 1)]$.

When this inequality holds, participation is rational in the short run. But as T increases or as $\Pr[\pi_T]$ rises (due to ideological shifts or quota increases), the inequality is reversed, and participation becomes ex post irrational. This dynamic captures the breakdown of trust in repeated authoritarian signaling games: initial incentive compatibility gives way to strategic regret and eventual strategic paralysis (Wintrobe, 1998; Kuran, 1995).

This framework can be used to explain other historical episodes. During the Cultural Revolution (1966–1976), early Red Guard participants enthusiastically attacked teachers, officials, and family members in response to Mao's calls for class struggle. Their participation was rewarded with symbolic prestige and political capital. However, by the early 1970s, many of these same Red Guards were purged as "ultra-leftist" deviants or factional conspirators (MacFarquhar & Schoenhals, 2006). Participation, once necessary for survival, became fatal as the regime reinterpreted its priorities.

The same logic reappears in the Xi Jinping era's anti-corruption campaign. Early enforcers such as Sun Zhengcai and Fu Zhenghua, who zealously executed Xi's orders to root out corruption, were themselves arrested and imprisoned once they became politically expendable (Pei, 2016). This confirms the structural feature of totalitarian logic emphasized by Xu (2025): the regime decomposes long-term, incentive-incompatible objectives (such as total ideological control) into a sequence of short-run, incentive-compatible acts. But once those acts serve their purpose, they are reinterpreted as signs of deviation. The system incentivizes loyalty, then punishes it.

This model generates several broader implications. First, in such a regime, there is no stable equilibrium strategy: both silence and participation may become punishable, depending on future reinterpretations. Second, mass participation is self-liquidating. The more individuals signal loyalty, the greater the pool of targets for retroactive punishment becomes. Third, agents eventually learn that all strategies are dangerous, resulting in strategic paralysis and the erosion of initiative. These implications generalize well beyond the Chinese context and may apply to other ideological totalitarian systems that rely on mass campaigns, such as Stalinist Russia, North Korea, and Pol Pot's Cambodia.

In sum, the dynamic signaling game with time-inconsistent punishment formalizes the logic by which authoritarian regimes use initial incentives to build participation and then destroy that participation to reinforce control. It explains how totalitarian institutions can temporarily align incentives to generate visible loyalty, only to revert those incentives to impose fear, silence, and obedience. It also explains why trust never accumulates in such systems—and why history is littered with the bodies of former loyalists who misread the signals of the regime they served.

3 The Epidemic Model of Political Participation and Purging

One of the most striking features of totalitarian political campaigns is their viral structure: participation spreads rapidly from individual to individual, often driven by peer pressure, local political incentives, or fear of being the last to act. But just as an epidemic eventually burns out by exhausting its pool of susceptible hosts, so too do political campaigns reach a saturation point—at which time the system turns inward, reclassifying former participants as deviant and purging them to reassert central control. This section formalizes that process through a modified SIR (Susceptible-Infected-Removed) model, commonly used in epidemiology but here repurposed to describe the diffusion and collapse of political participation under totalitarianism.

We define three population categories at time t: $S(t)$, the share of individuals who are susceptible to participation (i.e., have not yet joined the campaign); $I(t)$, the share actively participating (e.g., denouncing others, writing self-criticisms, attending rallies); and $R(t)$, the share who have been purged, removed, or otherwise neutralized by the regime. These proportions sum to unity: $S(t) + I(t) + R(t) = 1$. The dynamics are governed by a system of nonlinear differential equations:

$$\frac{dS}{dt} = -\beta S(t) I(t),$$

,

$$\frac{dI}{dt} = \beta S(t) I(t) - \gamma I(t),$$

$$\frac{dR}{dt} = \gamma I(t),$$

where $\beta > 0$ is the transmission rate of political participation—how quickly an individual's decision to participate influences others in their social or organizational network—and $\gamma > 0$ is the rate at which active participants are purged or removed from political life. These equations mirror the canonical SIR model used in modeling contagious diseases (Kermack & McKendrick, 1927), but here the "infection" is not biological but ideological, and the "removal" is political punishment or liquidation.

Initially, when $S(0)$ is close to 1 and $I(0)$ is small, the model predicts a rapid rise in $I(t)$. That is, as the regime launches a campaign and a few participants begin to act, their behavior spreads through peer pressure, fear of being labeled passive, or hope for material gain. This is consistent with historical patterns in China, such as the Cultural Revolution, where Red Guards rapidly multiplied through high school and university networks, or the Anti-Rightist Campaign, where intellectuals rushed to participate in "criticism sessions" to avoid suspicion.

However, as $S(t)$ declines and $I(t)$ rises, the model reaches a peak infection level, after which $\frac{dI}{dt} < 0$. Participation starts to decline—not because agents cease to participate voluntarily, but because participants begin to be purged. The mechanism of purging, modeled by the term $\gamma I(t)$, reflects the regime's tendency to first encourage activity and then punish that very activity when it has served its political purpose or when the regime requires scapegoats. This is precisely what happened in the Red Guard phase of the Cultural Revolution: after encouraging students to destroy the "Four Olds" and attack officials, Mao Zedong and the Party later turned on the Red Guards themselves, disbanding factions, sending youth to the countryside, and imprisoning key leaders (MacFarquhar & Schoenhals, 2006).

The model implies a grim symmetry: the same process that builds participation ensures its destruction. Once participation becomes widespread, it ceases to be a differentiator of loyalty. The regime, needing new enemies and unable to distinguish between true and false loyalty (Wintrobe, 1998), turns to previously active participants as new targets. Thus, the set $I(t)$ is first incentivized and later decimated. In the long-run limit as $t \to \infty$, we observe $I(t) \to 0$, $R(t) \to R^*$, and $S(t) \to S^*$, where S^* may remain nonzero if some individuals never participated and thus avoided both reward and risk.

This formalism helps explain how participation in totalitarian campaigns becomes a trap: individuals act to avoid being singled out or to curry favor, only to discover that the saturation of participation makes them indistinguishable from their peers—and thus equally purgeable. As Xu (2025) argues, the CCP perfected the strategy of first fragmenting society through staged mobilizations and then purging each group once their mobilization no longer served the regime's evolving goals. This strategy was deeply effective precisely because it decomposed a fundamentally incentive-incompatible long-run goal—absolute ideological conformity—into a sequence of local mobilizations that appeared short-run compatible for specific groups (e.g., peasants during land reform, students during Cultural Revolution, officials during anti-corruption campaigns).

The epidemic model also reveals that participation dynamics are path-dependent: the earlier one participates, the more influence they exert on others (through \beta), but also the more exposed they become to future reinterpretation and purging. This aligns with empirical observations that early participants in CCP campaigns—be they Red Guards, whistleblowers, or even campaign enforcers—often became some of the first victims of reversal or punishment. The model predicts that early action maximizes both short-run gain and long-run exposure, a fundamental tension at the heart of totalitarian politics.

Finally, this model generalizes well beyond the Chinese context. Similar contagion-purge dynamics occurred under Stalin's Great Terror, where initial denunciations were rewarded and later reversed, and under the Khmer Rouge, where mass mobilization into revolutionary units eventually led to mass internal liquidation (Short, 2004). What this model contributes is a precise temporal logic: a regime that encourages widespread action must, at saturation, convert participation into guilt, especially when ideological purity becomes indistinguishable from mass conformity.

Hence, the epidemic model provides a mathematically grounded explanation for the trajectory of totalitarian mass campaigns. It captures how campaigns grow explosively through social transmission, reach a peak of visibility and saturation, and then collapse inward through selective reinterpretation and political purging. When paired with the time-inconsistent signaling logic of the previous section, it reveals a powerful dynamic of totalitarian regimes: they require mass participation to destroy their enemies, but they require mass purging to preserve control. Participation is not safety—it is merely the first phase of danger.

4 The Quota-Driven Reinterpretation Game

The logic of totalitarian purging often rests not on concrete acts of opposition but on abstract criteria—shifting, retrospective, and deliberately vague. Participation in a campaign, prior loyalty, even silence, can be recategorized as guilt when the regime's political calculus demands it. This section models the phenomenon of quota-driven reinterpretation, where the state imposes endogenous targets for identifying political enemies and adjusts the criteria of guilt over time to meet these targets. Historically, this logic undergirded numerous CCP campaigns, including the Anti-Rightist Movement (1957), where a preassigned number of "rightists" had to be found regardless of their actual ideological deviance, and the Cultural Revolution, where nearly every work unit was required to uncover and punish a class enemy. This quota logic, institutionalized across Chinese history since the Qin dynasty and rationalized under Maoist governance, transforms political classification into a fluid mechanism of discipline, not a reflection of past behavior.

Let us formalize the state's problem. Let there be a population of N agents indexed by $i \in \{1, \ldots, N\}$. Each agent possesses a private type $\theta_i \in \{L, U\}$, as in previous sections, denoting loyal or unfaithful disposition. The regime cannot observe θ_i, but can observe a public action history $\mathcal{A}_i(t)$ up to time t. Let this history include political behaviors such as campaign participation, publications, speeches, voting records, and denunciations.

The regime assigns to each agent a guilt score $G_i(t) \in \mathbb{R}$, defined by a time-dependent mapping $G_i(t) = f_t(\mathcal{A}_i(t))$. Crucially, the function f_t evolves over time. This captures the phenomenon of retrospective redefinition: past actions that were politically acceptable—or even rewarded—at time t_0 can be reinterpreted as signs of deviation at time $t > t_0$. Mathematically, this means that the same action history $\mathcal{A}_i$ can yield different scores $G_i(t) \neq G_i(t')$ due to shifts in f_t. For example, expressing mild reformist opinions in 1956 may be assigned low G_i (1956) but reclassified as high-risk in 1957 under Mao's purge logic.

At each period t, the regime sets a purge quota $Q(t) \in \mathbb{N}$, specifying the number of individuals who must be labeled guilty and removed from positions of influence or subjected to punishment. This quota may be imposed from above (e.g., from the Central Committee or Politburo) or arise endogenously from the regime's internal security apparatus seeking to demonstrate vigilance.

The regime then selects a set $M(t) \subseteq \{1, \ldots, N\}$ such that $|M(t)| = Q(t)$, with members chosen as the $Q(t)$ agents with the highest $G_i(t)$ scores.

To operationalize this mathematically, we assume:

$M(t) = \text{Top-}Q(t)\{\ G_i(t) \mid i = 1, \ldots, N\}$.

Punishment is then applied to each $i \in M(t)$ with cost $\phi > 0$, as in previous models. The critical feature is that as $Q(t)$ increases over time, the regime must expand the pool of suspicious behaviors to maintain purge momentum. This forces the guilt scoring function f_t to become more inclusive, reclassifying previously benign or loyal actions as suspect. That is, for $t_2 > t_1$, we have:

$f_{t_2}(\mathcal{A}) \geq f_{t_1}(\mathcal{A})$ for a growing domain of $\mathcal{A}$.

This mechanism generates endogenous insecurity: even agents whose actions were once fully compliant now face increasing probability of punishment over time.

Historically, this model is directly supported by empirical evidence from the Anti-Rightist Campaign. Initially, only outspoken liberal critics were targeted. But by mid-1957, provincial authorities were ordered to meet explicit rightist quotas, and when obvious targets were exhausted, local cadres began to label mild reformists, previously loyal writers, and even earlier critics of rightists as "rightists" themselves (MacFarquhar, 1997). The result was an exponential growth in punishments—from fewer than 10,000 initially to over 550,000 within months. The guilt scoring function had to evolve rapidly to meet the centrally assigned quotas, resulting in persecution by reinterpretation.

A similar process occurred during the Cultural Revolution, when local work units were told that they had to find "one percent of class enemies" among their members. When obvious suspects were absent, the function f_t redefined guilt to include bad family background, insufficient enthusiasm in rallies, or past association with purged leaders. Thus, even loyalists could be targeted, and the only way to maintain safety was to constantly shift one's performance in line with an unpredictable and evolving ideological metric.

The strategic implication of the model is stark. In the presence of a time-evolving f_t and rising $Q(t)$, the agent's utility from any fixed strategy $a_i(t) \in \{0, 1\}$ becomes negative in expectation over a long enough time horizon. The expected probability of being included in $M(t)$ converges to a positive constant for every agent, due to the expanding definition of guilt. That is,

$\lim_{t\to\infty} \Pr(i \in M(t) \mid \mathcal{A}_i(t)) > 0 \quad \forall i.$

This means that there is no safe long-run strategy—not even total silence or perfect conformity. Every action can, at some point, be reinterpreted under a new scoring rule and used as grounds for punishment. In practical terms, this explains the pervasive atmosphere of fear and strategic paralysis that emerges in late-stage campaigns.

Xu (2025) identifies this process as a core innovation of the CCP's totalitarian technology: to disaggregate its incentive-incompatible end goal—absolute control over belief and behavior—into successive campaigns that appear compatible with the short-term interests of distinct groups. But because the regime always needs new enemies to target, it cannot fix the definition of loyalty. This is why the CCP repeatedly violated its own promises (e.g., land rights to peas-

ants, freedom to intellectuals) after achieving campaign objectives. Quotas, and their enforcement through reinterpretation, were necessary to mobilize society and suppress resistance in successive waves.

Moreover, the historical function of quotas in Chinese history long predates the CCP. As scholars of imperial governance have shown (Hucker, 1985; Ebrey, 1996), Qin and Han bureaucrats were routinely assessed based on how many criminals, tax delinquents, or heretics they uncovered. Failure to meet such numbers was interpreted as incompetence or disloyalty. The totalitarian modern state inherits and magnifies this quota logic, combining it with modern surveillance, propaganda, and ideological production. In effect, the CCP has transformed a dynastic disciplinary tool into a self-propelling, structurally paranoid political algorithm.

Therefore, the quota-driven reinterpretation game captures a core institutional logic of totalitarian governance. It formalizes the strategic use of variable punishment definitions and numerical enforcement targets to maintain a system of permanent instability, fear, and vulnerability. Agents cannot predict how their past will be read tomorrow, and therefore cannot construct stable expectations or trust in the system. In such regimes, loyalty is not a strategy—it is a liability. The more one conforms, the more material the regime accumulates to reinterpret when the purge quota demands it.

5 Simulation of All Three Models

To concretely illustrate the theoretical mechanisms developed in Sections 2 through 4, we simulate the dynamics of all three models using realistic parameter values chosen to reflect historically plausible political conditions. These simulations allow us to visualize how participation rises, trust collapses, and totalitarian control expands—despite being fundamentally built upon time-inconsistent incentives and strategic insecurity. Each simulation traces the trajectory of a typical political campaign under CCP-style totalitarianism, highlighting how apparent order evolves into internal disintegration and systemic repression.

We begin with the dynamic signaling game. Assume a population of $N = 1000$ agents, each with discount factor $\beta = 0.95$, and let the initial reward for participating in a campaign (e.g., speaking during Hundred Flowers) be $r_0 = 10$. The punishment cost is set to $\phi = 100$, a reflection of the severe consequences of political condemnation such as imprisonment, forced labor, or execution. The probability of retroactive punishment is modeled as an increasing function of time:

$\Pr[\text{punishment at } t] = 1 - e^{-\alpha t}$,

with $\alpha = 0.2$. This captures how, over time, the regime's need for scapegoats or ideological reclassification grows.

The simulation shows that at $t = 0$, participation is nearly universal. Over 90 of agents choose to signal loyalty due to the immediate reward and the low perceived risk of punishment. However, as t approaches 5, the expected penalty for prior participation surpasses the initial reward:

$\mathbb{E}[\text{future penalty}] = \beta^5 \cdot 100 \cdot (1 - e^{-1}) \approx 61.3 > r_0$.

Consequently, by $t = 5$, rational agents begin to withdraw from participation altogether. The campaign's strategic environment becomes toxic: past cooperation now signals risk, while silence remains suspicious. This reproduces the real-time collapse of trust and cooperation seen in 1957–1958, when many who had participated in the Hundred Flowers movement began to fear that every word they had spoken might be used against them.

Next, we simulate the epidemic model of campaign contagion and purging. We normalize the population size and set initial conditions $S(0) = 0.99$, $I(0) = 0.01$, $R(0) = 0$. The participation transmission rate is set to $\beta = 0.4$, reflecting how social and political contagion spreads rapidly through work units, universities, and family networks. The purge rate is set at $\gamma = 0.3$, indicating that once participation spreads, the regime actively begins to discipline and remove actors from political life.

The model reveals an early exponential rise in participation from $t = 0$ to $t = 5$, with $I(t)$ peaking at around 0.35. This reflects the early mass fervor of the Cultural Revolution when youth brigades and Red Guard factions quickly emerged. However, as $S(t) \to 0.6$, and the saturation point is reached, γ begins to dominate: $dI/dt < 0$, and removal accelerates. By $t = 10$, over 50 of the population has been purged or removed from political roles. At this point, active participation collapses, not because agents cease to comply, but because the regime uses them up. The historical analogy is clear: by the early 1970s, Mao's regime began dismantling the very Red Guards it had created, sending them to the countryside and labeling many as ultra-leftist deviants.

Finally, we simulate the quota-driven reinterpretation model. We assume the regime requires $Q(t) = 100 + 30t$ purges per time period to meet internal discipline or political demonstration goals. Each agent begins with a base guilt score $G_i(0) \sim \mathcal{N}(0.5, 0.1)$. The guilt score evolves according to:

$G_i(t+1) = G_i(t) + \varepsilon_i(t), \quad \varepsilon_i(t) \sim \mathcal{N}(0.05, 0.02)$,

representing reinterpretation drift, where past behavior is re-evaluated under shifting ideological criteria. The regime selects the top $Q(t)$ agents by $G_i(t)$ for punishment.

The simulation shows that by $t = 3$, over 50 of agents have been labeled guilty, including many with histories of participation. Because $Q(t)$ grows linearly and $G_i(t)$ evolves stochastically but upwardly biased, no agent can remain safe indefinitely. By $t = 10$, over 90 of the population has experienced punishment. This matches the trajectory of the Anti-Rightist Campaign, where early critics were followed by moderates, and eventually by even the enforcers of the campaign themselves. Similarly, under Xi Jinping's anti-corruption drive, tens of thousands of CCP officials—including high-ranking campaign leaders—were detained as the scope of guilt expanded.

Together, these simulations confirm a profound political logic. Short-run incentive compatibility—visible in early participation, mass signaling, and rapid campaign growth—is systematically undermined by the regime's evolving needs. As the regime redefines guilt, escalates quotas, and converts past loyalty into liability, all strategic behavior collapses. Agents can neither trust past rewards

nor project future safety. What appears as irrational betrayal by the regime is, in fact, structurally embedded in its design: the regime must first mobilize, then purge, and then reinterpret in order to sustain itself.

This simulation framework offers more than stylized dynamics; it reconstructs the internal rationality of CCP purging cycles. In doing so, it validates Chenggang Xu's (2025) argument that totalitarian institutions endure by sequentially transforming short-term incentives into long-term traps. Moreover, these results help explain why—despite repeated purges and mass trauma—such regimes can still mobilize participation: not because agents trust the system, but because in the early stages, the incentives are real. Only later do they discover that compliance has no memory.

6 Mapping the Models to Historical Chinese Campaigns

The formal models developed in this paper—dynamic signaling under time-inconsistent punishment, epidemic diffusion of participation, and quota-driven reinterpretation—are not mere theoretical constructions. They mirror with remarkable precision the actual structure and progression of political campaigns in the history of the Chinese Communist Party (CCP). From the early revolutionary years in Yan'an to the most recent anti-corruption campaigns under Xi Jinping, the same cyclical pattern has emerged: initial encouragement of participation, rapid spread of conformity and denunciation, reclassification of guilt, and mass purging of participants themselves. This section links each model to a sequence of concrete historical episodes, demonstrating that what might appear to be irrational or inconsistent behavior by the CCP is, in fact, a systematic and deeply rational institutional logic—one rooted in the structural dynamics of totalitarian survival.

The first clear instance of the dynamic signaling game emerged during the Yan'an Rectification Movement (1942–1945). CCP cadres and intellectuals were encouraged to engage in self-criticism, confess ideological confusion, and denounce "dogmatists." Initially, participation was rewarded: those who confessed or named others were promoted, rehabilitated, or protected. However, as the movement matured, those same confessions were reinterpreted as evidence of subversion or hidden disloyalty. Many confessors were executed or demoted, while those who had remained silent were now viewed as sly or evasive. The payoff structure changed dramatically over time: what had been a safe signal of loyalty was later treated as incriminating. This exactly matches the predictions of our dynamic signaling model, where early participation yields rewards, but time-inconsistent punishment retroactively criminalizes past cooperation.

The same model structure played out even more visibly during the Hundred Flowers Movement (1956–1957) and the Anti-Rightist Campaign (1957–1959). In 1956, Mao publicly encouraged intellectuals and party members to "let a hundred flowers bloom" and "let a hundred schools of thought contend." Hun-

dreds of thousands of Chinese citizens—including journalists, professors, and CCP members—responded, voicing criticism of bureaucracy, corruption, and ideological rigidity. But by the summer of 1957, the regime reversed its stance. Mao declared that these criticisms were "poisonous weeds," and launched the Anti-Rightist Campaign, retroactively using speeches and articles from the prior year as evidence of counterrevolutionary thinking. Over 550,000 people were labeled as "rightists," subjected to forced labor, loss of employment, imprisonment, or death (MacFarquhar, 1997; Dikötter, 2016). The dynamic signaling model captures this transformation: early rewards collapse into long-term punishment due to a political shift in the regime's interpretation function π_t. The result was the destruction of trust in political signaling for decades to come.

The epidemic model is vividly illustrated by the trajectory of the Cultural Revolution (1966–1976). When Mao launched the campaign by calling on youth to "bombard the headquarters," participation spread like wildfire. The Red Guards, composed of high school and university students, attacked teachers, officials, and even family members. Political loyalty became performative and contagious. Participation spread not merely from ideological conviction, but from fear of appearing passive or counterrevolutionary. This is precisely the structure modeled in our epidemic equations: as one agent participates, it increases the probability that their peers will follow, generating a self-reinforcing loop of visible conformity. Between 1966 and 1969, millions joined mass rallies, wrote confessions, destroyed temples and books, and denounced colleagues.

However, as our model predicts, the campaign soon reached saturation. By the early 1970s, the regime could no longer distinguish genuine loyalty from performative participation. Mao and the Party began to suppress the Red Guards themselves. Many were sent to the countryside in the "Up to the Mountains, Down to the Countryside" movement; others were imprisoned as ultra-leftist radicals. The number of those "removed" (in the model, R(t)) rose dramatically. By 1976, a vast proportion of early participants had been purged. The epidemic model's structure—rapid initial participation followed by long-term attrition through state purging—matches this trajectory with chilling accuracy.

The logic of quota-driven reinterpretation appears most clearly in the Anti-Rightist Campaign, the Cultural Revolution, and the ongoing anti-corruption campaign under Xi Jinping. In 1957, after the initial wave of intellectuals had been punished, provincial CCP offices were given explicit quotas for the number of "rightists" to be found in each institution, regardless of whether any existed. Local cadres were pressured to produce numerical results. The guilt function G_i(t) in our model was expanded: even praising the Party insufficiently, failing to criticize others, or previously denouncing rightists too zealously became reasons for suspicion. The function $f_t(\mathcal{A}_i(t))$ was no longer based on absolute behavior but evolved to ensure the quota $Q(t)$ could be filled. This led to a second wave of punishments that swept in moderate reformers, conformists, and previous enforcers—just as our model predicts when $Q(t)$ increases and f_t expands.

A similar pattern unfolded during Xi Jinping's anti-corruption drive, beginning in 2013. Originally targeting obvious abusers of public funds, the campaign

evolved into an instrument of political centralization. Over 4.7 million officials were investigated or punished by 2023. Many high-ranking Party elites—such as Sun Zhengcai, Zhou Yongkang, and even former anti-corruption czars—were arrested. As in our quota model, guilt scoring functions drifted: ties to rival factions, prior silence, or even old displays of loyalty became evidence of deviation. The ongoing need to demonstrate ideological control meant the campaign could not end with a fixed number of guilty officials; instead, it demanded continuous reinterpretation to justify further purges. Our simulation in Section 5, which showed that over 90% of agents eventually face punishment as $Q(t) \to N$, is mirrored in the real-world logic of factional elimination and total Party discipline under Xi.

Taken together, these historical cases confirm that the mathematical dynamics we modeled are not incidental—they are structural. Each CCP campaign follows a similar arc: short-run encouragement of participation, mass spread of conformity and denunciation, growing political saturation, shifting definitions of guilt, rising purge quotas, and eventual betrayal of the very groups that made the campaign possible. Participation is rational only in the early phase, when the reward is salient and the risk appears negligible. But once the regime's repressive apparatus reactivates, every prior action becomes a liability. As Xu (2025) argues, this is not merely a pattern—it is a strategy. The CCP decomposes a long-term, fundamentally incentive-incompatible goal (total control of thought and behavior) into sequential political movements, each of which appears temporarily rational to the group it mobilizes, but culminates in that group's elimination. Loyalty is manufactured, weaponized, and then destroyed.

Thus, the Chinese case provides perhaps the most complete empirical realization of our models. It is a regime that has survived not despite undermining trust, but by institutionalizing its collapse. The logic of strategic reinterpretation, viral participation, and quota-driven purging is neither accidental nor avoidable—it is a defining feature of totalitarian governance.

7 Conclusion – The Mathematical Logic of Political Collapse

The models developed in this paper reveal a foundational instability embedded in the architecture of totalitarian rule. By analyzing three interlinked mechanisms—time-inconsistent signaling, epidemic participation with mass purging, and quota-driven reinterpretation of guilt—we have shown that totalitarian political campaigns are structurally unable to sustain long-term incentive compatibility. Instead, they operate through a cyclical logic of short-term mobilization followed by strategic betrayal. This logic explains both the internal coherence of such regimes during the early stages of a campaign and their eventual tendency to consume even their most loyal adherents. Our simulations, grounded in realistic political parameters, reinforce the stark finding that no long-run safe strategy exists in a totalitarian system governed by evolving ide-

ological criteria and enforced purging quotas.

This conclusion is not merely theoretical. As we have demonstrated through extensive historical mapping, every major Chinese Communist Party campaign since 1942 has followed this structural logic. The Yan'an Rectification, Hundred Flowers and Anti-Rightist Campaigns, Cultural Revolution, and Xi Jinping's anti-corruption purges all began with inducements to participate—ranging from promises of reform and liberation to protection or promotion. But each campaign eventually transformed participation into evidence of guilt, leveraging reinterpreted histories, shifting ideological benchmarks, and escalating purge quotas to destroy individuals once deemed loyal. These campaigns exhibit the same mathematical form: initially incentive-compatible behavior becomes retroactively criminalized as the regime re-optimizes its punishment function over time.

This behavior is not irrational. It is the result of a coherent institutional strategy first observed in early dynastic China and perfected in modern totalitarian regimes. As Chenggang Xu (2025) has argued, the CCP operationalizes a system in which fundamentally incompatible long-run goals—total ideological conformity, perpetual regime security, absolute control over information—are broken into stages, each offering short-term incentive compatibility to selected social groups. The Party then mobilizes those groups to eliminate rivals, only to purge them once their role is complete. Our models show that such strategies are sustainable not despite their internal contradictions but because they are sequenced through deception, quotas, and redefinition.

Importantly, this paper offers a generalizable framework for understanding authoritarian decay. Although our focus has been China, the models are not confined to this case. Similar patterns of participation followed by reinterpretation and purging have emerged in Stalinist Russia, Cambodia under the Khmer Rouge, and North Korea under successive Kims. What all of these regimes share is a core structure: they cannot credibly commit to fixed rules of loyalty and protection because their survival depends on internal repression, not stable cooperation. They destroy trust not as a policy failure, but as a condition of rule.

The implication is profound. In such systems, political equilibrium cannot stabilize. Agents rationally participate early, only to be destroyed later. Strategic silence becomes suspicious. Excessive loyalty becomes incriminating. Eventually, the regime faces a population that no longer believes in signaling, no longer acts spontaneously, and no longer trusts even its own history. What remains is a brittle structure maintained through coercion, fear, and recursive reinterpretation. When purges can no longer be sustained, and quotas can no longer be met, the regime faces either paralysis or implosion.

This is the mathematical logic of political collapse under totalitarianism. It is not merely that the system becomes unjust. It is that the system becomes strategically unsolvable. Our models show that every signal deteriorates into noise, every action becomes risky, and every prior loyalty is weaponized. In the long run, totalitarianism eats its own equilibrium.

References

• Bueno de Mesquita, B., Smith, A., Siverson, R. M., & Morrow, J. D. (2003). The Logic of Political Survival. Cambridge, MA: MIT Press.

• Bueno de Mesquita, E. (2005). The quality of information and authoritarian politics. American Journal of Political Science, 49(3), 530–543.

• Dikötter, F. (2010). Mao's Great Famine: The History of China's Most Devastating Catastrophe, 1958–1962. New York: Walker & Company.

• Dikötter, F. (2016). The Cultural Revolution: A People's History, 1962–1976. London: Bloomsbury Publishing.

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

• Ebrey, P. B. (1996). The Cambridge Illustrated History of China. Cambridge University Press.

• Fudenberg, D., & Tirole, J. (1991). Game Theory. Cambridge, MA: MIT Press.

• Hucker, C. O. (1985). A Dictionary of Official Titles in Imperial China. Stanford University Press.

• Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society A, 115(772), 700–721.

• Kuran, T. (1995). Private Truths, Public Lies: The Social Consequences of Preference Falsification. Cambridge, MA: Harvard University Press.

• Lü, X. (2000). Cadres and Corruption: The Organizational Involution of the Chinese Communist Party. Stanford University Press.

• MacFarquhar, R. (1997). The Origins of the Cultural Revolution: Volume 3, The Coming of the Cataclysm 1961–1966. Oxford University Press.

• MacFarquhar, R., & Schoenhals, M. (2006). Mao's Last Revolution. Cambridge, MA: Harvard University Press.

• Maskin, E., & Tirole, J. (2001). Markov perfect equilibrium: I. Observable actions. Journal of Economic Theory, 100(2), 191–219.

• McCarty, N., & Meirowitz, A. (2007). Political Game Theory: An Introduction. Cambridge University Press.

• Pei, M. (2016). China's Crony Capitalism: The Dynamics of Regime Decay. Cambridge, MA: Harvard University Press.

• Schoenhals, M. (1996). China's Cultural Revolution, 1966–1969: Not a Dinner Party. Armonk, NY: M. E. Sharpe.

• Short, P. (2004). Pol Pot: Anatomy of a Nightmare. New York: Henry Holt and Co.

• Wintrobe, R. (1998). The Political Economy of Dictatorship. Cambridge University Press.

• Xu, C. (2025). Institutional Genes: Origins of China's Institutions and Totalitarianism. Cambridge University Press.