Conflict of Interest Declaration for “Purges and Predictability”

Matteo Camboni and Samuel Kapon have no conflicts of interest to disclose.

Purges and Predictability

Sam Kapon
UC Berkeley, Haas BPP

Matteo Camboni
UWisconsin–Madison*

August 15, 2026

Abstract

We study a novel dynamic game in which a principal can purge a group of agents, sequentially targeting its members for elimination. Agents can organize a protest that, if sufficiently large, may end the purge permanently. When targeting is predictable, next-period targets coordinate with current targets, limiting the purge to a few agents and at most one period. When targeting is unpredictable, currently untargeted agents face diffuse risk of near-term elimination, weakening their incentive to protest. Unpredictability typically benefits the principal by allowing multi-period purges that eliminate a larger share of agents. We illustrate these results with historical elite purges.

1 Introduction

Authoritarian leaders purge their enemies, their friends, and the masses. These purges often unfold gradually, with unpredictable criteria for selecting targets (Gregory, 2009). Participation in repression and past loyalty offer no guarantee of safety: authoritarian leaders routinely turn against old allies and powerful insiders in their quest to centralize power. Potential victims understand that they too may be targeted and, especially early in the process, often collectively possess the capacity to resist. Yet in many historical purges, such as those perpetrated by Stalin or Saddam Hussein, elites remained largely compliant.

*Kapon: skapon@berkeley.edu; Camboni: camboni@wisc.edu.

We provide a theory of how the predictability of future targeting shapes this compliance and, more broadly, the dynamics of sequential purges. When future targeting is predictable, for example because the leader's priorities can be inferred from observable characteristics (faction, rank, institutional affiliation, or documented past conduct), elites who expect to be targeted soon have an incentive to join current targets in resistance against the leader. This mechanism generates blocking coalitions that limit purges to a few agents and at most one period. By contrast, when future targeting is unpredictable, the risk of near-term elimination is spread across all untargeted elites, weakening their individual incentives to join a collective protest. Unpredictability can therefore sustain multi-period purges that eliminate a larger share of agents.

We formalize this mechanism in a dynamic infinite-horizon game between a principal and a finite group of forward-looking agents whom she may repeatedly target for elimination (or equivalently, purge). While in power, the principal's flow payoff is single-peaked in the number of surviving agents, capturing a trade-off between consolidating authority and retaining a base large enough to govern. Agents receive a flow payoff while alive and zero thereafter. In each period, the principal targets a subset of surviving agents for elimination. After observing the target set, all surviving agents simultaneously choose whether to protest. A protest that reaches a known participation threshold succeeds with positive probability. A successful protest permanently stops the purge and leaves the principal with zero continuation payoff. If such a protest fails, targeted and protesting agents are eliminated, and the game continues with the remaining agents. We focus on environments in which defeat is sufficiently costly that the principal prefers forgoing elimination to choosing a target set that induces a protest with a positive probability of success.

To abstract from coordination frictions and isolate the role of target uncertainty, we restrict attention to subgame-perfect equilibria (SPE) that are robust to profitable joint one-shot deviations by coalitions of agents, which we call coalition-proof SPE. We compare two polar benchmarks for target predictability: one in which agents can identify who would be targeted next if the purge continued, and one in which all untargeted agents perceive the same risk of future targeting. In the baseline model, we capture these environments through *predictable-order* and *unpredictable-order* coalition-proof SPE, as if the principal followed a prespecified priority order or selected targets uniformly at random, respectively. As the foundation in Section 6 clarifies, however, the

economic distinction behind these labels can be interpreted as concerning agents' information, not the principal's literal targeting procedure.

The contrast between the two environments is stark. Under predictable-order targeting, the purge lasts at most one period, and the number of agents eliminated remains below the protest threshold. Indeed, agents who anticipate being eliminated next prefer to join current targets in a protest, even if doing so entails a risk of immediate elimination. Anticipating such protests, the principal eliminates the maximum number of agents that avoids a potentially successful protest in the first period and then stops permanently. In this sense, there will be an infinite sequence of *blocking population sizes* that occur at intervals equal to the protest threshold, at which any further elimination would induce a potentially successful protest. Predictability therefore protects agents by allowing them to identify and assemble blocking coalitions.

Under unpredictable-order targeting instead perceived near-term risk of elimination is spread across all untargeted agents, so even the agents who will in fact be targeted next perceive the same risk as every other untargeted agent. This fragments opposition and makes blocking coalitions harder to assemble, allowing the principal to sustain longer sequences of eliminations. We show that all coalition-proof SPE with unpredictable-order targeting induce the same path of target counts and population sizes. This path is governed by a finite sequence of blocking population sizes, analogous to the blocking population sizes under predictable-order targeting, at which any further elimination would induce a potentially successful protest. These sizes are independent of the initial population. For each initial population above the principal's ideal size, call the largest blocking size weakly below it the *final blocking size*. The principal sequentially targets and eliminates agents until the population reaches the final blocking size, after which elimination stops permanently. Until then, the principal follows what we call the *greedy* elimination strategy in all but at most two periods, targeting as many agents as possible while ensuring that the targeted agents alone cannot mount a potentially successful protest.

Intuitively, slower elimination could raise untargeted agents' continuation value and discourage them from joining a protest, but two forces limit this motive. First, when the current population is far above the final blocking size, untargeted agents face a low near-term risk of being targeted, since in any period the principal optimally targets fewer agents than the protest threshold. Untargeted agents are therefore unwilling to join a protest, so delay is unnecessary. At the other

extreme, when the population is close to the final blocking size, targeting more agents increases each untargeted agent's probability of belonging to the final blocking coalition (i.e., the group that is never targeted), reaching one when the number of untargeted agents is exactly the final blocking size. Faster elimination therefore weakens protest incentives, so delay is again unnecessary. Slower elimination can help only in a narrow intermediate region, where the continuation value under greedy elimination would fall below the value of protesting.[1] Consequently, genuine delay can occur at most once along the elimination path; the only other departure from greedy elimination is the final adjustment needed to reach the final blocking size exactly.

Our equilibrium characterizations yield several comparative statics (Propositions 4 and 5). Under predictable-order targeting, blocking population sizes occur at intervals equal to the protest threshold. Hence the number of agents eliminated remains uniformly bounded below the protest threshold regardless of the initial population. Under unpredictable-order targeting, consecutive blocking sizes are at least as far apart and their distance grows at higher population sizes, allowing for more elimination. For initial populations large enough, unpredictable-order targeting yields a higher payoff for the principal. We further show that under unpredictable-order targeting, the blocking population sizes approximately scale linearly with the protest threshold. A final implication of our analysis is that decreasing the protest threshold can lead to an increase in number of agents the principal can eliminate.

The broader lesson is that unpredictability can sustain a dynamic form of divide and conquer even when agents can coordinate profitable joint deviations. The principal cannot commit to spare survivors after current targets are eliminated, and target predictability determines whether this continuation threat is concentrated on identifiable agents or diffused across the surviving group. This perspective helps explain why authoritarian purges may appear arbitrary even when rulers have underlying priorities, and why insiders may remain compliant despite knowing that compliance offers little protection.

We illustrate the mechanism through several historical episodes. During Stalin's purge of

[1] Within the unpredictable-order equilibrium class, slower elimination cannot be sustained by punishing deviations with worse continuation payoffs, because the principal's equilibrium continuation payoff is uniquely pinned down after every history. Outside this class, history-dependent punishments can support a broader set of equilibrium values; see Section 7.

the Red Army officer corps, senior officers understood that loyalty and participation in past repression offered little protection, yet future targets remained difficult to predict and effective resistance failed to emerge (Zakharov and Sonin, 2024, Montefiore, 2003). Saddam Hussein's 1979 single-day purge of the Ba'ath Party leadership provides a compressed illustration of the same logic: names were disclosed sequentially while officials not yet named remained compliant (Matthews, 2018). The fall of Robespierre and the events of 9 Thermidor illustrates the cost of mismanaging expectations. Once a sufficiently large set of revolutionary leaders believed they might be targeted soon, they coordinated against Robespierre (Linton, 2013).

The remainder of the paper is organized as follows. Section 2 discusses the related literature. Section 3 presents the model and solution concept. Section 4 characterizes equilibrium under predictable and unpredictable-order targeting and compares their implications for the share of agents purged and the principal's payoff. Section 5 relates the mechanism to historical purges. Section 6 provides the heterogeneous-priority foundation for the two benchmark environments. Section 7 discusses extensions, and Section 8 concludes.

2 Literature

Our paper relates to several strands of research in political economy. A first strand studies repression and elite purges in authoritarian regimes. Tyson (2018) analyzes the agency problem between an autocrat and her repressive apparatus. Montagnes and Wolton (2019) study mass purges as a device to create accountability among subordinates. Zakharov and Sonin (2024) provide empirical evidence from the Great Terror and distinguish between informed and preventive purges. By contrast, we abstract from agency frictions within the regime and from the ruler's information about potential coup plotters and focus instead on how uncertainty about future purge targets shapes protest incentives among elites and the endogenous path of eliminations pursued by the ruler.

Closely related is Che et al. (2026), who study divide-and-conquer when potential victims are uncertain about the ultimate scope of an aggressor's campaign. As in our model, uncertainty weakens coordination among potential victims, but the object of uncertainty differs: they hold the aggressor's ranking of victims fixed and study uncertainty about how many will ultimately be

targeted, whereas we hold the desired number of targets fixed and commonly known and study uncertainty about the targeting order. Also related is the analysis in Zeilberger (2024), which considers a version of De Mesquita et al. (2005) in which elites face uncertainty about their future inclusion in the ruling coalition. There, uncertainty plays a fundamentally different role: it makes it harder for the leader to maintain elite support because elites are uncertain about whether they will remain in the coalition. As a result, uncertainty harms, rather than benefits, the principal.

Our paper is also closely related to the literature on coalition formation and bargaining with an endogenous status quo, including Bernheim et al. (2006), Acemoglu et al. (2008), Diermeier and Fong (2009), Acemoglu et al. (2012), Diermeier and Fong (2011), Diermeier et al. (2017), Ali et al. (2019), Iaryczower and Oliveros (2023), and Ali et al. (2023). In these models, forward-looking agents internalize how current redistribution or elimination decisions affect their future bargaining positions, which disciplines equilibrium proposals. Acemoglu et al. (2008) study a dynamic environment in which a ruling coalition is constructed over time through supermajority-approved elimination proposals. A central result is that the process converges in a single transition to a unique self-enforcing coalition that cannot be successfully challenged in continuation play. Diermeier and Fong (2011) analyze an infinite-horizon agenda-setting model with an evolving status quo and show that the possibility of future revision induces non-proposers to protect one another, limiting the proposer's ability to form minimal winning coalitions and yielding more egalitarian outcomes. Ali et al. (2023) consider a finite-horizon agenda-setting model with an evolving status quo and provide conditions under which the agenda-setter can achieve her preferred policy even when voters are sophisticated.[2]

Relative to this literature, our main contribution is to show how uncertainty about the order of targeting fundamentally shapes the elimination path by weakening agents' forward-looking incentives to coordinate on blocking coalitions. When the targeting order is fixed and known, elimination collapses to a single period, echoing Acemoglu et al. (2008). When the targeting order is unpredictable, instead, longer elimination chains can be sustained. In doing so, we provide a

[2]At the heart of the difference between our predictable-order elimination environment and Ali et al. (2023) is that, in their setting, the principal retains proposal power even after a failed proposal, whereas a successful protest in our model permanently ends the principal's ability to eliminate agents. Appendix A.1 describes the connection in greater detail.

novel understanding of how randomness can be strategically valuable for the principal, characterize a class of subgame-perfect equilibria in this environment, and connect the mechanism to historical episodes of repression.

Our focus on the effects of randomization also connects our paper to results about mixed-strategy equilibria in the bargaining literature, such as Kalandrakis (2004) and Diermeier and Fong (2009). In Diermeier and Fong (2009), mixing over future coalition partners allows the agenda-setter to reach her most preferred policy within three periods, in contrast to pure-strategy equilibria in which she may be severely constrained. Beyond the difference in substantive focus, the basic mechanics of this mixing, as well as the extreme proposal power it confers, are different from the randomization in our setting.[3]

Our paper also relates to work on dynamic collective action, such as Shadmehr and Bernhardt (2019) and Gieczewski and Kocak (2024). Shadmehr and Bernhardt (2019) study how revolutionary vanguards shape participation in collective action, while Gieczewski and Kocak (2024) show that expectations of future protest can reduce citizens' willingness to protest today. Our mechanism highlights a complementary force: predictable future targeting can strengthen protests by identifying the agents with the greatest incentive to join those currently targeted, whereas unpredictability weakens that force by diffusing the perceived risk of being targeted in the future.

Finally, our coalition-proof refinement is connected to Ali and Liu (2025), who introduce the solution concept of perfect coalitional equilibria (PCE). While we study a dynamic rather than a repeated game (since the set of eliminated agents evolves over time), our coalition-proofness is a natural analogue to the PCE concept introduced in Ali and Liu (2025).

3 Model

We study a dynamic game in which a principal seeks to eliminate a subset of agents, whose only protection is the ability to organize a protest that can permanently halt further eliminations.

Time, players, and state. Time is discrete and infinite horizon, indexed by $t \in \{0, 1, \dots\}$. Initially, there is a *principal* and a finite collection of homogeneous agents $\mathcal{N}_0 = \{1, ..., N_0\}$.

[3] See Appendix Section A.1 for a more detailed discussion of the connection.

For each period t, let $\mathcal{N}_t \subseteq \mathcal{N}_0$ denote the set of surviving agents, i.e., those not eliminated in earlier periods, and let $N_t \equiv |\mathcal{N}_t|$ denote its size. Let $\Omega_t \in \{0,1\}$ be an indicator variable capturing whether a successful protest has occurred before period t: $\Omega_t = 0$ means that no successful protest has yet occurred, and $\Omega_t = 1$ means that a successful protest has already occurred. Initially, $\Omega_0 = 0$. All players are risk-neutral and share discount factor $\delta \in (0,1)$.

In each period t, each surviving agent receives flow payoff 1, and each eliminated agent receives flow payoff 0. If $\Omega_t = 0$, the principal's flow payoff at population size n is $u(n)$, where $u : \mathbb{N} \to \mathbb{R}_+$ is strictly single-peaked at N^*:

$$u(n+1) > u(n) \quad \text{for } n < N^*, \qquad u(n+1) < u(n) \quad \text{for } n \geq N^*.$$

If $\Omega_t = 1$, the principal's flow payoff is zero. This payoff structure captures a trade-off between consolidating power and retaining a sufficiently large base from which to govern.

Actions and state transitions. At the beginning of each period t with $\Omega_t = 0$, the principal chooses a set $\mathcal{K}_t \subseteq \mathcal{N}_t$ of agents to *target* for elimination, and we write $k_t \equiv |\mathcal{K}_t|$. After observing $\mathcal{K}_t$, each agent $i \in \mathcal{N}_t$ decides whether to protest or not, $a_i \in \{0,1\}$, with action $a_i = 1$ denoting *protest* and $a_i = 0$ denoting *not protest*. Let $\mathcal{P}_t \subseteq \mathcal{N}_t$ denote the set of protesters in period t, and let $p_t \equiv |\mathcal{P}_t|$ denote the size of the protest.

A protest in period t is *potentially successful* if its size is at least $\overline{k}$, i.e., $p_t \geq \overline{k}$, for $\overline{k} \in \mathbb{N}$ with $\overline{k} \geq 2$. Denote by $\hat{\mathcal{P}}_t$ the set of participants in a potentially successful protest. Such a protest succeeds if and only if an exogenous binary state $\omega_t \in \{0,1\}$, realized after actions are taken, equals 1; otherwise it fails. The states are independent across periods, with $\Pr(\omega_t = 1) = \rho \in (0,1)$.[4]

If the protest succeeds, the game proceeds to period $t+1$ with

$$\mathcal{N}_{t+1} = \mathcal{N}_t \qquad \text{and} \qquad \Omega_{t+1} = 1.$$

[4] In our setting, the risk of elimination following a failed protest is the only cost of participation. If $\rho = 1$ and $N_0 \geq \overline{k}$, any targeted agent could form a surely successful protest with $\overline{k} - 1$ others. They all would be strictly better off, as long as they perceive any positive probability of elimination if they do not protest, and so coalition-proofness would rule out any equilibrium in which the principal is expected to ever eliminate at least $\overline{k}$ agents. A direct cost of protesting would play a similar role to failure risk.

If no protest succeeds, all targeted agents and all participants in a potentially successful protest are permanently eliminated, so[5]

$$\mathcal{N}_{t+1} = \mathcal{N}_t \setminus (\mathcal{K}_t \cup \hat{\mathcal{P}}_t) \qquad \text{and} \qquad \Omega_{t+1} = 0.$$

In each period t such that $\Omega_t = 1$, the principal cannot target any agent and agents do not protest. Accordingly, we set $\mathcal{K}_t = \emptyset$, $\mathcal{P}_t = \emptyset$, $\Omega_{t+1} = \Omega_t = 1$, and $\mathcal{N}_{t+1} = \mathcal{N}_t$. If $\mathcal{K}_t = \emptyset$, we assume that protest does not occur, so $\mathcal{P}_t = \emptyset$.[6]

Histories. Given the initial set of agents $\mathcal{N}_0$, a *history* up to period t is

$$h^t \equiv (\mathcal{K}_s, \mathcal{P}_s, \omega_s)_{s=0}^{t-1}.$$

Denote by H the set of all histories. Each history $h^t \in H$ induces a state $(\mathcal{N}_t(h^t), \Omega_t(h^t))$, where

$$\Omega_t(h^t) = \max_{0 \le s < t} \{\mathbf{1}\{p_s \ge \overline{k}\}\, \omega_s\} \qquad \textit{and} \qquad \mathcal{N}_t = \mathcal{N}_0 \setminus \bigcup_{0 \le s < t : \Omega_{s+1} = 0} (\mathcal{K}_s \cup \hat{\mathcal{P}}_s).$$

We suppress the dependence of the state on h^t when it causes no confusion.

A *history with proposal* (i.e., after the principal's move at t) is

$$\ell^t \equiv \big(h^t, \mathcal{K}_t\big).$$

Denote by L the set of all histories with proposal.

Strategies. A (behavioral) strategy for the principal is a mapping

$$\sigma^p : H \to \Delta\big(2^{\mathcal{N}_0}\big)$$

such that, after any history $h^t \in H$, the support of $\sigma^p(h^t)$ is contained in $2^{\mathcal{N}_t(h^t)}$, and if $\Omega_t(h^t) = 1$, $\sigma^p(h^t)$ assigns probability 1 to $\emptyset$.

[5] An alternative specification would also eliminate all agents who choose to protest, even when participation remains below $\overline{k}$. This would lead to identical results if protest decisions were made sequentially. With simultaneous protest decisions, some proposal subgames under this alternative specification need not admit a pure-strategy coalition-proof equilibrium as we define it, and our specification above avoids this issue.

We also note that the formulation in our paper leads to identical results as a formulation in which agents who protest but not targeted are not eliminated, but instead pay a fixed cost to protest.

[6] Allowing agents to protest when no elimination is proposed would not affect the analysis.

A (behavioral) strategy for agent i is a mapping

$$\sigma^i : L \to \Delta(\{0,1\}),$$

with the convention that $\sigma^i(\ell^t) = 0$ whenever $i \notin \mathcal{N}_t$ or $\Omega_t = 1$ or $\mathcal{K}_t = \emptyset$.[7]

Continuation payoffs. Fix a strategy profile $\sigma = (\sigma^p, (\sigma^i)_{i \in \mathcal{N}_0})$. Let $\mathbb{E}_\sigma$ denote expectation with respect to the distribution of the process $(\mathcal{N}_t, \Omega_t)_{t \geq 0}$ induced by σ and $(\omega_t)_{t \geq 0}$. The principal's continuation payoff after history h^t is

$$U^p(\sigma \mid h^t) = \mathbb{E}_\sigma \Big[\sum_{s=t}^{\infty} \delta^{s-t} \mathbf{1}\{\Omega_s = 0\} u(N_s) \;\Big|\; h^t \Big].$$

For agent i, the continuation payoff after history with proposal ℓ^t is

$$U^i(\sigma \mid \ell^t) = \mathbb{E}_\sigma \Big[\sum_{s=t}^{\infty} \delta^{s-t} \mathbf{1}\{i \in \mathcal{N}_s\} \;\Big|\; \ell^t \Big].$$

We impose two assumptions.

Assumption 1. *Agents prefer participating in a protest that succeeds with probability ρ, to waiting one period and then being eliminated for sure:*

$$1 < \frac{\rho}{1-\delta}.$$

Assumption 1 ensures that the payoff of being eliminated next period, $1+\delta$, is lower than $1 + \frac{\rho\delta}{1-\delta}$, the expected payoff of participating today in a protest of size at least $\overline{k}$; the protest succeeds with probability ρ, yielding survival thereafter, and fails with probability $1-\rho$, leading to immediate elimination. If Assumption 1 fails, the model is uninteresting, as only targeted agents would have an incentive to protest.

Assumption 2. *The principal strictly prefers to remain forever at the initial population size than to choose any action that induces a protest of size $p_t \geq \overline{k}$, i.e.,*

$$u(N_0) > (1-\rho) u(N^*).$$

Assumption 2 ensures that the principal prefers not eliminating any agents to triggering a potentially successful protest. It captures settings in which the consequences of a successful protest are severe, for example because the ruler risks being overthrown or killed.

[7] With abuse of notation, we denote by a_i the mixed action that assigns probability 1 to $a_i \in \{0,1\}$

Sequential optimality and solution concept. Fix a strategy profile $\sigma = (\sigma^p, (\sigma^i)_{i \in \mathcal{N}_0})$. The principal's strategy σ^p is *sequentially optimal* if, for every history h^t with induced state $(\mathcal{N}_t, \Omega_t)$, and for every alternative strategy $\tilde{\sigma}^p$ that agrees with σ^p up to $t-1$,

$$U^p(\sigma^p, \sigma^{-p} \mid h^t) \;\geq\; U^p(\tilde{\sigma}^p, \sigma^{-p} \mid h^t).$$

Similarly, agent i's strategy σ^i is *sequentially optimal* if, for every history with proposal ℓ^t such that $i \in \mathcal{N}_t$, and every alternative strategy $\tilde{\sigma}^i$ that agrees with σ^i up to $t-1$,

$$U^i(\sigma^i, \sigma^{-i} \mid \ell^t) \;\geq\; U^i(\tilde{\sigma}^i, \sigma^{-i} \mid \ell^t).$$

A strategy profile σ is a *subgame perfect equilibrium (SPE)* if every player's strategy is sequentially optimal after every history at which that player moves.

Coalition-proof refinement. We refine SPE by imposing robustness to joint one-shot deviations by coalitions of agents.

Definition 1. *A strategy profile σ is a* coalition-proof SPE *if:*

1. *σ is an SPE;*
2. *there exists no history with proposal ℓ^t and nonempty coalition $C \subseteq \mathcal{N}_t$ that has a joint one-shot deviation at ℓ^t that strictly increases the payoff of every $i \in C$;*
3. *for every agent i and history with proposal ℓ^t such that $i \in \mathcal{N}_t$, if agent i is indifferent between protesting and not protesting, then $\sigma^i(\ell^t)$ assigns probability 1 to not protest.*

The third condition only resolves nongeneric indifferences and is incorporated into the definition to avoid additional notation.[8]

We use coalition-proofness to abstract from coordination failures at the protest stage, which are important but distinct from the mechanism we seek to isolate. To see the issue, suppose that $N_t - N^* \geq \overline{k}$. There is always an SPE in which the principal eliminates all $N_t - N^*$ agents immediately. This is because no agent can reach the participation threshold unilaterally if they expect no one else to protest. Coalition-proofness eliminates such equilibria.

[8] Mixed strategies for agents are ruled out except for a nongeneric set of parameters even without the tie-breaking rule, by the possibility of joint deviations.

To study how uncertainty about future purge targets shapes protest incentives and the endogenous path of eliminations, we focus on two classes of coalition-proof SPE:[9]

- A *predictable-order coalition-proof SPE* is a coalition-proof SPE for which there exists a strict total order $\succ$ on $\mathcal{N}_0$ such that, after every history with $\Omega_t = 0$, conditional on targeting k_t agents, the principal targets the k_t highest-ranked surviving agents according to $\succ$.
- An *unpredictable-order coalition-proof SPE* is a coalition-proof SPE such that, after every history with $\Omega_t = 0$, conditional on targeting k_t agents, the principal chooses uniformly at random among the k_t-size subsets of $\mathcal{N}_t$.

To avoid confusion, we emphasize that these are both classes of *coalition-proof SPE*: we do not restrict the deviations available to the principal.

The two benchmarks capture how perceived near-term risk of eliminated is distributed across survivors: a predictable order concentrates it on identifiable future targets, whereas an unpredictable order diffuses it across untargeted agents. In Section 5, we relate the two classes to a number of historical elite purges.

Foundation for the equilibrium classes. Section 6 provides a microfoundation for the two equilibrium classes in a richer environment in which agents differ in the cost they impose on the principal. The principal targets more costly agents first in both cases, but which outcome emerges depends on agents' information. When these costs are common knowledge, the predictable-order outcome emerges; when they are observed only by the principal and all untargeted agents perceive the same elimination risk, the unpredictable-order outcome emerges. Thus, the two equilibrium classes isolate the effect of transparency about future targets: what matters is not whether the principal literally randomizes or uses a fixed order, but how well agents can predict who will be targeted next. We should expect the predictable-order outcome when the leader's priorities can be inferred from observable characteristics such as faction, rank, institutional affiliation, or documented past conduct, and the unpredictable-order outcome when her ranking depends on privately observed assessments.

[9] See Section 7 for a brief discussion of the set of all coalition-proof SPE that can arise.

4 Coalition-Proof Subgame Perfect Equilibria

4.1 Predictable-order targeting

Under predictable-order elimination, agents who expect to be targeted next join current protesters. As a result, the purge collapses to a single period, and eliminates at most $\overline{k} - 1$ agents.

Fix a priority order $\succ$ and let $E_t := \max\{0, N_t - N^*\}$ be called the *excess population* in period t. For each nonnegative integer E, let $r(E) := E \bmod \overline{k} \in \{0, \ldots, \overline{k} - 1\}$ denote the remainder when E is divided by $\overline{k}$, and set $r_t := r(E_t)$.

Proposition 1 (Predictable-order targeting)**.** *Under Assumptions 1 and 2, in every predictable-order coalition-proof SPE with priority order $\succ$, the principal targets the r_0 highest-ranked agents in period 0 and then stops permanently. No protest succeeds, and $N_t = N_0 - r_0$ for any $t \geq 1$.*

Thus, the purge lasts at most one period, and is limited to eliminating no more than $\overline{k} - 1$ agents, independently of the initial population size. This result echoes the one-step transition in Acemoglu et al. (2008), with agents blocking the principal because they anticipate becoming future targets. As we show next, this conclusion fails under unpredictable-order elimination.

Sketch of the proof. Predictability allows next-period targets (if any) to identify themselves, giving them an incentive to join current targets in protest and constraining the principal. We use this idea in a proof by induction on the excess population size, denoted E, showing that the principal targets the $r(E)$ highest-ranked survivors and then stops permanently. If $E = 0$, the principal stops, as any further elimination would lower her payoff. If $1 \leq E < \overline{k}$, she eliminates all $r(E) = E$ excess agents and reaches her ideal size N^* in one period: the targets are too few to mount a potentially successful protest alone, while untargeted agents anticipate no further elimination and therefore prefer not to join.

Next suppose that $E = q\overline{k} + r$ for some $q \geq 1$, where $r := r(E)$, and that the claim holds for every smaller excess population. Targeting $m > r$ agents induces a potentially successful protest, either by the targets alone when $m \geq \overline{k}$, or with the participation of untargeted agents when $m \in \{r+1, \ldots, \overline{k}-1\}$. Indeed, in the latter case, the induction hypothesis implies that the $r(E - m) = \overline{k} + r - m$ highest-ranked untargeted agents would be eliminated next period if the current

targets were eliminated. If the resulting protest had fewer than $\overline{k}$ participants, Assumption 1 implies that a coalition of current and prospective targets could profitably deviate by joining it and bringing participation to $\overline{k}$. Coalition-proofness therefore implies that the proposal induces a potentially successful protest. Assumption 2 thus makes every target count above r suboptimal for the principal. If $r = 0$, the principal therefore stops. If instead $r > 0$, targeting exactly r agents induces no such protest because $E - r = q\overline{k} < E$ has remainder zero, so the induction hypothesis implies that no further elimination occurs. Untargeted agents therefore do not join, while the $r < \overline{k}$ targets alone cannot mount a potentially successful protest. Targeting fewer agents merely postpones reaching $q\overline{k}$ agents and is strictly worse because u is strictly decreasing above N^*. Hence, the principal targets exactly the $r(E)$ highest-ranked surviving agents and then stops permanently.

Example 1. *Consider the case $N^* = 2$ and $\overline{k} = 3$. For $N_0 \in \{3, 4\}$, $E_0 = N_0 - 2 \in \{1, 2\}$ and $r(E_0) = E_0$, so the principal eliminates one agent if $N_0 = 3$ or two agents if $N_0 = 4$, leaving $N^* = 2$ agents surviving. At $N_0 = 5$, however, $E_0 = 3$ and $r(E_0) = 0$, so the principal eliminates no agents. To see why, note that targeting $m \in \{1, 2\}$ agents leaves $3 - m$ excess agents who know that they would be eliminated next period and therefore prefer to join the protest today rather than wait to be eliminated. Thus, by Assumption 1 and coalition-proofness, a potentially successful protest of size $m + (3 - m) = 3$ forms, making a proposal to eliminate 1 or 2 agents suboptimal by Assumption 2. Similarly, the principal would not target three or more agents, since coalition-proofness implies that the targets alone would mount a potentially successful protest. At $N_0 \in \{6, 7\}$, $E_0 = N_0 - 2 \in \{4, 5\}$ and $r(E_0) = E_0 - 3 \in \{1, 2\}$, so the principal eliminates one agent if $N_0 = 6$ or two agents if $N_0 = 7$, reaching $N_1 = 5$ in one step and then stopping. A similar argument shows that the principal targets no agents at $N_0 = 8$, and the pattern repeats.*

As we show next, under unpredictable-order targeting the purge can last for multiple periods, and its depth can grow with the initial population.

4.2 Unpredictable-order targeting

When future targets are unpredictable, the risk of near-term targeting is spread uniformly across all untargeted agents, weakening their incentives to join current targets in protest. As we show,

this can sustain purges over multiple periods and eliminate many more agents than predictable-order targeting can.

Because current targets always benefit from a potentially successful protest, the key question is whether untargeted agents would join them. Their decision depends on the probability and timing of future elimination if the principal is allowed to continue the purge. Both are endogenous and depend on the principal's continuation strategy and on whether further targeting would induce a potentially successful protest.

As a first step, we show that every unpredictable-order coalition-proof SPE induces the same path of target counts, and that the equilibrium target count and agents' continuation values depend only on the current population size, and that the principal's continuation payoff is strictly decreasing in population size above N^* (Lemma 1 in the appendix). Combined with Assumption 2, this implies that the principal targets as many agents as possible without inducing a potentially successful protest.

Let $\kappa(n)$ denote the equilibrium target count with n survivors, and let $V(n)$ denote an agent's ex ante continuation payoff before the principal chooses the target count. As under predictable-order targeting, the principal targets $\kappa(n) = 0$ when $n \leq N^*$ and $\kappa(n) = n - N^*$ when $0 < n - N^* < \overline{k}$, without inducing any protest.

Suppose instead that $n \geq N^* + \overline{k}$. Targeting at least $\overline{k}$ agents induces a potentially successful protest by the targets alone, so consider targeting $m < \overline{k}$ agents. If the principal's proposed elimination is implemented without a potentially successful protest, an untargeted agent obtains $1 + \delta V(n-m)$, whereas joining a potentially successful protest yields $1 + \delta\rho/(1-\delta)$. Coalition-proofness and the tie-breaking rule therefore imply that the proposal avoids a potentially successful protest if and only if

$$1 + \delta V(n-m) \geq 1 + \delta \frac{\rho}{1-\delta}$$

If this inequality fails, a coalition of $\overline{k}$ targeted and untargeted agents would strictly benefit from protesting. Thus, for all $n \geq N^* + \overline{k}$ the principal optimally targets $\kappa(n) = 0$ agents if $V(n-m) < \frac{\rho}{1-\delta}$ for all $m \in \{1, \ldots, \overline{k} - 1\}$ and otherwise targets

$$\kappa(n) = \max \left\{ m \in \{1, \ldots, \overline{k} - 1\} : V(n-m) \geq \frac{\rho}{1-\delta} \right\}.$$

We now describe our main result, which we break into two propositions to aid readability.

Proposition 2 (Blocking population sizes)**.** *Under Assumptions 1 and 2, all unpredictable-order coalition-proof SPE induce the same path of target counts and population sizes* $(k_t, N_t)_{t\geq 0}$*. Moreover, there exists a unique finite set* $\mathcal{B} = \{B_0, B_1, \ldots, B_n\} \subseteq \mathbb{N}$*, independent of* u *and* N_0*, such that* $N^* = B_0 < B_1 < \cdots < B_n$*, and the following holds. For any initial population size* $N_0 \in \mathbb{N}$*, the equilibrium path reaches population size*

$$B^*_{N_0} := \begin{cases} N_0, & \text{if } N_0 < N^*, \\ \max\{B_j \in \mathcal{B} : B_j \leq N_0\}, & \text{if } N_0 \geq N^*. \end{cases}$$

in finite time, after which no further eliminations occur. Along the equilibrium path, no untargeted agent protests, and no protest succeeds. Equilibrium strategies are Markovian in N_t*.*

A population size $N_t = B$ is terminal either because the principal has reached or passed her ideal point, $B \leq N^*$, or because every further elimination would lower the continuation value of untargeted agents enough to induce them to join a protest. In the latter case, to avoid the risk of a successful protest, the principal refrains from targeting any agents once the population hits this threshold B. Crucially, if a size $\overline{k}$ protest is triggered by targeting k agents starting from a population size $N_t = B$, then targeting $k + m$ agents from a population size $N_t = B + m$ does so as well: both proposals leave the same number of agents $B - k$ after the current elimination, and hence generate the same continuation value for untargeted agents. Consequently, starting from any $N_t \geq B$, the principal can never eliminate more than $N_t - B$ agents without triggering a potentially successful protest.

The set $\mathcal{B}$ consists of the ideal population size N^* and precisely these endogenous blocking population sizes above it, at which any further targeting would make future elimination risk sufficiently high that untargeted agents would join the current targets in protest. Starting from any initial population $N_0 \geq N^*$, the principal eliminates agents until reaching the largest threshold $B \in \mathcal{B}$ such that $B \leq N_0$, at which point she stops permanently. Therefore, diffusion weakens resistance but does not eliminate it.

Our next result characterizes the speed of elimination. We begin by defining a greedy elimination strategy as the maximum pace consistent with avoiding the risk of a successful protest when untargeted agents do not join.

Definition 2. *We say that the principal uses the* greedy *elimination strategy in period t if she targets $\overline{k}-1$ agents.*

Thus, under the greedy strategy, the principal targets as many agents as possible while ensuring that the targeted agents alone cannot mount a successful protest. With this definition in hand, we can state our next result.

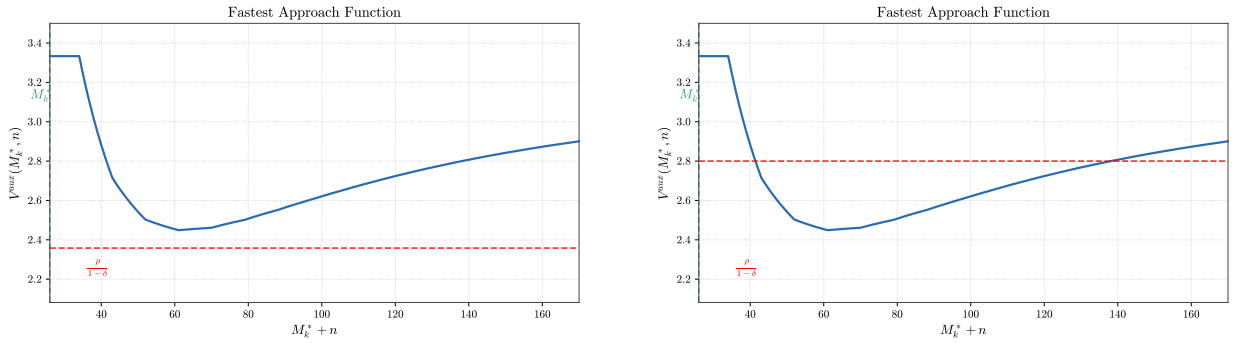
Proposition 3 (Minimal delay)**.** *Along the unique unpredictable-order coalition-proof SPE path of target counts and population sizes from N_0 to $B^*_{N_0}$, the principal uses the greedy elimination strategy in every period except at most two.*

Intuitively, a slower pace of elimination could help the principal by raising the continuation value of untargeted agents and thereby discourage them from joining a protest. Two forces limit this motive. First, when the current population is close to the blocking coalition size $B^*_{N_0}$, eliminating more agents today raises each untargeted agent's probability of belonging to the final coalition, and that probability reaches one exactly when the remaining untargeted agents form the final blocking coalition size. In this region close to the final blocking coalition, the continuation value of untargeted agents increases with the number of agents eliminated, so the principal has no incentive to delay and will target $k_t = \min\{N_t - B^*_{N_0}, \overline{k}-1\}$ agents. Second, when the population is far above the final blocking coalition size $B^*_{N_0}$, untargeted agents know that they are unlikely to be targeted soon, because the principal never targets more than $\overline{k}-1$ agents per period, since doing so would trigger a potentially successful protest. As a result, although their continuation value declines as elimination proceeds, that decline is too small for them to find it optimal to join a protest, even if they expect elimination to proceed at the maximum speed (i.e., greedily).

Taken together, we show these two forces imply that, under the greedy strategy, the continuation value of untargeted agents when the principal faces no risk of successful protest—which we call the *fastest approach function*—is quasi-convex in N_t along the path from N_0 to $B^*_{N_0}$, reaching its minimum at $n_{min} > B^*_{N_0}$.[10] This quasi-convexity implies that there is at most one contiguous region around n_{min} in which untargeted agents find it optimal to join a protest. However,

[10] More precisely, we define the fastest approach function from N_0 to $B^*_{N_0}$ as the continuation value of untargeted agents when the principal uses the greedy strategy in every period, except perhaps in the final period, in which the principal uses $(N_0 - B^*_{N_0}) \bmod (\overline{k}-1)$ to exactly reach $B^*_{N_0}$.

either this region is small enough that there is at most one period of delay, or it contains another element of $\mathcal{B}$, contradicting the fact that $B^*_{N_0}$ is the final blocking coalition size.



(a) No intersection

(b) Intersection

Figure 1: Fastest approach function (V^{aux}) v.s. value of protesting

Proof sketch and constructing B_j. We sketch here the recursive construction of thresholds B_j, which underlies the proof of both Propositions 2 and 3; these are in fact proved simultaneously. We initialize the construction with $B_0 = N^*$, and suppose that we have found values $(B_0, \ldots, B_{j-1})$ satisfying Propositions 2 and 3. To construct B_j, we compute the fastest approach function as defined in the previous paragraph, and if it never falls below the payoff from protesting, then we set $B_j = \infty$.[11] Otherwise, B_j is the first integer from which every feasible positive elimination would induce a population size at which the fastest approach function falls below the payoff from protesting. The fastest approach function in these two main cases is displayed in Figure 1, where we denote the fastest approach function by V^{aux}. It is straightforward to verify that for any $N \in (B_{j-1}, B_j)$, the greedy strategy is optimal for the principal, and hence the fastest approach function indeed describes the continuation value of an untargeted agent. [12]

[11] More generally, we also set $B_j = \infty$ when the fastest approach function falls below the protest payoff only over a region that the principal can jump over in a single period.

[12] This explanation is valid except for the possible one period of delay mentioned in the previous paragraph. In that case, the fastest approach function falls beneath the payoff to protesting but only for an interval of length smaller than $\overline{k}$. We formalize and deal with this case in the proof. In that case, the fastest approach function serves as a lower bound on the continuation value of an untargeted agent, and the solution in this case produces one of the non-greedy periods of elimination in Proposition 3.

Example 2. *Consider $\delta = 9/10$, $\rho = 3/4$, $\bar{k} = 3$, and $N^* = 3$. Then $B_0 = 3$; we now describe how to construct B_1. At $N_0 \in \{4, 5\}$, the principal eliminates $N_0 - 3$ agents, avoids protest, and hence $N_t = 3$ for all $t \geq 1$. Next, consider $N_0 = 6$ and suppose the principal targets 2 agents. An untargeted agent who is pivotal for forming a protest of size $\overline{k}$ obtains $1 + \frac{\delta}{1-\delta} \times \rho = 7.75$ from protesting and $1 + \delta\left[\frac{1}{1-\delta} \times \frac{3}{4} + \frac{1}{4}\right] \approx 8$ from not protesting. Thus, the principal eliminates 2 agents without protest, leaving $N_1 = 4$; she then eliminates 1 agent, leaving $N_2 = 3$, and stops. Next, consider $N_0 = 7$ and suppose that the principal proposes to eliminate 2 agents. A pivotal untargeted agent again receives 7.75 from protesting and $1 + \delta\left[\frac{1}{1-\delta} \times \frac{3}{5} + \frac{2}{5}\right] \approx 6.8$ from not protesting. Thus, she strictly prefers to protest if pivotal, so eliminating 2 agents induces a potentially successful protest, which is suboptimal for the principal.*

The natural question is whether, at $N_0 = 7$, the principal could avert protest by targeting only 1 agent. However, a pivotal untargeted agent who does not protest when 1 agent is targeted receives $1 + \delta\left[\frac{1}{1-\delta}\frac{3}{6} + (1+\delta)\frac{1}{6} + \frac{2}{6}\right] \approx 6.1$, less than she receives when 2 agents are targeted. Because her payoff from joining a potentially successful protest is unchanged, targeting only 1 agent also induces a potentially successful protest. Therefore, the principal cannot avert protest by targeting only 1 agent. Targeting at least 3 agents induces one by the targets alone and so the principal proposes no elimination, and $B_1 = 7$.

To understand why targeting fewer agents does not avert protest, observe that whether the principal targets 1 or 2 agents, only 3 agents survive the purge. Reducing the targets from 2 to 1 decreases an untargeted agent's chance of surviving the purge. When the population is close to B_0, as it is here, this survival effect dominates, so targeting fewer agents cannot reduce her willingness to protest. For sufficiently large populations, however, targeting more agents decreases an untargeted agent's continuation value. As the proof shows, this region of the value function does not determine the cutoffs in $\mathcal{B}$.

4.3 Comparative statics and the value of unpredictability

Our first comparative static result shows how unpredictability *typically* increases the total number of agents eliminated, which we call *purge depth*, and raises the principal's payoff.

For fixed primitives, Proposition 1 implies that an equivalent notion to the blocking population sizes from unpredictable-order elimination can be defined under predictable-order elimina-

tion as

$$\mathcal{B}^F := \left(B_q^F\right)_{q=0}^{\infty} = \left(N^* + q\overline{k}\right)_{q=0}^{\infty}.$$

The depth of a purge starting from any $N_0 \geq N^*$ is the distance between N_0 and the largest blocking size weakly below it.[13]

Since blocking population sizes under predictable- and unpredictable-order generally do not coincide, no point-wise ranking of purge depth is available: for instance, starting from $N_0 = B_j \in \mathcal{B}$ we have no elimination under unpredictable-order targeting but typically a positive purge depth under predictable-order targeting, apart for the case where $B_j \in \mathcal{B}^F$ as well.[14] Nevertheless, the spacing between blocking population sizes provides a useful comparison of purge depth across the two regimes.

Under predictable-order elimination, agents are able to form blocking coalitions with their neighbors in the elimination order. Hence, blocking population sizes recur every $\overline{k}$ agents, $B_{q+1}^F - B_q^F = \overline{k}$. This limits the purge depth to at most $\overline{k} - 1$ agents. Under unpredictable-order elimination, instead, the risk of being in the next targeted group is diffuse rather than concentrated among a few untargeted agents, weakening their incentive to join the protest. As a result, consecutive unpredictable-order blocking sizes are at least as far apart as their predictable-order counterparts, allowing for deeper purges and, at least for large N_0, benefiting the principal.

Proposition 4. *Suppose Assumptions 1 and 2 hold. Let $(B_0, ..., B_n)$, with $n < \infty$ and $B_0 = N^*$, denote the sequence of blocking population sizes under unpredictable-order coalition proof SPE in Proposition 2, and, by convention, set $B_{n+1} = \infty$. Then,*

- *For every $j \in \{0, \ldots, n\}$ and every $q \in \mathbb{N}$, $B_{j+1} - B_j \geq \overline{k} = B_{q+1}^F - B_q^F$, with $B_{j+1} - B_j > \overline{k}$ if and only if $B_j \geq \frac{\delta}{1-\rho} - 1$.*

- *There exists $\bar{N}$ independent of N_0 and u such that, if $N_0 > \bar{N}$, the principal's payoff is strictly*

[13] Of course, purge depth under predictable-order elimination coincides with r^* as defined in Section 4.1.

[14] Recall example 2, with $N^* = 3$, $\delta = 9/10$, $\rho = 3/4$, $\bar{k} = 3$. We found that $B_1 = 7$, i.e., the first blocking threshold after $B_0 = 3$ occurs with 7 agents remaining. Suppose that $N_0 = 7$. Under unpredictable-order coalition proof SPE, the principal is unable to eliminate any agents, so $N_t = N_0 = 7$ for all $t \geq 1$. But under predictable-order coalition proof SPE, since $r^* = 1$, the principal is able to eliminate 1 agent, so $N_t = 6$ for all $t \geq 1$. Thus, the principal strictly prefers predictable-order to unpredictable-order coalition-proof SPE.

higher under unpredictable-order coalition-proof SPE than under predictable-order coalition-proof SPE.

The first part implies that if $B_j \geq \frac{\delta}{1-\rho} - 1$, then $B_j + \overline{k} < B_{j+1}$, so that if $N_0 \in \{B_j + \overline{k}, ..., B_{j+1} - 1\}$, unpredictable-order purge depth is at least $\overline{k}$, whereas predictable-order purge depth is at most $\overline{k} - 1$. Except for the possibility that the principal's first elimination is non-greedy or that N_0 is within $\overline{k} - 1$ of $B^*_{N_0}$—in which case predictable-order coalition-proof SPE may or may not lead to higher payoff than unpredictable-order coalition-proof SPE—the principal will therefore have a higher payoff under unpredictable-order than predictable-order elimination. The second part of the proposition shows that these two qualifiers can be dropped for N_0 large enough. Unpredictability is therefore typically valuable, and unambiguously valuable to the principal when N_0 is large enough.

The next comparative static examines how the principal's strength, measured by the largest protest she can withstand with certainty, $\overline{k} - 1$, affects the blocking thresholds and hence purge depth under unpredictable-order elimination.

Denote by $\mathcal{B}(\overline{k}) = (B_0, B_1, ..., B_n)$ the tuple of blocking thresholds in Proposition 2 when the protest threshold parameter is $\overline{k}$. Recall that as long as Assumption 2 is satisfied, $\mathcal{B}(\overline{k})$ is independent of u; we therefore use $\mathcal{B}(\overline{k})$ to refer to the tuple of thresholds that prevails under Assumption 2.

Proposition 5. *Assume $N^* = 0$, and Assumption 1 holds. For every integer $\kappa \geq 1$ there exists $\overline{N}(\kappa, \delta, \rho)$ such that, for all $\overline{k} > \overline{N}(\kappa, \delta, \rho)$, the tuple $\mathcal{B}\big(\kappa(\overline{k} - 1) + 1\big)$ has the same number of components as $\mathcal{B}(\overline{k})$ and*

$$\frac{\mathcal{B}\big(\kappa(\overline{k} - 1) + 1\big)}{\overline{k}} = \frac{\kappa\, \mathcal{B}(\overline{k})}{\overline{k}} + o(1), \tag{1}$$

componentwise as $\overline{k} \to \infty$, except on a non-generic set of (ρ, δ).

In words, when $N^* = 0$, scaling up the principal's strength $\overline{k} - 1$ by κ scales up every blocking threshold by approximately κ.[15]

One interesting implication of Proposition 5 is that, even for N_0 significantly above $\overline{k}$, *decreasing* $\overline{k}$ (strengthening the agents, since fewer are needed to succeed) can lead to an *increase*

[15] A version of the result can be stated for the case $N^* > 0$, but is slightly more cumbersome.

in the number of agents the principal is able to eliminate.[16] This is substantively related to results in Diermeier et al. (2017), showing how increasing supermajority requirements may increase expropriation.[17]

5 Uncertainty in authoritarian regimes: examples

Authoritarian regimes—in which purges, enemy elimination, and expropriation take center stage—can be highly unpredictable. During the Great Terror, orchestrated by Stalin in the 1930s, the identities of individual targets were often difficult to predict. Purges of the broad population were implemented *by quota*, without clear criteria for selecting individuals beyond broad categories (Gregory, 2009). Mobutu Sese Seko, president of the Democratic Republic of Congo (Zaire, under Mobutu), was known to regularly shuffle ministers in and out of office, evidently to stop them from amassing independent power; as Haskin (2005) writes,

> *But, as many authors show, it was the uncertainty of Mobutu's system that encouraged politicians and officeholders to plunder and steal as they did. With Mobutu's constant reshuffling of his cabinet and all-too-frequent dismissals, it was unclear how long any of them would be allowed access to sources of wealth.*

These cases illustrate a general uncertainty that pervades authoritarian regimes, especially as they eliminate opposition. Masha Gessen, writing in the *New York Times* in 2026,[18] crystallizes the uncertainty faced by those living under terror regimes:

> *When we think of the terror regimes of the past, it is tempting to superimpose a logical narrative on them, as though totalitarian leaders had an extermination to-do list and worked their way through it methodically. This, I think, is how most people understand Martin Niemöller's classic poem "First They Came." In reality, though, the people living under those regimes never knew which group of people would be designated an enemy of the state next.*

[16] If $N_0 = \overline{k}$, decreasing $\overline{k}$ increases purge depth from 0 to 1, and the statement holds.

[17] Even under predictable-order elimination, it is possible to recover the result that decreasing $\overline{k}$, strengthening the agents since fewer are needed to succeed, can increase the number of agents the principal is able to eliminate. For instance, suppose $N = 5$ and $\overline{k} = 5$. Then $r^* = 0$ and no agents can be eliminated. If we reduce $\overline{k}$ to 4, then $r^* = 1$ and one agent can be eliminated.

[18] https://www.nytimes.com/2026/01/24/opinion/state-terror-has-arrived.html

Our paper places this uncertainty at the center of the analysis, showing how unpredictability regarding the identity of future purge targets undermines resistance. At the same time, historical episodes suggest that mismanaging the expectations of potential purge targets can have severe consequences for authoritarian leaders themselves. During the Reign of Terror of the French Revolution, Maximilien Robespierre, along with other members of the Committee of Public Safety, orchestrated the elimination of (among others) Girondins, Dantonists and Hébertists, groups spanning the ideological spectrum. Robespierre himself was eventually arrested and executed by a coalition that include members of the Committee of Public Safety, who feared their own downfall. Linton (2013) writes about the Thermidorian reaction

> *Most of the members of the Committee of Public Safety along with the Committee of General Security had actually signed the arrest warrant for the Dantonists ... Now they feared that such tactics might be used against themselves. Several of them were eyeing Robespierre warily, afraid that he was about to precipitate another purge ... and that their own names might be on his list.*

and quoting Pierre Louis Prieur, a member of the Committee of Public Safety,[19]

> *Storm clouds banked up over our heads, and at each moment we came closer to an inevitable crisis. However the work of the Committee did not suffer. In the midst of our anguish, uncertain whether the hour would come that would see us sent before the Revolutionary Tribunal, to go from there to the scaffold, without perhaps having the time to say goodbye to*

[19]Robespierre is commonly believed to have made an error by failing to narrow the scope of the purge effectively publicly announced on 9 Thermidor, leaving too many with the belief that they may be on his list of immediate targets. It is also established that a number of his enemies were either clearly identified by Robespierre or could identify themselves. Linton (2013) writes, "...At the same time he [Robespierre] accused several of his fellow Committee members, more or less openly, of being part of the plot against him. He compounded this extraordinarily dangerous public declaration by refusing to give the names of the men on the list of those whose arrest he sought, leaving many of his terrified listeners to speculate whether they themselves, or their friends, appeared on it..." Thus, it seems plausible that even naming the list of immediate purge targets which, according to Saint-Just's undelivered speech, was confined to just Jean-Marie Billaud-Varenne and Jean-Marie Collot d'Herbois, near-term fear by others (e.g., Joseph Fouché, Jean-Lambert Tallien who spoke out and coordinated on or immediately after the speech of 9 Thermidor) been concentrated rather than uniform, and a fundamental factor in Robespierre's downfall.

our families and our friends . . . We ended by becoming so accustomed to these inextricable situations, that we pursued our daily task, so that business should not be left pending, as though we had a whole lifetime before us, when it was perfectly likely that we would not see the sun rise tomorrow.

While providing direct evidence of the leader's strategic use of uncertainty and the effect of uncertainty on collective action is difficult, we suspect both are present in the authoritarian regimes we highlight. In what follows, we discuss three episodes in more detail. The 1937–1938 purge of Red Army officers and Saddam Hussein's 1979 Ba'ath Party purge illustrate unpredictability over the identity of future targets. Zimbabwe's 2017 military coup provides a contrast: targeting became predictable at the level of an identifiable group, helping potential future targets coordinate.

We make an important caveat before proceeding. When interpreting these episodes, we must distinguish at least two *sources* of uncertainty: (a) the desired number of targets, and (b) *who* will be eliminated (our focus). We chose the first two cases because, while the desired number of targets was almost surely uncertain in each, the identity of who would be eliminated appears highly uncertain as well.[20]

1937-38 Purges of Officers in the Red Army. Between 1937 and 1938, Stalin purged thousands of officers from the Red Army, including many high ranking officers. While scholars have attributed substantial discretion over the choice of targets to Stalin (and the People's Commissariat for Internal Affairs, NKVD) directly, the archives which may host internal documents revealing the precise nature and scope of Stalin's discretion are unavailable to the public. From the recent analysis in Zakharov and Sonin (2024), however, we learn about the characteristics of those among the 1,844 general-grade army officers who were targeted for arrest and execution (at least 780 of whom were executed) and who, as noted in Zakharov and Sonin (2024) likely presented a substantial risk of collective action and resistance to the purge. As the authors show, being purged is not substantially correlated with one's superior or subordinate being purged,

[20] Another type of purge for which our model yields limited insight is the simultaneous execution of opponents who are caught completely by surprise, such as Hitler's Night of Long Knives in 1934. We view such events akin to a "preemptive surprise attack" in war.

suggesting that it would have been difficult to predict future purge targets based on past purge targets. The authors propose a network based theory to differentiate *informed* purges—the leader knows at least one coup plotter's identity—from *preventive* purge—the leader does not know the identity of any coup plotters, and purges to prevent the formation of a coup plot—and argue that the evidence is more consistent with preventive as opposed to informed purges.

The theory we develop in this paper produces a rationale for why purges whose future targets are difficult to predict, as in a preventive purge in the language of Zakharov and Sonin (2024), more successfully undermine collective action and resistance than those for which future targets are easy to predict, as in an informed purge. It also suggests that the weak empirical predictability of purge targets documented by Zakharov and Sonin (2024) among high-ranking army officers could have been the result of a *deliberate* attempt to maintain uncertainty about future targets, thereby reducing the likelihood of collective resistance. While direct evidence on intent is scarce, some qualitative observations are consistent with this interpretation.

First, there is evidence that high-ranking regime insiders often understood that participation in repression did not guarantee, or even make likely, future safety. As Corps Commander Ivan Belov, a member of the Military Collegium and one of the judges presiding over the 1937 trial of Marshal Tukhachevsky, remarked, "*Tomorrow I'll be put in the same place.*" (Montefiore, 2003, p. 230). Second, this fear was not misplaced, as Belov was arrested in January 1938 and executed in July 1938. Several other members of the tribunal, including marshals Vasily Blyukher and Alexander Yegorov and corps commanders Yakov Alksnis, Pavel Dybenko, and Nikolai Kashirin, were also arrested and executed in subsequent purge waves (Montefiore, 2003, Conquest, 2008).[21] Third, despite this fear and despite holding substantial coercive power, these high-ranking insiders did not mount any effective resistance or coup attempt prior to their arrests and execution.[22]

Consistent with our model, this suggests that Stalin successfully maintained insiders' perceived risk of becoming future targets sufficiently small relative to the risks associated with op-

[21] Of the high-ranked officials involved in the trial only a few, such as Marshal Semyon Budyonny and Army Commander Boris Shaposhnikov, survived (Montefiore, 2003, Conquest, 2008).

[22] Following Stalin's death, Soviet authorities formally overturned many purge-era convictions and posthumously rehabilitated numerous senior military officers who had been accused of participating in a military conspiracy, including Tukhachevsky, Blyukher, Yegorov, Belov and several other senior officers involved in the Tukhachevsky trial. These rehabilitation decisions established that the charges against them had been unfounded.

posing the regime, making continued compliance individually optimal for those not yet targeted.

Saddam Hussein's Ba'athist Purge 1979. In this section, we detail Saddam Hussein's 1979 Ba'ath party purge. In doing this, our goal is to simply establish the existence of a nearly simultaneous purge in which the leader chooses to maintain uncertainty over future purge targets, even within the highly condensed sequence of events that unfolded over a single day. This illustrates one of the core premises of our model: that uncertainty over future purge targets undermines collective action, and a leader would prefer to maintain uncertainty rather than to make the ranking of potential purge targets known ex-ante.

In 1979, Saddam Hussein, having newly ascended to the Iraqi presidency, executed a televised purge of dozens of members of the Ba'ath party leadership in Baghdad's al-Khuld conference center. As described in Matthews (2018),

> *Muhie Mashhadi, visibly broken from torture by the secret police, was pushed onstage by his security handlers and forced to read a "confession" that implicated himself and other members of the RCC in a Syrian plot to take over the Iraqi Ba'ath party. Mashhadi then proceeded to slowly read a prepared list of co-conspirators, many of whom were assembled in the hall. As names were called, each of those named were confronted by men in suits and pulled from the conference hall by Hussein's security forces.*

While it is impossible to know whether those eventually called out by Mashhadi could have mounted any resistance if they were all informed simultaneously of their supposed guilt, nearly a third of the present members of the Ba'ath party were arrested (over 60 members), 21 of whom were then executed. By removing the "guilty" one-by-one, Hussein ensured that no large group who *knew* they would be named could coordinate resistance. While it is certainly true that some would have predicted a higher probability of their own "guilt" than others, the structure of the purge was such that no one was certain until their name was called, at which point they were immediately removed from the room by security forces.

Zimbabwe's 2017 political purge and military coup. Zimbabwe provides a useful contrast: as targeting became more predictable, potential future victims could identify one another and

coordinate. The following account of the episode draws throughout on Tendi (2020). In December 2014, Robert Mugabe, then the President of Zimbabwe, purged Vice President Joice Mujuru and several of her supporters from the ruling ZANU PF party. He subsequently marginalized liberation-era officials while elevating younger politicians commonly labelled Generation 40 (G40). By July 2017, he had publicly announced plans to "retire" members of the military leadership; senior officers also feared arrest and prosecution. Because the risk of being purged tracked observable characteristics—military rank and association with the liberation generation—the threatened group was increasingly identifiable (Tendi, 2020, pp. 51–56). The process became more concrete on November 6, 2017, when Mugabe dismissed Vice President Emmerson Mnangagwa, a prominent representative of the liberation-era faction within ZANU PF. Lieutenant General Philip Sibanda, then acting commander of the Zimbabwe Defence Forces, alerted Defence Forces commander Constantino Chiwenga and arranged military assistance for Mnangagwa to leave the country. Although Sibanda and Chiwenga were not close and had served in rival liberation armies, they cooperated. On November 13, after Mugabe refused to meet with the generals, Chiwenga appeared alongside dozens of senior army and air-force officers. He described the ongoing "purging and cleansing process" as "targeting mostly members associated with our liberation history" and warned that the military would intervene. Later that day, according to one participant, Chiwenga told officers commanding key military formations: "If anyone tries to fire or arrest one of us, that is an attack on us all." The officers supported intervention. The following night, the military launched Operation Restore Legacy, placed Mugabe under house arrest, and ultimately ended his thirty-seven-year rule.

In the language of our model, Mnangagwa's dismissal identified a current target, while Mugabe's public threat against the military leadership and his visible marginalization of liberation-era officials identified a relatively narrow set of likely future targets. The precise order of future removals remained unknown, but exposure was sufficiently concentrated for threatened insiders to recognize one another and treat an attack on any member as a threat to the group. The case therefore lies between our two polar benchmarks but is considerably closer to predictable-order elimination: predictability arose not from an announced list, but because targeting followed observable political and institutional characteristics. Although target predictability did not alone cause the coup, the episode provides unusually direct evidence for our mechanism: potential fu-

ture victims recognized their common exposure and coordinated before they could be removed separately.[23]

6 Target priorities

In the baseline model, the principal's payoff depends only on the number of surviving agents. Consequently, conditional on eliminating a given number of agents, the principal is indifferent over their identities. In this section, we allow agents to differ in the costs they impose on the principal. We show that the predictable-order and unpredictable-order targeting environments studied above arise naturally depending on whether agents observe these costs.

Environment. For any ideal population size $N^* \geq 0$, normalize u so that $u(n) - u(n-1) \geq 1$ for all $n \in \{1, \ldots, N^*\}$. Assume that, when $\Omega_t = 0$, the principal's per-period payoff is

$$u(N_t) - \sum_{i \in \mathcal{N}_t} v_i,$$

where $(v_i)_{i \in \mathcal{N}_0}$ are i.i.d. draws from a continuous distribution F with full support on $[0, 1]$.[24] Agents with larger v_i are more costly to retain. When $\Omega_t = 1$, the principal's per-period payoff remains zero. The following condition is a convenient analogue of Assumption 2:

Assumption 3. *The principal prefers forgoing elimination to choosing a target set that induces a protest of size $p_t \geq \overline{k}$:*

$$(1 - \rho)u(N^*) < u(N_0) - N_0.$$

Indeed, retaining all agents gives the principal at least $u(N_0) - N_0$ in each future period, whereas choosing a target set that induces such a protest gives her an expected payoff of at most $(1 - \rho)u(N^*)$ in each future period. Assumption 3 therefore imposes the stated preference uniformly over realizations of $(v_i)_{i \in \mathcal{N}_0}$. All other elements of the environment are unchanged.

[23] The episode concerns political elimination through dismissal, arrest, or prosecution rather than certain execution. Tendi (2020) also emphasizes the generals' political ambitions, Mugabe's declining authority, attachment to the liberation struggle, and fear of prosecution. He identifies Mugabe's refusal to meet with the generals on November 13 as the immediate catalyst.

[24] Together with the normalization of u, the upper bound $v_i \leq 1$ ensures that, absent protest constraints, the principal almost surely wishes to eliminate every excess agent but not to reduce the population below N^*.

6.1 Known priorities

Consider first the case in which the realization $(v_i)_{i\in\mathcal{N}_0}$ is common knowledge. Because F is continuous, ties occur with probability zero. Let $\succ_v$ denote the strict total order induced by the realized values (v_i), so that

$$i \succ_v j \qquad \text{if and only if} \qquad v_i > v_j.$$

When the principal's priorities are common knowledge, equilibrium play reduces to predictable-order elimination with order $\succ_v$. The reason is simple. Holding fixed the number of targets, the principal strictly prefers to target agents with the highest v_i. When $N^* < N_t < N^* + \overline{k}$, she can eliminate all remaining excess agents without triggering a potentially successful protest and therefore eliminates those with the highest v_i. Proceeding by backward induction as in the predictable-order elimination case, this implies that, starting from any initial population with $N_0 > N^*$, the principal eliminates $E_0 \bmod \overline{k}$ agents in the first period and then stops. By the same argument, these are the agents with the highest v_i. Thus under common knowledge of priorities, coalition-proof SPE are predictable-order coalition-proof SPE.

Proposition 6. *Suppose the realization $(v_i)_{i\in\mathcal{N}_0}$ is common knowledge. Then the unique equilibrium path of elimination in coalition-proof SPE coincides with the unique equilibrium path of eliminations under predictable-order coalition-proof SPE with order $\succ_v$.*

6.2 Unknown priorities

We now consider the case in which the principal observes $(v_i)_{i\in\mathcal{N}_0}$ but agents do not. Agents share a common prior under which (v_i) is drawn i.i.d. from F. Because F is continuous, the realizations are almost surely distinct and therefore induce a strict ranking of agents. In particular, for every subset $\mathcal{N}_t \subseteq \mathcal{N}_0$, the surviving agents can be uniquely ordered according to their v_i values. Agents observe the history of targeting decisions and protests but not the realizations (v_i).

This defines a dynamic game with incomplete information. We focus on a refined class of perfect Bayesian equilibria, which we denote by rPBE.

Definition 3. *A* refined perfect Bayesian equilibrium *(rPBE) is a perfect Bayesian equilibrium satisfying the following three additional conditions.*

1. ***Coalition-proofness.*** *At any history with proposal ℓ^t, no coalition $C \subseteq \mathcal{N}_t$ of agents has a joint one-shot deviation that strictly improves every agent's payoff.*

2. ***Exchangeability of untargeted agents.*** *At any history with proposal $\ell^t = (h^t, \mathcal{K}_t)$, on or off the equilibrium path, the posterior distribution over the vector of types $(v_i)_{i \in \mathcal{N}_t \setminus \mathcal{K}_t}$ is invariant to permutations of identities in $\mathcal{N}_t \setminus \mathcal{K}_t$.*

3. ***Tie-breaking.*** *At any history with proposal $\ell^t = (h^t, \mathcal{K}_t)$, if a surviving agent $i \in \mathcal{N}_t$ is indifferent between protesting and not protesting, then i protests with probability zero.*

The first restriction aligns the analysis with the baseline model and allows us to abstract from coordination issues, focusing instead on the role of uncertainty. The second is a reduced-form unpredictability restriction, ruling out differential inferences among agents who receive the same treatment. In particular, conditional on not being targeted, agents do not infer that one untargeted survivor is more exposed than another. Without the exchangeability condition, proposals could be interpreted as revealing arbitrary information about the principal's ranking, vastly expanding the set of behavior that could be sustained in an equilibrium. The third restriction imposes the same non-protest-favoring tie-breaking rule used in the baseline model.

Proposition 7. *Suppose that only the principal observes $(v_i)_{i \in \mathcal{N}_0}$. Then every rPBE induces the same path of target counts and population sizes $(k_t, N_t)_{t \geq 0}$ as the unique unpredictable-order coalition-proof SPE path in the baseline model with the same initial population N_0. Moreover, no protest succeeds on path and, whenever the principal targets k_t agents in period t, she targets the k_t surviving agents with the highest values of v_i.*

Propositions 6 and 7 clarify the interpretation of the two benchmark environments studied above. Predictable-order elimination arises when the principal's priorities are commonly understood, so future targets are predictable. Unpredictable-order elimination arises when the principal has a determinate priority ranking but, conditional on surviving the current period, untargeted agents remain symmetric in their beliefs about future exposure. What matters for the logic of the model is therefore not whether the principal literally randomizes over targets, but whether agents can correctly anticipate future targets.

This interpretation is natural in many applications. In authoritarian environments, insiders may understand that the principal has underlying priorities, grudges, or private information about potential opponents, while remaining uncertain about how those considerations map into future purge targets. From the perspective of any one untargeted agent, the relevant risk is then diffuse: although the principal is not literally randomizing, future targets are sufficiently unpredictable that the same coordination problem arises as under unpredictable-order elimination. Because unpredictable-order coalition-proof SPE typically lead to higher payoff than predictable-order coalition-proof SPE, we should not expect to observe a leader clarifying their ranking.

7 Discussion

History-dependent threats and equilibrium multiplicity. We focus on predictable-order and unpredictable-order coalition-proof SPE. Outside these classes, a broader set of equilibrium paths and payoffs can be sustained through history-dependent threats. For instance, a coalition-proof SPE in which the principal targets no agent for x periods and thereafter follows the unique unpredictable-order coalition-proof SPE path can be sustained by the threat that any acceleration in targeting triggers a switch to the predictable-order coalition-proof SPE associated with some arbitrary ordering. Indeed, if N_0 is sufficiently large relative to x and δ is sufficiently close to one, the loss from triggering the predictable-order continuation outweighs the immediate benefit from accelerating elimination. Nevertheless, restricting attention to the fixed- and unpredictable-order classes is not simply a tractability device. As shown in Section 6, these two environments arise endogenously, depending on whether agents observe the heterogeneous costs they impose on the principal, so the elimination order is pinned down by the information structure rather than wielded as a history-dependent threat.

Elimination technology. In the baseline model, targeted agents can resist by joining the protest before eliminations are carried out. Combined with Assumption 2, this imposes a clear upper bound of $\overline{k}-1$ on the number of agents the principal can optimally target in any period. In some applications, however, targeted agents may be eliminated before they can react (e.g., assassinations). The timing would then be reversed: the principal first eliminates her chosen targets,

after which the remaining agents decide whether to protest. If the principal could eliminate an arbitrary number of agents, the problem would then become trivial: as in our baseline setting with $\overline{k} > N_0$, the principal would eliminate all agents she wishes, reducing the population size to $\min\{N_0, N^*\}$ in one period. A more realistic alternative is that the principal faces a per-period targeting capacity, $\bar{m} \in \mathbb{N}$, reflecting constraints on personnel, intelligence, transportation, detention, or execution capacity. Under such a capacity constraint, a substantively equivalent version of our results apply, except replacing the maximum elimination size beneath the protest constraint $\overline{k} - 1$ with the technological constraint $\bar{m}$. The main qualitative mechanism remains: predictable-order targeting allows future victims to coordinate before their turn, whereas unpredictability diffuses perceived risk and weakens resistance, although the precise blocking thresholds and protest constraints must be recomputed because current targets no longer participate in the protest.

8 Conclusion

In this paper we have studied a game between a principal and finite group of agents who she seeks to eliminate. We have constructed two classes of equilibria: predictable-order coalition proof SPE and unpredictable-order coalition proof SPE. In the former, agents are eliminated according to a known order, while in the latter, agents are eliminated according to uniform random draws each period. We showed that a unique path of eliminations exists in each case, with the principal typically preferring unpredictable-order elimination. In the unique unpredictable-order coalition-proof SPE path, the principal adopts a greedy elimination strategy before stopping at an endogenously determined threshold. We connected our results to historical purge episodes, arguing that uncertainty over future purge targets was pervasive in these episodes and served to undermine collective resistance.

A number of avenues for future work could prove fruitful. First, our paper has focused on a "non-steady state" environment, in which the principal starts with a population size that is too large, and eventually reaches a population size from which no further eliminations occur. In many authoritarian regimes however, purges recur. One way to study recurrent purges is to introduce replenishment: some number of agents are born each period. In authoritarian politics, leaders

often replace powerful ministers with weaker ones. Over time however, weaker ministers become more powerful, posing a coup risk to the leader. As such, the leader may eventually attempt to purge the replacement minister, in which case the type of uncertainty we have studied in this paper could prove useful. We leave to future work the study of such dynamics.

Second, our protest technology is simple—no success below a threshold and fixed probability of success above a threshold—which allows for tight characterizations but may lead us to ignore some important features of the environments we study. Numerical simulations show that more complex protest technology leads to significantly more complex equilibrium behavior, and future work studying these dynamics may yield new insight.

More broadly, novel insights into coalition-formation could emerge from combining our notion of unpredictable-order elimination with the rich heterogeneity present in Acemoglu et al. (2008) and follow-on work. Finally, while substantial empirical work on elite purges exists (see our use of Zakharov and Sonin 2024), there are many historical purge episodes in which targets faced uncertainty over their future inclusion in the purge. Assembling the data necessary to quantify this uncertainty is an exciting avenue for future work.

References

Acemoglu, D., G. Egorov, and K. Sonin (2008): "Coalition formation in non-democracies," *The Review of Economic Studies*, 75, 987–1009.

——— (2012): "Dynamics and stability of constitutions, coalitions, and clubs," *American Economic Review*, 102, 1446–1476.

Ali, S. N., B. D. Bernheim, A. W. Bloedel, and S. C. Battilana (2023): "Who controls the agenda controls the legislature," *American Economic Review*, 113, 3090–3128.

Ali, S. N., B. D. Bernheim, and X. Fan (2019): "Predictability and power in legislative bargaining," *The Review of Economic Studies*, 86, 500–525.

Ali, S. N. and C. Liu (2025): "Coalitions in repeated games," *Working Paper*.

Bernheim, B. D., A. Rangel, and L. Rayo (2006): "The power of the last word in legislative policy making," *Econometrica*, 74, 1161–1190.

Che, Y.-K., J. Huang, and W. Lim (2026): "First They Came for the Others: A Theory of Divide-and-Conquer," Tech. rep.

Conquest, R. (2008): *The great terror: A reassessment*, Oxford University Press.

De Mesquita, B. B., A. Smith, R. M. Siverson, and J. D. Morrow (2005): *The logic of political survival*, MIT press.

Diermeier, D., G. Egorov, and K. Sonin (2017): "Political economy of redistribution," *Econometrica*, 85, 851–870.

Diermeier, D. and P. Fong (2009): "Endogenous limits on proposal power," Tech. rep., Discussion Paper.

——— (2011): "Legislative bargaining with reconsideration," *The Quarterly Journal of Economics*, 126, 947–985.

Gieczewski, G. and K. Kocak (2024): "Collective procrastination and protest cycles," *American Journal of Political Science*.

Gregory, P. R. (2009): *Terror by quota: state security from Lenin to Stalin:(an archival study)*, Yale University Press.

Haskin, J. M. (2005): *The tragic state of the Congo: From decolonization to dictatorship*, Algora Publishing.

Iaryczower, M. and S. Oliveros (2023): "Collective hold-up," *Theoretical Economics*, 18, 1063–1100.

Kalandrakis, A. (2004): "A three-player dynamic majoritarian bargaining game," *Journal of Economic Theory*, 116, 294–14.

Linton, M. (2013): *Choosing terror: Virtue, friendship, and authenticity in the French revolution*, OUP Oxford.

Matthews, A. S. (2018): *Conflict among Comrades: Elite Purges and Political Violence in Authoritarian Regimes*, Baton Rouge, LA: Louisiana State University.

Montagnes, B. P. and S. Wolton (2019): "Mass purges: Top-down accountability in autocracy," *American Political Science Review*, 113, 1045–1059.

Montefiore, S. S. (2003): *Stalin: The Court of the Red Tsar*, London: Weidenfeld & Nicolson.

Shadmehr, M. and D. Bernhardt (2019): "Vanguards in revolution," *Games and Economic Behavior*, 115, 146–166.

Tendi, B.-M. (2020): "The Motivations and Dynamics of Zimbabwe's 2017 Military Coup," *African Affairs*, 119, 39–67.

Tyson, S. A. (2018): "The agency problem underlying repression," *The Journal of Politics*, 80, 1297–1310.

Zakharov, A. and K. Sonin (2024): "The Anatomy of the Great Terror: A Quantitative Analysis of the 1937-38 Purges in the Red Army," *University of Chicago, Becker Friedman Institute for Economics Working Paper*.

Zeilberger, T. (2024): "The Logic of Political Survival Revisited: Consequences of Elite Uncertainty Under Authoritarian Rule," *arXiv preprint arXiv:2408.01887*.

A Appendix

A.1 Further discussion of literature

Connection to Ali et al. (2023): To see more clearly the connection between our model and Ali et al. (2023), consider the following model, which is a variant of our model made to be a special case of Ali et al. (2023). There is an agenda-setter and 3 agents. There are 2 periods of proposals that the agenda-setter makes, with period indexed $t \in \{0, 1\}$, and payoffs only accrue after the last period of proposals. The policy proposed by the principal is the set of agents to be eliminated. Each agent's payoff is $\mathbf{1}\{\text{NOT ELIMINATED}\} + \epsilon \times \text{NUMBER OF OTHER ELIMINATED AGENTS}$ and the

principal's payoff is NUMBER OF ELIMINATED AGENTS$+\epsilon\times$SUM OF INDEXES OF ELIMINATED AGENTS, for very small ϵ. Each period, there is a status quo set of eliminated agents (with eliminating no agents as the period 0 status quo), and the agenda-setter can propose any revision to the set of eliminated agents, and if at least 2 agents approve (not including the agenda-setter) then the proposal passes, and otherwise it fails.

First, consider the case in which a failed proposal does not stop the agenda-setter from future proposals, as in the setting of Ali et al. (2023). Applying the logic from Ali et al. (2023), the agenda-setter will propose the favorite improvement operator each period to eliminate 2 out of 3 agents.

Suppose instead that a failed proposal stops the principal from making any future proposals. Then, since the status quo in period 0 is to eliminate no agents, the agenda-setter is only able to eliminate agent $i = 3$, which he passes with a proposal in the final period. Why? First, observe that if no past proposal has failed, then in the final period, the principal will successfully propose to eliminate the highest indexed agent who would not be eliminated according to the status quo. Then, suppose that in period 0 the agenda-setter proposes to eliminate at least one agent. Then all of the agents who are proposed to be eliminated, as well as the highest indexed non-eliminated agent, would vote against the proposal, since in the event that the proposal fails they will get payoff at least 1, while in the event that the proposal succeeds they will get payoff no more than 2ϵ, which is assumed very small. As a result, the agenda-setter is only able to eliminate agent $i = 3$ in the last period.

Connection to mixed strategy equilibria in bargaining As mentioned in the literature review of Section 2, the randomization resulting from mixed strategy equilibria in bargaining is qualitatively different than the randomization we study. The core idea in a 3 agent, split-the-dollar, version of the model is as follows (from Diermeier and Fong (2009)). Suppose status-quo for the agenda-setter is 0, and for the others is $\frac{1}{2}$ each. Then consider the following behavior. The agenda-setter randomly selects one agent, and proposes an adjustment that gives the entire 1 to that agent; the agenda-setter and the selected agent vote in favor. In period 2, the agenda-setter then proposes 1 for herself, the agenda-setter and agent whose status-quo payoff in period 2 is 0 vote in favor. It is straightforward to argue that once the status quo payoff for one of the

non-agenda-setter agents has reached 0, the agenda-setter can redistribute the entire 1 to herself. Taking that as given, the path can be sustained as an equilibrium path because if the contemporaneously selected agent rejects, she anticipates that the other agent will eventually be selected, in which case that agent will get the entire 1 for a single period, and then both will get 0 everafter.

A.2 Proof of Proposition 1

Proof. Recall that $E_t := \max\{0, N_t - N^*\}$ and $r(E) := E \bmod \overline{k}$. Conditional on a target count m, a predictable-order equilibrium selects the m highest-ranked survivors, although deviations are unrestricted. It therefore suffices to establish the optimal target count.

We prove by strong induction on E that, at every history with $\Omega_t = 0$ and $E_t = E$, the principal targets $r(E) < \overline{k}$ agents and then stops permanently. The proposal induces no potentially successful protest.

Base step ($E < \overline{k}$). If $E = 0$, the principal targets $r(E) = 0$ agents: any elimination would move the population farther from N^* and lower her continuation payoff. If instead $1 \leq E < \overline{k}$, targeting all $E = r(E)$ excess agents leaves N^* survivors, after which no further elimination occurs. Anticipating this, untargeted agents are unwilling to bring participation to $\overline{k}$, while the $E < \overline{k}$ targets cannot reach the threshold alone. Thus, the proposal induces no potentially successful protest. Targeting more either results in the principal's removal or reduces the population below N^*, while targeting fewer leaves more than N^* survivors next period, yields a lower next-period flow, and cannot outperform the maximal flow $u(N^*)$ thereafter. Hence, the principal targets all $E = r(E)$ excess agents immediately.

Inductive step. Write $E = q\overline{k} + r$, where $q \geq 1$ and $r := r(E)$, and assume the claim for every smaller value of E. Any target count $m \geq \overline{k}$ induces a potentially successful protest by the targets alone. Now consider $r < m < \overline{k}$. If the induced protest were subthreshold, only the m targets would be eliminated, so the next-period excess population would satisfy

$$E - m = (q-1)\overline{k} + (\overline{k} + r - m) < E, \qquad r(E-m) = \overline{k} + r - m \geq \overline{k} - m.$$

By the induction hypothesis, that many highest-ranked untargeted survivors would be targeted next period. Since $r(E-m) \geq \overline{k} - m$, enough of them can join the m current targets to make the

protest potentially successful. Assumption 1 makes this joint deviation strictly profitable, contradicting coalition-proofness. Hence, every proposal with $m > r$ induces a potentially successful protest and is suboptimal by Assumption 2.

If $r = 0$, the principal therefore stops permanently. If instead $r > 0$, targeting exactly r agents leaves the smaller excess population $E - r = q\overline{k}$, whose remainder is zero. By the induction hypothesis, no further elimination occurs. Untargeted agents are therefore unwilling to join a potentially successful protest, while the $r < \overline{k}$ targets cannot mount one alone. The proposal is thus implemented. Any proposal with $m \in \{1, \dots, r-1\}$ targets is either blocked or, by the induction hypothesis, implies that the principal targets another $r - m$ agents next period and then stops permanently. The latter path reaches the same terminal population as targeting all r agents immediately, but one period later. Because u is strictly decreasing above N^*, targeting all r agents immediately is strictly better. If the principal targets no agent, the state is unchanged and the same upper bound applies next period; inaction therefore either delays the reduction by r or forgoes it, and is strictly worse. Hence, the principal targets exactly $r(E)$ agents and then stops permanently. By predictable-order targeting, these are the $r(E)$ highest-ranked survivors. Applying the claim at E_0 proves the proposition. □

Proof of Propositions 2 and 3

We prove Propositions 2 and 3 together.

We can proceed by backward induction to uniquely determine the number of targeted agents and the value functions for the government and for the agents when the surviving population size is $n > 0$, before the government's targeting choice.

Lemma 1 (Characterization under unpredictable-order)**.** *Any unpredictable-order coalition-proof equilibrium induces the same mappings from the current population size $n \in \mathbb{N}$ to (i) the number of targeted agents $\kappa(n) \in \{0, \dots, n\}$, and (ii) the value functions for the government $V^{gov}(n)$ and for the agents $V(n)$, before the realization of who is targeted. Moreover, $V^{gov}(n)$ is strictly decreasing*

in n on $\{n \in \mathbb{N} : n \geq N^*\}$, and the following recursive relationships hold:[25]

$$\kappa(n) = \max\left\{\{0\} \cup \left\{k \in \{1, \ldots, \overline{k}-1\} : 1 + \delta V(n-k) \geq 1 + \frac{\rho\delta}{1-\delta} \text{ and } n-k \geq N^*\right\}\right\}.$$

$$V^{gov}(n) = u(n) + \delta V^{gov}(n - \kappa(n)),$$

$$V(n) = 1 + \delta V(n - \kappa(n))\frac{n - \kappa(n)}{n}.$$

Furthermore, $\kappa(n)$, $V^{gov}(n)$, and $V(n)$ are all independent of N_0, the initial population size at $t = 0$.

Proof. First note that if $n \leq N^*$ the government optimally sets $k(n) = 0$.

Second, suppose $n > N^*$ and denote by $m = n - N^*$ the number of excess agents the government would like to eliminate. The government can eliminate all remaining excess agents in one period whenever $m < \overline{k}$. Hence, for all $n \in \{N^*, ..., N^* + \overline{k} - 1\}$, $\kappa(n) = m$, $V(n) = 1 + \delta\frac{N^*}{N^*+m}(\frac{1}{1-\delta})$, and $V^{gov}(n) = u(n) + \frac{\delta u(N^*)}{1-\delta}$, strictly decreasing in $n \in \{N^*, ..., N^* + \overline{k} - 1\}$.

Now suppose $\overline{n} \geq N^* + \overline{k} - 1$ and that $\kappa(n)$, $V^{gov}(n)$, and $V(n)$ are uniquely determined for all $n \leq \overline{n} \in \mathbb{N}$, with $V^{gov}(n)$ strictly decreasing over $n \in \{N^*, ..., \bar{n}\}$. We show that $\kappa(n)$, $V^{gov}(n)$, and $V(n)$ are also uniquely determined at $n = \overline{n} + 1$, and that $V^{gov}(\overline{n} + 1) < V^{gov}(\overline{n})$.

By Assumption 2, the government strictly prefers setting $k = 0$ (thereby avoiding any protest) to choosing any $k > 0$ that would trigger a protest that (weakly) exceeds size $\overline{k}$. Thus, for $n \geq N^* + \overline{k}$, the government chooses $k > 0$ only if the untargeted agents do not protest, i.e.

$$1 + \delta V(\overline{n} + 1 - k) \geq 1 + \frac{\rho\delta}{1-\delta}. \tag{2}$$

To establish that the government chooses $k > 0$ when it is feasible to do so without triggering a size $\overline{k}$ protest, suppose the government targets no agents for $l > 0$ periods, after which it chooses $k \in \{1, ..., \overline{k}\}$. Then, the government's payoff would be

$$\begin{aligned}
&u(\overline{n}+1)\sum_{i=0}^{l}\delta^i + \delta^{l+1}V^{gov}(\overline{n}+1-k) \\
<&u(\overline{n}+1) + u(\overline{n}+1-k)\sum_{i=1}^{l}\delta^i + \delta^{l+1}V^{gov}(\overline{n}+1-k). \\
\leq&u(\overline{n}+1) + \delta V^{gov}(\overline{n}+1-k)
\end{aligned}$$

[25] The principal never optimally targets $k_t \geq \overline{k}$ agents in a single period. Otherwise, the targeted agents alone can (and will) form a protest of size at least $\overline{k}$; by Assumption 3, such a proposal is strictly suboptimal.

where the last expression captures the payoff the government would get by targeting k today. As a result, the government chooses $k > 0$ if it is feasible, and otherwise chooses $k = 0$.

The government's problem at $\overline{n} + 1 \geq N^* + \overline{k}$ can be written as,

$$\max_{k \in S(\overline{n}+1)} V^{gov}(\overline{n} + 1 - k),$$

where

$$S(\overline{n} + 1) = \left\{k \in \{1, \ldots, \overline{k} - 1\} : 1 + \delta\, V(\overline{n} + 1 - k) \geq 1 + \tfrac{\rho\delta}{1-\delta}\right\}$$

if $S(\bar{n} + 1) \neq \emptyset$ and otherwise the government sets $k = 0$ and $V^{gov}(\overline{n} + 1) = u(\overline{n} + 1)\frac{1}{1-\delta}$

Because $V^{gov}(n)$ is strictly decreasing over $\{N^*, ..., \overline{n}\}$ this implies that the principal sets

$$\kappa(\overline{n} + 1) = \max\left\{\{0\} \cup \left\{S(\overline{n} + 1)\right\}\right\}$$
$$V^{gov}(\overline{n} + 1) = u(\overline{n} + 1) + \delta V^{gov}(\overline{n} + 1 - \kappa(\overline{n} + 1)),$$

and, by anonymity,

$$V(\overline{n} + 1) = 1 + \delta V(\overline{n} + 1 - \kappa(\overline{n} + 1))\frac{\overline{n} + 1 - \kappa(\overline{n} + 1)}{\overline{n} + 1}.$$

If $\kappa(\overline{n} + 1) = 0$, then $V^{gov}(\overline{n} + 1) = \frac{u(\overline{n}+1)}{1-\delta} < \frac{u(\overline{n})}{1-\delta} \leq V^{gov}(\overline{n})$. If $\kappa(\overline{n} + 1) \geq 1$,

$$\begin{aligned} V^{gov}(\overline{n} + 1) &= u(\overline{n} + 1) + \delta V^{gov}(\overline{n} + 1 - \kappa(\overline{n} + 1)) \\ &< u(\overline{n}) + \delta V^{gov}(\overline{n} + 1 - \kappa(\overline{n} + 1)) \\ &\leq V^{gov}(\overline{n}) \end{aligned}$$

where the last inequality follows from the fact that if the government can target $\kappa(\overline{n} + 1)$ agents starting from $\bar{n} + 1$ without triggering protests by untargeted agents, then it can also target $\kappa(\overline{n} + 1) - 1$ agents starting from $\bar{n}$ without triggering protests by untargeted agents. This is because the decision of untargeted agents to protest or not depends only on the number of agents left after those targeted are eliminated, not on the starting point (see (2)).

In either case, $V^{gov}(\overline{n} + 1) < V^{gov}(\overline{n})$, completing the proof. □

Proof of Proposition 2 and 3. Consider any initial population size $N_0 \in \mathbb{N}$. If $N_0 \leq N^*$, the proof is complete, since the principal has reached her ideal point among feasible population sizes, since

the population cannot grow. As a result, suppose $N_0 > N^*$ for the remainder of the proof. Denote by $\mathcal{X}$ the set of integers x such that, in any coalition-proof subgame perfect equilibrium, an active government does not eliminate any further agents once the number of survivors equals x, i.e., such that $\kappa(x) = 0$.[26] Since the government never eliminates any agent when the current population size is $n \leq N^*$, every $n \in \{0, ..., N^*\}$ belongs to $\mathcal{X}$ and no untargeted agent would ever oppose the elmination of k agents from a starting population of size n when $n - k \leq N^*$.[27] Hence, by Lemma 1, $\kappa(n) = n - N^*$ agents are targeted and eliminated with probability 1 whenever $n \in \{N^*, ..., N^* + \overline{k} - 1\}$.

Moreover, in any coalition-proof equilibrium, population sizes in $\mathcal{X}$ are terminal: starting from any initial population $N_0 \geq 0$, the process reaches the final population size

$$x_{N_0} \equiv \max\{x \in \mathcal{X} : x \leq N_0\}.$$

For every $N_0 \in \{0, ..., N^* + \overline{k} - 1\}$ the result is trivial and x_{N_0} is reached in at most one period. The result, however holds also for $N_0 > \overline{k} - 1$. First, by definition, for any population size $n \in \{0, ..., N_0\}$ such that $n \notin \mathcal{X}$, the government would target $\kappa(n) \geq 1$ agents without triggering a size $\overline{k}$ protest.[28] Second, the government does not try to reduce the surviving population below $x \in \mathcal{X}$ from any $n \geq x$. To see why, observe that if $n - x \geq \overline{k}$, the government does not reduce the population below x in a single step: doing so would require targeting at least $n - x + 1 > \overline{k}$ agents, which would trigger a protest of size at least $\overline{k}$ (since $\overline{k}$ targeted agents would protest) which, by Assumption 2, the government strictly prefers to avoid. Thus, the only n from which it might be optimal to try to reach $< x$ agents is for n s.t. $n - x < \overline{k}$. But note that for any $x > N^*$, $\kappa(x) = 0$ implies that targeting any number $k > 0$ of agents would trigger a size $\overline{k}$ protest starting from population size x, and therefore $1 + \delta V(s) < 1 + \frac{\rho\delta}{1-\delta}$ for all $s \in \{x - \underline{k}, \ldots, x - 1\}$, where $\underline{k} \equiv \overline{k} - 1$.[29] This, in turn, ensures $\kappa(n) \leq n - x$ for all $n \in \{x + 1, ..., x + \underline{k}\}$, since the government refrains from triggering a protest of size at least $\overline{k}$ by Assumption 2. Hence, starting

[26] Note this is well defined by Lemma 1.

[27] Note that under Assumption 1, if $N^* = 0$, then the smallest strictly positive element of $\mathcal{X}$ would be $\overline{k}$. At $N_t = \overline{k}$, agents would protest the proposed elimination of any agent, which by Assumption 2 is undesirable for the principal. On the other hand, if $N_t < \overline{k}$, the principal could eliminate all the agents with no risk of protest.

[28] By Assumption 2, triggering such a protest would be suboptimal, contradicting $\kappa(n) \geq 1$.

[29] Recall that in our definition of coalition-proof equilibrium untargeted agents would not protest when indifferent.

from any initial $N_0 \in \mathbb{N}$, the principal eliminates agents until the population size reaches x_{N_0} (within at most $N_0 - x_{N_0}$ periods) and then eliminates no more after reaching x_{N_0}.

If the government targets $k < \overline{k}$ agents when the population size is $n \in \{0, ..., N_0\}$, then untargeted agents protest in a coalition proof equilibrium if and only if $1 + \delta V(m + x - k) < 1 + \frac{\rho\delta}{1-\delta}$, where $m = n - x$.

Anticipating this, as we have shown, the government optimally chooses

$$\kappa(x+m) = \max\Big\{\{0\}\cup\big\{k \in \{1, \ldots, \overline{k}-1\} : 1+\delta\, V(x+m-k) \geq 1+\tfrac{\rho\delta}{1-\delta},\ \text{and}\ x+m-k \geq N^*\big\}\Big\}.$$

We will now show that V is quasi-convex.

To proceed, let $V^{\text{aux}}(x, m)$ denote the expected value for an agent who is alive together with $x+m-1$ others, assuming that in each period exactly $\min\{\overline{k}-1, m\}$ agents, randomly selected, are eliminated (not merely targeted), until reaching x agents remaining, after which no more eliminations occur. More precisely $V^{aux}(x, m)$ is defined by $V(x + m) = 1 + \delta V^{aux}(x, m)$, where $V(x + m)$ is the agent's value as defined in Lemma 1 induced by the path just described. Along this elimination path, $\underline{k} \equiv \overline{k} - 1$ agents are eliminated in each of the first $\lfloor m/\underline{k} \rfloor$ periods,[30] followed by the elimination of $m - \lfloor m/\underline{k} \rfloor\, \underline{k} \leq \underline{k}$ agents in period $\lfloor m/\underline{k} \rfloor + 1$ (leading to a population size of x), after which no further elimination occurs. Under this assumption, each agent has probability $\frac{\min\{m,\underline{k}\}}{x+m}$ of being eliminated in the current period (period 1) and probability $\frac{x}{x+m}$ of never being eliminated. Moreover, if $m > \underline{k}$, the agent has probability $\frac{\underline{k}}{x+m}$ of being eliminated in each period $t \in \{1, \ldots, \lfloor m/\underline{k} \rfloor\}$, and probability $\frac{m-\underline{k}\,\lfloor m/\underline{k} \rfloor}{x+m}$ of being eliminated in period $\lfloor m/\underline{k} \rfloor + 1$. As a result,

$$V^{\text{aux}}(x, m) = V^{\text{auxx}}(x, m, \lfloor m/\underline{k} \rfloor),$$

where

$$\begin{aligned} V^{\text{auxx}}(x, m, q) &:= \frac{1}{1-\delta} \times \frac{x}{x+m} + \left(\frac{m - q\,\underline{k}}{x+m}\right) \frac{1-\delta^{\,q}}{1-\delta} + \frac{\underline{k}}{x+m} \sum_{i=1}^{q} \frac{1-\delta^{i-1}}{1-\delta}, \\ &= \frac{1}{(1-\delta)^2} \times \frac{a_q\, m + b_q}{m+x} \end{aligned}$$

with $a_q = (1-\delta)\big(1-\delta^{\,q}\big)$ and $b_q = (1-\delta)x - \underline{k} + \underline{k}\,\delta^{\,q}\big[1 + (1-\delta)q\big]$.

[30] $\lfloor m/\underline{k} \rfloor$ denotes the largest integer not exceeding $m/\underline{k}$.

Additionally, note that, for every $q \in \mathbb{N}$

$$V^{\text{aux}}(x,m) = V^{\text{auxx}}(x,m,q) \quad \text{for all } m \in I_q := \{\, m \in \mathbb{N} :\ q\,\underline{k} \le m \le (q+1)\,\underline{k} \,\}.$$

This holds, by definition, when $q\,\underline{k} \le m < (q+1)\,\underline{k}$, but also at $m = (q+1)\,\underline{k}$ since

$$V^{auxx}(x,(q+1)\underline{k},q) = V^{auxx}(x,(q+1)\underline{k},q+1), \tag{Bdry Matching}$$

as implied by

$$\begin{aligned} V^{auxx}(x,(q+1)\underline{k},q) - V^{auxx}(x,(q+1)\underline{k},q+1) &= \big((a_q - a_{q+1})\,m + (b_q - b_{q+1})\big)\Big|_{m=(q+1)\,\underline{k}} \\ &= (1-\delta)^2\,\delta^q\big(-m + (q+1)\,\underline{k}\big)\Big|_{m=(q+1)\,\underline{k}} = 0. \end{aligned}$$

Monotonicity within I_q. Fix any $q \in \mathbb{N}$. For any $m \in I_q \setminus \{(q+1)\underline{k}\}$, the forward difference satisfies

$$\Delta V^{\text{aux}}(x,m) := V^{\text{aux}}(x,m+1) - V^{\text{aux}}(x,m) = \frac{1}{(1-\delta)^2} \cdot \frac{a_q\,x - b_q}{(m+x)(m+1+x)}.$$

Hence the sign of $\Delta V^{\text{aux}}(x,m)$ is constant on each block I_q and equals the sign of

$$S(q) := a_q\,x - b_q = \underline{k} - \delta^{\,q}\Big[(1-\delta)\,x + \underline{k}\big(1 + (1-\delta)q\big)\Big].$$

Thus $V^{\text{aux}}(x,m)$ is monotone in m within every integer block I_q.

Single crossing and quasi-convexity. For each $q \in \mathbb{N}$, a direct calculation gives

$$S(q+1) - S(q) = \delta^{\,q}(1-\delta)^2\big[x + \underline{k}(1+q)\big] > 0, \tag{Diff-S}$$

so $\{S(q)\}_{q\ge 0}$ is strictly increasing in q. As a result, there exists $q^* \in \{0,1,\ldots,\infty\}$ such that $S(q) < 0$ for $q < q^*$ and $S(q) \ge 0$ for $q \ge q^*$.

Combining this single-crossing property with the fact that $V^{aux}(m,x)$ is monotone on I_q with sign equal to $S(q)$ and equation (Bdry Matching), we find that $V^{\text{aux}}(x,m)$ is (discrete) quasi-convex in m. For every initial population size N_0 denote the minimum attainable $V^{aux}(x,m)$ as

$$\underline{v}^{\text{aux}}_{N_0} \equiv \min_{m\in\{0,1,\ldots,N_0-x\}} V^{\text{aux}}(x,m),$$

and denote by $m^*(N_0,x)$ the minimum minimizer in the problem above. Importantly, note that $m^*(N_0,x) \ge \min\{N_0 - x, \underline{k}\}$ whenever $N_0 \ge N^* + \overline{k}$. We can see this by observing that $S(0) = -(1-\delta)x$, so that $V^{aux}(x,\underline{k}) < V^{aux}(x,s)$ for all $s \in \{0,\ldots,\underline{k}-1\}$.

Mapping back to the original problem Given an initial population size of $N_0 \in \mathbb{N}$, let

$$x = \max\{\, s \in \mathcal{X} :\ s \leq N_0 \,\} \geq \min\{N^*, N_0\}$$

and consider any $m \in \{0, \ldots, N_0 - x\}$.[31] From Lemma 1, any government strategy in a coalition-proof equilibrium is represented by a function $\tilde{\kappa} : \mathbb{N}^2 \to \mathbb{N}$ such that $\tilde{\kappa}(x, m) \equiv \kappa(x + m)$

By the quasi-convexity of V^{aux}, the lower contour set

$$B := \left\{ s \in \{0, ..., N_0 - x\} :\ 1 + \delta\, V^{\text{aux}}(x, s) < \tfrac{\rho}{1-\delta} \right\}$$

is either a consecutive integer interval,

$$B = \{\underline{m},\ \underline{m} + 1,\ \ldots,\ \overline{m}\} \quad \text{for some} \quad 0 \leq \underline{m} \leq \overline{m} \leq N_0 - x$$

or the empty set, i.e., $B = \emptyset$. Then, by Lemma 1 and under our tie-breaking rule in favor of not protest in the definition of coalition-proof equilibrium, we argue now that every coalition-proof equilibrium induces the same, unique $\tilde{\kappa}(x, m)$, with the following properties:[32]

1. If $m \leq \underline{k}$, then $\tilde{\kappa}(x, m) = m$, i.e., the government eliminates all m excess agents, leaving x agents.

2. If $N_0 - x \geq m > \underline{k}$ and $B = \emptyset$, then $\tilde{\kappa}(x, m) = \underline{k}$ and

$$1 + \delta V^{\text{aux}}(x, m) = V(x + m).$$

3. If $N_0 - x \geq m > \underline{k}$ **and** $B = \{\underline{m}, ..., \overline{m}\} \neq \emptyset$ **and** $m \notin \{\underline{m} + \underline{k}, ..., \overline{m} + \underline{k}\} \cap \{0, ..., N_0 - x\}$, then $\tilde{\kappa}(x, m) = \underline{k}$ and

$$1 + \delta V^{\text{aux}}(x, m) \leq V(x + m).$$

4. If $N_0 - x \geq m > \underline{k}$ **and** $B = \{\underline{m}, ..., \overline{m}\} \neq \emptyset$ **and** $m \in \{\underline{m} + \underline{k}, ..., \overline{m} + \underline{k}\} \cap \{0, ..., N_0 - x\}$, then $\tilde{\kappa}(x, m) = m - (\bar{m} + 1)$ and

$$1 + \delta V^{\text{aux}}(x, m) \leq V(x + m).$$

Note, as we show in the proof below, $\tilde{\kappa}(x, m)$ as defined here cannot be negative.

[31] Recall that $\mathcal{X} \subseteq \mathbb{N}$ is the set of integers x such that, in any coalition-proof subgame perfect equilibrium, an active government cannot successfully eliminate any further agents once the number of survivors equals x, i.e., such that $\kappa(x) = 0$

[32] We say that a coalition-proof equilibrium *induces* a $\tilde{\kappa}(x, m)$ if at every history at which $N = x + m$, the principal's strategy is to target $\tilde{\kappa}(x, m)$ agents.

Call this the ***4-point characterization of*** $\tilde{\kappa}$

That $\tilde{\kappa}(x,m)$ induced by any coalition-proof equilibrium must satisfy point 1 is immediate. Furthermore, that $\tilde{\kappa}(x,m)$ induced by any coalition-proof equilibrium must satisfy point 2 follows from the fact that if $B=\emptyset$ and hence $1+\delta\,\underline{v}^{\text{aux}}_{N_0}\geq\frac{\rho}{1-\delta}$, then the fastest approach path induces no protest of size at least $\overline{k}$, and is the principal's most preferred path among those with $\tilde{\kappa}\leq\underline{k}$.

We now argue that any coalition-proof equilibrium induces $\tilde{\kappa}(x,m)$ satisfies points 3 and 4. Recall that

$$B=\left\{s\in\{0,...,N_0-x\}:\ 1+\delta\,V^{\text{aux}}(x,s)<\tfrac{\rho}{1-\delta}\right\}$$

and by discrete quasi-convexity of V^{aux} is a consecutive integer interval (and by the premise of points 3 and 4 that $B\neq\emptyset$),

$$B=\{\underline{m},\,\underline{m}+1,\,\ldots,\,\overline{m}\}\quad\text{for some}\quad 0\leq\underline{m}\leq\overline{m}\leq N_0-x.$$

By point 2, we know that $\tilde{\kappa}(x,m)=\min\{\underline{k},m\}$ and $1+\delta V^{\text{aux}}(x,m)=V(x+m)$ for all $m\in\{0,...,\ldots,\underline{m}-1\}$. As a consequence, $\tilde{\kappa}(x,m)=\min\{\underline{k},m\}$ and $1+\delta V^{\text{aux}}(x,m)=V(x+m)$ also for all $m\in\{\underline{m},\ldots,\underline{m}+\underline{k}-1\}\cap\{0,...,N_0-x\}$.

Hence, for any population size $n>\underline{m}$, government attempts to eliminate k agents would be met with a protest of size at least $\overline{k}$ whenever

$$m-k\in\{\underline{m},\ldots,\min\{\underline{m}+\underline{k}-1,\overline{m}\}\}.$$

Consequently, if $\{\underline{m}+\underline{k},\ldots,\overline{m}+\underline{k}\}\cap\{0,\ldots,N_0-x\}\neq\emptyset$, then $\overline{m}<\underline{m}+\underline{k}-1$ (if not, the definition of x would be contradicted); otherwise, point 4 is vacuous. In that case, $N_0-x<\underline{m}+\underline{k}$, and the preceding argument already gives

$$\tilde{\kappa}(x,m)=\min\{\underline{k},m\}\quad\text{and}\quad V(x+m)=1+\delta V^{\text{aux}}(x,m)$$

for every $m\in\{0,\ldots,N_0-x\}$. Hence point 3 holds as well. Thus, for the remainder of the argument suppose $\{\underline{m}+\underline{k},\ldots,\overline{m}+\underline{k}\}\cap\{0,\ldots,N_0-x\}\neq\emptyset$.

If $\overline{m}<\underline{m}+\underline{k}-1$, however, $\tilde{\kappa}(x,m)=\min\{\underline{k},m\}$ and $1+\delta V^{\text{aux}}(x,m)=V(x+m)$ also for all $m\in\{\overline{m},\ldots,\underline{m}+\underline{k}-1\}\cap\{0,...,N_0-x\}$, and no $s\in B$ can ever be reached on the equilibrium path. Hence, by Lemma 1,

$$\tilde{\kappa}(x,m)=m-(\overline{m}+1)\quad\text{for all } m\in\{\underline{m}+\underline{k},\ldots,\overline{m}+\underline{k}\}\cap\{0,...,N_0-x\},$$

and consequently

$$V(x+m) = 1 + \delta V\big(x + m - \tilde{\kappa}(x,m)\big)\frac{x+m-\tilde{\kappa}(x,m)}{x+m} > 1 + \delta V^{\text{aux}}\big(x,m\big).$$

Since $1+\delta V^{aux}(x,m) \geq \frac{\rho}{1-\delta}$ for all $m \in \{\overline{m}+1, ..., N_0 - x\}$, this implies $1+\delta V(x+m) \geq 1 + \frac{\rho\delta}{1-\delta}$ for all $m \in \{\overline{m}+1, \ldots, \overline{m}+\underline{k}\} \cap \{0, ..., N_0 - x\}$. Thus, by induction on m, when $\overline{m}+\underline{k}+1 \leq N_0 - x$, Lemma 1 implies $\tilde{\kappa}(x,m) = \underline{k}$ (and $V(x+m) \geq 1+\delta V^{\text{aux}}(x,m)$) for all $m \in \{\overline{m}+\underline{k}+1, ..., N_0 - x\}$. Points 3 and 4 follow immediately.

To conclude the proof of Proposition 2, we observe three things. First, observe that the 4 point characterization of $\tilde{\kappa}(x,m)$ implies that for any initial number of agents N_0, there is a point $x \leq N_0$ and time t^*, such that

- t^* is the first time that x agents is reached
- principal eliminates $\underline{k}$ agents each period, except possibly at $t^* - 1$ and at most one other period
- after t^*, no more eliminations occur.

Second, we observe that the uniqueness of the coalition-proof path of eliminations follows from the uniqueness of $\tilde{\kappa}(x,m)$ satisfying our characterization.

Third and finally, we observe that it is straightforward to construct a coalition-proof SPE given 1 and our characterization of $\tilde{\kappa}(x,m)$.

These observations conclude the proof of Propositions 2 and 3

□

Proof of Proposition 6

Proof. Without loss of generality, assume

$$v_1 > v_2 > \cdots > v_{N_0}.$$

We proceed by induction on the number of surviving agents. For any surviving set $\mathcal{N}_t$, let $\mathcal{H}(\mathcal{N}_t, m)$ denote the set of the m agents in $\mathcal{N}_t$ with the highest values of v_i.

Base step. Consider any history h^t with $\Omega_t = 0$ and $N_t \in \{N^*, \dots, N^* + \overline{k} - 1\}$. In this case the principal eliminates exactly $m = N_t - N^*$ agents, those in $\mathcal{H}(\mathcal{N}_t, m)$, in period t and targets no agent afterward. Indeed, eliminating all (and only) the agents in $\mathcal{H}(\mathcal{N}_t, m)$ leaves exactly N^* survivors, after which the principal has no incentive to eliminate additional agents[33]. Hence targeting $\mathcal{H}(\mathcal{N}_t, m)$ does not induce a protest of size at least $\overline{k}$ and yields the highest feasible payoff for the principal starting from $\mathcal{N}_t$. Therefore, in any coalition-proof SPE starting at $\mathcal{N}_t$, the principal eliminates exactly the agents in $\mathcal{H}(\mathcal{N}_t, m)$ and no agent afterward.

Inductive step. Fix $n \geq 1$ and suppose that for every history with $N_t \in \{N^* + (n-1)\overline{k}, \dots, N^* + n\overline{k} - 1\}$, the principal targets and eliminates exactly $\mathcal{H}\big(\mathcal{N}_t,\ N_t - \big(N^* + (n-1)\overline{k}\big)\big)$, that is, the $N_t - (N^* + (n-1)\overline{k})$ agents with the highest values of v_i in $\mathcal{N}_t$, and no agents afterward. Consider now any history h^t with $\Omega_t = 0$ and $N_t \in \{N^* + n\overline{k}, \dots, N^* + (n+1)\overline{k} - 1\}$. Define $m = N_t - (N^* + n\overline{k}) \in \{0, \dots, \overline{k} - 1\}$.

First we show that the principal cannot target more than m agents. Suppose instead that she targets a set $\mathcal{S}$, with $\overline{k} > |\mathcal{S}| \equiv S > m$ agents.[34] Unless the principal is blocked, the next-period population size becomes $N_{t+1} = N_t - S < N^* + n\overline{k}$. In that case, by the induction hypothesis, in period $t+1$ the principal would target and eliminate exactly $\mathcal{H}\big(\mathcal{N}_t \setminus \mathcal{S},\ N_t - S - (N^* + (n-1)\overline{k})\big)$, namely the $N_t - S - (N^* + (n-1)\overline{k})$ surviving agents with the highest values of v_i. But then, by Assumption 1, those agents strictly prefer to join the current protest rather than wait to be eliminated next period. Together with the agents in $\mathcal{S}$, they would therefore form a protest coalition of size at least $\overline{k}$. By Assumption 3, the principal strictly prefers to avoid such a protest. Hence, in equilibrium, she cannot target more than m agents.

Applying the preceding argument to the boundary case $m = 0$ shows that, at any history with $N_t = N^* + n\overline{k}$, the principal targets no agents. On the other hand, the principal can target m agents without inducing any significant protest: their elimination leads to a population size of $N_t - m = N^* + n\overline{k}$, after which, by the boundary case $m = 0$ just established, no further eliminations occur. Moreover, by applying the same argument to the continuation state, targeting less than m agents today still does not allow the principal to eliminate more than m agents in

[33] Since $v_i < 1$, any further elimination would result in a flow payoff lower than $u(N^*) - \sum_{i \in \mathcal{N}_t \setminus \mathcal{H}(\mathcal{N}_t, m)} v_i$.

[34] By Assumption 3, the principal never targets $\overline{k}$ or more agents: doing so would necessarily trigger a protest of size at least $\overline{k}$.

total; so the principal strictly prefers targeting all m agents in the first period.

It remains to show that, if the principal targets exactly m agents, those agents must be the elements of $\mathcal{H}(\mathcal{N}_t, m)$. However, this follows because these are precisely the m agents whose removal yields the largest immediate reduction in $\sum_{i\in\mathcal{N}_t} v_i$. This completes the induction.

Applying the result at $t = 0$, let $m^* \in \{0, \dots, \overline{k} - 1\}$ be such that $N_0 - m^* \in \{N^*, N^* + \overline{k}, N^* + 2\overline{k}, \dots\}$. Then the principal eliminates exactly m^* agents in period 0, namely the agents with the highest values of v_i, that is, agents $\{1, \dots, m^*\}$. No protest succeeds, and no further eliminations occur. □

Proof of Proposition 7

Proof. Let $\kappa(\cdot)$ and $V(\cdot)$ denote the target-count function and the agent value function in the unpredictable-order environment, as characterized in Lemma 1. For any finite set $S \subseteq \mathcal{N}_0$ and any $m \leq |S|$, let $H(S, m)$ denote the set of the m agents in S with the highest values of v_i. Because F is continuous, this set is almost surely unique.

We prove by induction the following stronger statement.

Claim. Consider any history h^t with $\Omega_t = 0$ such that $|\mathcal{N}_t| = n$, and suppose that the posterior over $(v_i)_{i\in\mathcal{N}_t}$ is exchangeable. Then:

1. every surviving agent's ex ante continuation payoff at that history equals $V(n)$;
2. the principal targets exactly $\kappa(n)$ agents;
3. conditional on targeting $\kappa(n)$ agents, she targets exactly the agents in $H(\mathcal{N}_t, \kappa(n))$;
4. no protest blocks the principal.

The claim applies at $t = 0$ because the prior is i.i.d. and hence exchangeable. Moreover, if at some proposal history $\ell^t = (h^t, K_t)$ no untargeted agent protests, then after the proposal is implemented the next-period survivor set is exactly $\mathcal{N}_{t+1} = \mathcal{N}_t \setminus K_t$. Since Definition 3 imposes exchangeability of the posterior over untargeted agents at every proposal history, the posterior over $(v_i)_{i\in\mathcal{N}_{t+1}}$ is again exchangeable. Thus the class of histories covered by the claim is closed under the continuation relevant for the equilibrium path.

Base case 1: $n \leq N^*$.

If $n \leq N^*$, the principal targets no agent. Indeed, removing any agent when the current population is weakly below N^* weakly lowers her flow payoff, and strictly lowers it almost surely because $v_i \leq 1$ and F is continuous. Hence no further elimination occurs, no untargeted agent protests, and every surviving agent's continuation payoff is

$$\frac{1}{1-\delta} = V(n).$$

Base case 2: $N^* < n \leq N^* + \overline{k} - 1$.

In this range the principal can target exactly $n - N^*$ agents while keeping the target set strictly smaller than $\overline{k}$. If she does so, the next-period population is N^*, after which no further elimination occurs, so no untargeted agent protests.

She cannot do better by targeting more than $n - N^*$ agents, because that would leave fewer than N^* survivors and is suboptimal given $v_i \leq 1$. She also cannot do better by targeting fewer than $n - N^*$ agents, because the remaining excess is still strictly smaller than $\overline{k}$ and could therefore be removed in the next period without inducing protest by untargeted agents; postponing those eliminations by one period is strictly worse than carrying them out immediately. Hence the principal targets exactly $n - N^*$ agents. Conditional on targeting that many agents, protest incentives depend only on the number of survivors, so the principal strictly prefers to target the agents in $H(\mathcal{N}_t, n - N^*)$.

Because the posterior over the current survivors is exchangeable, each agent is ex ante equally likely to belong to the target set. Therefore each surviving agent is untargeted with probability N^*/n, and her ex ante continuation payoff is

$$1 + \delta \frac{N^*}{n} V(N^*) = V(n),$$

where the equality is exactly the recursion in Lemma 1 for this range of n.

Inductive step.

Fix $n \geq N^* + \overline{k}$ and suppose the claim holds for every smaller population size $\{0, ..., n-1\}$. Consider any history h^t with $\Omega_t = 0$, $|\mathcal{N}_t| = n$, and exchangeable posterior over $(v_i)_{i \in \mathcal{N}_t}$.

Step 1: continuation value after any proposal of size $k < \overline{k}$.

Take any proposal $K \subseteq \mathcal{N}_t$ with $|K| = k < \overline{k}$, and let

$$U \equiv \mathcal{N}_t \setminus K, \qquad |U| = n - k.$$

By Definition 3, conditional on the proposal history $\ell^t = (h^t, K)$, the posterior over $(v_i)_{i \in U}$ is exchangeable. Since types are almost surely distinct, this implies that every identity in U is equally likely to occupy any rank in the ordering of types within U.

If no untargeted agent protests, then the next-period survivor set is exactly U, whose posterior is still exchangeable, so the induction hypothesis applies at population size $n-k$. Therefore, from the next period onward, the principal targets exactly $\kappa(n-k)$ agents, conditional on that number she targets the set $H(U, \kappa(n-k))$, and each surviving agent's ex ante continuation payoff equals $V(n-k)$.

Hence every untargeted agent's continuation payoff from not protesting at the current proposal history is exactly $1 + \delta V(n-k)$. Crucially, this depends only on the number of survivors $n-k$, not on the identities of the currently targeted agents.

Step 2: feasible target counts are the same as in the unpredictable-order environment.

First, the principal never chooses $k \geq \overline{k}$, because then the targeted agents alone can form a protest of size at least $\overline{k}$, which she strictly prefers to avoid by Assumption 3.

Second, she never chooses k such that $n - k < N^*$, because leaving fewer than N^* survivors is suboptimal given $v_i \leq 1$.

Now consider any $k \in \{1, \ldots, \overline{k} - 1\}$ with $n - k \geq N^*$. By Step 1, every untargeted agent gets $1 + \delta V(n-k)$ from not protesting. Any k such that

$$1 + \delta V(n-k) < 1 + \frac{\rho\delta}{1-\delta},$$

cannot be part of a coalition proof equilibrium: it would induce untargeted agents to join the protest of targeted agents in a size $\overline{k}$ protest that the principal wants to avoid.

Conversely, $k \in \{1, \ldots, \overline{k} - 1\}$ such that

$$1 + \delta V(n-k) \geq 1 + \frac{\rho\delta}{1-\delta},$$

are *feasible* (i.e., avoid successful protests) as untargeted agents do not join.

Thus the set of feasible target counts at population size n is exactly

$$\{0\} \cup \left\{k \in \{1, \ldots, \overline{k} - 1\} : n - k \geq N^* \text{ and } 1 + \delta V(n-k) \geq 1 + \frac{\rho\delta}{1-\delta}\right\},$$

which is exactly the feasible set in Lemma 1.

Step 3: conditional on a feasible count, the principal targets the highest types.

Fix a feasible count $q < \overline{k}$ and suppose the principal targets some set $\mathcal{K} \subseteq \mathcal{N}_t$ with $|\mathcal{K}| = q$. Suppose $\mathcal{K} \neq \mathcal{H}(\mathcal{N}_t, q)$. Then there exist $i \in \mathcal{N}_t \setminus \mathcal{K}$ and $j \in \mathcal{K}$ such that $v_i > v_j$. Define

$$\mathcal{K}' \equiv (\mathcal{K} \setminus \{j\}) \cup \{i\}, \qquad \mathcal{S} \equiv \mathcal{N}_t \setminus \mathcal{K}, \qquad \mathcal{S}' \equiv \mathcal{N}_t \setminus \mathcal{K}'.$$

By Step 2, feasibility depends only on the resulting number of survivors $n-q$, not on the identities of the targeted agents. Thus $\mathcal{K}'$ is feasible whenever $\mathcal{K}$ is feasible. Moreover,

$$\mathcal{S}' = (\mathcal{S} \setminus \{i\}) \cup \{j\},$$

so $|\mathcal{S}'| = |\mathcal{S}|$ and $\sum_{\ell \in \mathcal{S}'} v_\ell < \sum_{\ell \in \mathcal{S}} v_\ell$. From period $t+1$ onward, the set of feasible target counts depends only on the current number of survivors, by Step 2. Since $|\mathcal{S}'| = |\mathcal{S}| = n - q$, starting from $\mathcal{S}'$, the principal can replicate any future sequence of target counts that is feasible starting from $\mathcal{S}$. Along that replicated sequence, let $\mathcal{S}'_r$ and $\mathcal{S}_r$ denote the survivor sets after r continuation periods starting from $\mathcal{S}'$ and $\mathcal{S}$, respectively. Then $\sum_{\ell \in \mathcal{S}'_r} v_\ell \leq \sum_{\ell \in \mathcal{S}_r} v_\ell$ for every $r \geq 0$, with strict inequality at $r = 0$. Therefore the principal's continuation payoff is strictly higher under $\mathcal{K}'$ than under $\mathcal{K}$, contradicting optimality of $\mathcal{K}$. Hence, conditional on targeting q agents, the principal targets exactly the agents in $\mathcal{H}(\mathcal{N}_t, q)$.

Step 4: among feasible counts, the principal chooses the largest one.

Fix two feasible counts

$$0 \leq k < k' \leq \overline{k} - 1.$$

By Step 3, if the principal targets k agents today, the next-period survivor set is

$$S_k \equiv \mathcal{N}_t \setminus H(\mathcal{N}_t, k),$$

while if she targets k' agents today, the next-period survivor set is

$$S_{k'} \equiv \mathcal{N}_t \setminus H(\mathcal{N}_t, k').$$

Hence

$$S_{k'} \subset S_k \qquad \text{and} \qquad |S_{k'}| = n - k' < n - k = |S_k|.$$

Consider the equilibrium continuation path after choosing k today, and let

$$n_1 = n - k > n_2 > \cdots$$

denote the successive population sizes along that path.

Case 1: the path never reaches a population weakly below $n - k'$.

Suppose

$$n_s > n - k' \qquad \text{for every } s \geq 1.$$

Then along the path that starts with k, the total number of eliminated agents is always strictly smaller than k'. By Step 3, each future survivor set therefore strictly contains $S_{k'}$.

Now consider the alternative plan in which the principal targets k' agents today and then stops forever. From period $t+1$ onward, this leaves the survivor set $S_{k'}$ forever. Since every survivor set along the path that starts with k strictly contains $S_{k'}$, and the principal's per-period payoff is strictly higher when the surviving set is smaller, this alternative yields a strictly higher continuation payoff. Hence k cannot be optimal.

Case 2: the path first reaches a population weakly below $n - k'$.

Let $r \geq 2$ be the first index such that $n_r \leq n - k'$. By minimality of r, we have $n_{r-1} > n - k$'.

By Step 3, when a total of $n - n_r$ agents have been eliminated from $\mathcal{N}_t$, the survivor set is

$$S_{n_r} \equiv \mathcal{N}_t \setminus H(\mathcal{N}_t, n - n_r),$$

independently of how those eliminations are distributed over time. Therefore the path that starts with k reaches the continuation state S_{n_r} for the first time at period $t + r$.

Now the move from population n_{r-1} to population n_r is feasible on that path. By Step 2, feasibility depends only on the destination population size, so any proposal that leaves exactly n_r survivors and targets fewer than $\overline{k}$ agents is feasible. Since $n_{r-1} > n - k'$ and $n_{r-1} - n_r < \overline{k}$, it follows that $0 \leq (n - k') - n_r < \overline{k}$. Hence, after targeting k' agents today, the principal can in period $t + 1$ target exactly $(n - k') - n_r$ additional agents, namely the highest-v_i agents in $S_{k'}$, and thereby reach the same continuation state S_{n_r} by period $t + 2$.

Therefore, the path that starts with k' reaches the same continuation state as the path that starts with k, but weakly earlier. If $r = 2$, both paths reach S_{n_r} at the same date, but the path that starts with k' already yields a strictly higher payoff at period $t+1$ because $S_{k'} \subset S_k$. If $r > 2$, the path that starts with k' both yields a strictly higher payoff at period $t+1$ and reaches S_{n_r} strictly earlier. In either case, choosing k' strictly dominates choosing k.

Since this argument applies to every feasible pair $k < k'$, the principal chooses the largest feasible count. By Step 2, that count is exactly $\kappa(n)$.

Step 5: agent values.

By Step 4, the principal targets exactly $\kappa(n)$ agents. By Step 3, conditional on this target count, she targets exactly the agents in $\mathcal{H}(\mathcal{N}_t, \kappa(n))$. Since the posterior over $(v_i)_{i\in\mathcal{N}_t}$ is exchangeable, each current survivor is ex ante equally likely to belong to that target set. Therefore each survives the current proposal with probability $\frac{n-\kappa(n)}{n}$. Because no untargeted agent protests, the ex ante continuation payoff of any surviving agent is

$$1 + \delta \frac{n - \kappa(n)}{n} V(n - \kappa(n)).$$

By Lemma 1, this equals $V(n)$. This proves the claim for population size n and completes the induction.

Applying the claim at the initial history yields the proposition: along the equilibrium path, the number of agents targeted in each period and protest outcomes coincide with those in the unpredictable-order environment with the same initial population N_0, and conditional on targeting k_t agents in period t, the principal targets the k_t surviving agents with the highest values of v_i. □

A.3 Proof of Proposition 4

Proof. Fix B_j. For every $m \in \{1, \dots, \overline{k} - 1\}$, starting from $B_j + m$ the principal can target m agents and reach B_j. Untargeted agents then anticipate no further elimination and do not protest. Hence none of these population sizes is blocking, so $B_{j+1} - B_j \geq \overline{k}$.

Moreover, since B_j is terminal we have $V(B_j) = \frac{1}{1-\delta}$, and from $B_j + m$ the principal targets exactly m agents, so the recursion in Lemma 1 gives $V(B_j + m) = 1 + \frac{\delta B_j}{(1-\delta)(B_j+m)}$, which is decreasing in m. Thus, from $B_j + \overline{k}$, the easiest positive proposal to sustain targets $\overline{k} - 1$

agents and leaves $B_j + 1$ survivors. By the protest condition, $B_j + \overline{k}$ is blocking if and only if $V(B_j + 1) < \frac{\rho}{1-\delta}$, or, equivalently,

$$B_j < \frac{\delta}{1-\rho} - 1 = \widehat{N}(\delta, \rho).$$

Therefore the next gap equals $\overline{k}$ if $B_j < \widehat{N}(\delta, \rho)$ and is strictly larger otherwise. Finally, maximality of B_n implies $B_n \geq \widehat{N}(\delta, \rho)$; otherwise, $B_n + \overline{k}$ would be another blocking size.

The second part of the proposition follows from the fact that under unpredictable-order coalition-proof SPE the principal's strategy at a given population size N_t is independent of N_0, by Lemma 1. As a result, for any parameterization, once $N_0 > B_n + \overline{k}$ and lies beyond the finitely many population sizes at which the principal applies a non-greedy elimination, the principal eliminates at least $\overline{k} - 1$ agents immediately as well as at least $\overline{k}$ agents in aggregate, achieving a strictly higher payoff than under predictable-order coalition-proof SPE. □

A.4 Proof of Proposition 5

Recall. Fix the primitives of Section 4.2, write $\underline{k} = \overline{k} - 1$, and set

$$p \;:=\; \frac{\rho}{1-\delta},$$

the value at which an untargeted agent is indifferent to joining a protest (Assumption 1; protest set displayed in the proof of Proposition 2). We make $\underline{k}$ an explicit argument of V^{auxx} and V^{aux}, i.e. $V^{\text{auxx}}(x, m, q; \underline{k})$ and $V^{\text{aux}}(x, m; \underline{k})$.[35] Monotonicity of V^{aux} on each block I_q is governed by

$$S(q; x, \underline{k}) = a_q x - b_q = \underline{k} - \delta^q \left[(1-\delta)x + \underline{k}\big(1 + (1-\delta)q\big)\right],$$

which is strictly increasing in q by (Diff-S), with single crossing $q^*(x; \underline{k}) = \min\{q : S(q; x, \underline{k}) \geq 0\}$. The protest (lower-contour) set $B(x; \underline{k}) = \{m : 1 + \delta V^{\text{aux}}(x, m; \underline{k}) < p\}$ is, by quasi-convexity, a contiguous interval $\{\underline{m}, \ldots, \overline{m}\}$ or empty (proof of Proposition 2); on the decreasing branch its lower endpoint is the ceiling $\underline{m}(x; \underline{k}) = \lceil \widehat{m}(x; \underline{k}) \rceil$ of the unique continuous crossing solving $1 + \delta V^{\text{aux}}(x, \widehat{m}; \underline{k}) = p$. By the *4-point characterization of* $\tilde{\kappa}$ in the proof of Proposition 2,

[35] Since $V^{\text{auxx}}(x, (q+1)\underline{k}, q; \underline{k}) = V^{\text{auxx}}(x, (q+1)\underline{k}, q+1; \underline{k})$ by (Bdry Matching), $V^{\text{aux}}(\cdot)$ extends continuously in m, we abuse notation and write $V^{\text{aux}}(x, m; \underline{k})$ for this continuous extension).

the tuple $\mathcal{B}(\overline{k}) = (B_0, \ldots, B_n)$ of Proposition 2 satisfies, on sections where $\overline{m}_j - \underline{m}_j \geq \underline{k} - 1$ holds,

$$B_j(\overline{k}) = B_{j-1}(\overline{k}) + \underline{m}_j(\overline{k}) + \underline{k}, \qquad \underline{m}_j(\overline{k}) = \underline{m}\big(B_{j-1}(\overline{k}); \underline{k}\big). \tag{3}$$

More precisely, the 4-point characterization in Proposition 2 gives the following recursive stopping rule. Given $B_{j-1}(\overline{k})$, consider the lower-contour set

$$B(B_{j-1}(\overline{k}); \underline{k}) = \{m : 1 + \delta V^{aux}(B_{j-1}(\overline{k}), m; \underline{k}) < p\}.$$

If this set is empty, the recursive construction stops. If it is nonempty, write

$$B(B_{j-1}(\overline{k}); \underline{k}) = \{\underline{m}_j, \underline{m}_j + 1, \ldots, \bar{m}_j\}.$$

Then a new finite threshold is generated if and only if

$$\bar{m}_j - \underline{m}_j \geq \underline{k} - 1.$$

In that case,

$$B_j(\overline{k}) = B_{j-1}(\overline{k}) + \underline{m}_j + \underline{k}.$$

If instead

$$\bar{m}_j - \underline{m}_j < \underline{k} - 1,$$

the recursive construction stops.

We impose

$$N^* = 0, \qquad \text{so that} \qquad B_0 = 0, \quad B_1 = \max\{N^*, \overline{k}\} = \overline{k} \tag{4}$$

Lemma 2 (Joint homogeneity in $(x, m, \underline{k})$)**.** *For every $\lambda > 0$, every $q \in \mathbb{N}$, and all (x, m),*

$$V^{\text{auxx}}(\lambda x, \lambda m, q; \lambda \underline{k}) = V^{\text{auxx}}(x, m, q; \underline{k}).$$

Consequently $S(q; \lambda x, \lambda \underline{k}) = \lambda\, S(q; x, \underline{k})$, so q^* and the protest-set block index are invariant under $(x, \underline{k}) \mapsto (\lambda x, \lambda \underline{k})$, and the continuous crossing is positively homogeneous of degree one, $\widehat{m}(\lambda x; \lambda \underline{k}) = \lambda\, \widehat{m}(x; \underline{k})$.

Proof. In the closed form, b_q scales by λ since $b_q(\lambda x; \lambda \underline{k}) = (1 - \delta)\lambda x - \lambda \underline{k} + \lambda \underline{k} \delta^q [1 + (1 - \delta)q] = \lambda b_q(x; \underline{k})$, the numerator $a_q(\lambda m) + b_q(\lambda x; \lambda \underline{k})$ scales by λ, and the denominator $(\lambda m + \lambda x)$

scales by λ; the ratio is unchanged. Then $S(q;\lambda x,\lambda\underline{k}) = a_q\lambda x - b_q(\lambda x;\lambda\underline{k}) = \lambda S(q;x,\underline{k})$ has the same sign, so q^* and the decreasing branch are unchanged, as is the block index defined by $1+\delta V^{\text{auxx}}(x,(q+1)\underline{k},q;\underline{k}) < p$. If $\widehat{m}$ solves $1+\delta V^{\text{aux}}(x,\widehat{m};\underline{k}) = p$, then by the displayed identity (applied to the continuous extension) $\lambda\widehat{m}$ solves $1+\delta V^{\text{aux}}(\lambda x,\cdot\,;\lambda\underline{k}) = p$; uniqueness on the decreasing branch gives $\widehat{m}(\lambda x;\lambda\underline{k}) = \lambda\widehat{m}(x;\underline{k})$. □

By Lemma 2, $\widehat{m}(x;\underline{k})/\underline{k}$ depends on $(x,\underline{k})$ only through $\beta = x/\underline{k}$; write

$$g(\beta;\rho,\delta) \;:=\; \frac{\widehat{m}(\beta\underline{k};\underline{k})}{\underline{k}}$$

for this *normalized crossing*. The crossing sits in the block I_{q_ℓ}, where $q_\ell = q_\ell(\beta;\delta) = \lfloor g(\beta;\rho,\delta)\rfloor < q^*$; on that block, multiplying the equation $1+\delta V^{\text{aux}}(x,\widehat{m};\underline{k}) = p$ by $(\widehat{m}+x)$ yields an equation that is affine in $\widehat{m}$ with a constant coefficient on $\widehat{m}$ independent of x, so the normalized root g is affine in β there. Hence g is piecewise linear and continuous, its kinks occurring exactly where the crossing meets a block endpoint $\widehat{m}\in\underline{k}\mathbb{Z}$, i.e. where $g(\beta;\rho,\delta)\in\mathbb{Z}$. For $j\geq 1$, write $\beta_j(\overline{k}) = B_{j-1}(\overline{k})/\underline{k}$. Going forward, we suppress dependence of $g(\beta;\rho,\delta)$ on (ρ,δ).

For later use, also define

$$v^c(\beta;\delta) := \inf_{\mu\geq 0} V^{aux}(\beta\underline{k},\mu\underline{k};\underline{k}).$$

which, by Lemma 2, is independent of $\underline{k}$. Whenever $1+\delta v^c(\beta;\delta) < p$ and hence $B(\beta\underline{k};\underline{k})$ is non-empty denote by $\widehat{m}_\ell(\beta\underline{k};\underline{k})$ and $\widehat{m}_u(\beta\underline{k};\underline{k})$, the smallest and largest solutions of $1+\delta V^{aux}(\beta\underline{k},m;\underline{k}) = p$. Define

$$w(\beta;\rho,\delta) := \frac{\widehat{m}_u(\beta\underline{k};\underline{k}) - \widehat{m}_\ell(\beta\underline{k};\underline{k})}{\underline{k}}.$$

which again, by Lemma 2, is independent of $\underline{k}$. As with g, we suppress dependence of w on (ρ,δ) going forward.

Proof of Proposition 5. Write $\underline{k} = \overline{k}-1$, and for the scaled cutoff $\overline{k}' := \kappa(\overline{k}-1)+1$ put $\underline{k}' = \overline{k}'-1 = \kappa\underline{k}$.

The ratios converge to a limit that does not depend on the cutoff, conditional on continuation. By Lemma 2, the construction depends on the cutoff only through β, so throughout this step $\overline{k}$ denotes a generic cutoff. At any step at which the recursive construction generates a new finite

threshold, equation (3) applies. Thus, if $B_j(\overline{k})$ is finite, dividing (3) by $\underline{k}$ and using $\underline{m} = \lceil \widehat{m} \rceil$ and $\widehat{m}/\underline{k} = g$ gives

$$\beta_{j+1}(\overline{k}) = \beta_j(\overline{k}) + \frac{\underline{m}_j(\overline{k})}{\underline{k}} + 1 = \beta_j(\overline{k}) + g(\beta_j(\overline{k})) + 1 + O\left(\tfrac{1}{\underline{k}}\right),$$

the $O(1/\underline{k})$ being the integer-rounding error. By (4), $\beta_1(\overline{k}) = \overline{k}/\underline{k} = 1 + 1/\underline{k} \to 1 =: \beta_1^\infty$. Define the *limiting recursion* by $\beta_{j+1}^\infty = \beta_j^\infty + g(\beta_j^\infty) + 1$, the recursion continuing at step j if and only if $1 + \delta v^c(\beta_j^\infty; \delta) < p$ and $w(\beta_j^\infty; \rho, \delta) \geq 1$, and terminating otherwise. It is straightforward to show that it terminates after finitely many steps; let $n^\infty < \infty$ be such that the limiting sequence is $(\beta_1^\infty, \ldots, \beta_{n^\infty+1}^\infty)$.

Same length. By Proposition 2, the recursion appends a new finite entry B_j exactly when the trough of the fastest-approach value at stopping point B_{j-1} lies strictly below p and $\bar{m}_j(\overline{k}) - \underline{m}_j(\overline{k}) \geq \underline{k} - 1$; otherwise the the tuple terminates. Fix $j \geq 1$ and suppose, for the moment, that the following holds:

$$\textit{for all sufficiently large } \overline{k}, \mathcal{B}(\overline{k}) \textit{ contains the entry } B_{j-1}(\overline{k}), \textit{ and } \beta_j(\overline{k}) \to \beta_j^\infty. \tag{H_j}$$

This hypothesis is verified by induction at the end of this step. As in the previous step, $\overline{k}$ denotes a generic cutoff: by Lemma 2, v^c, w, and g depend only on β. By Lemma 2 the *continuous* trough is a function $v^c(\beta; \delta)$ of the ratio β alone; the discrete trough at B_{j-1} differs from $v^c(\beta_j; \delta)$ by $O(1/\underline{k})$,[36] hence converges to $v^c(\beta_j^\infty; \delta)$ as $\overline{k} \to \infty$, and whenever $1 + \delta v^c(\beta_j^\infty; \delta) < p$,

$$\frac{\bar{m}_j(\overline{k}) - \underline{m}_j(\overline{k})}{\underline{k}} \to w(\beta_j^\infty; \rho, \delta).$$

The condition $\bar{m}_j(\overline{k}) - \underline{m}_j(\overline{k}) \geq \underline{k} - 1$ can be written

$$\frac{\bar{m}_j(\overline{k}) - \underline{m}_j(\overline{k})}{\underline{k}} \geq 1 - \frac{1}{\underline{k}}.$$

Since

$$\frac{\bar{m}_j(\overline{k}) - \underline{m}_j(\overline{k})}{\underline{k}} \to w(\beta_j^\infty; \rho, \delta),$$

if $w(\beta_j^\infty; \rho, \delta) > 1$, then $\bar{m}_j(\overline{k}) - \underline{m}_j(\overline{k}) \geq \underline{k} - 1$ holds for all sufficiently large $\overline{k}$, while if $w(\beta_j^\infty; \rho, \delta) < 1$, then $\bar{m}_j(\overline{k}) - \underline{m}_j(\overline{k}) \geq \underline{k} - 1$ fails for all sufficiently large $\overline{k}$. When $w(\beta_j^\infty; \rho, \delta) =$

[36] In fact, the two coincide: within each block $m \in [q\underline{k}, (q+1)\underline{k}]$, $\frac{\partial}{\partial m} V^{\text{aux}}(x, m; \underline{k}) = \frac{S(q;x,\underline{k})}{(1-\delta)^2(m+x)^2}$, so by single crossing of S the continuous infimum is attained at the block boundary $m = q^*(x; \underline{k})\,\underline{k} \in \mathbb{N}$, an integer.

1, the limit alone does not decide the finite inequality. Accordingly, let $\mathcal{E}$ be the set of primitive pairs (δ, ρ) for which, at some β_j^∞, either $1 + \delta v^c(\beta_j^\infty; \delta) = p$, or both $1 + \delta v^c(\beta_j^\infty; \delta) < p$ and $w(\beta_j^\infty; \rho, \delta) = 1$. It is straightforward to show that $\mathcal{E}$ has measure zero. For the rest of the proof, fix $(\rho, \delta) \notin \mathcal{E}$. Outside $\mathcal{E}$, both continuation conditions hold with strict inequality at β_j^∞ for every $j \leq n^\infty$, and at $j = n^\infty + 1$ the failing condition fails strictly.

Outside this null set, each of the finitely many continuation conditions for $j \leq n^\infty$ holds with a *strict* inequality at β_j^∞, and the reverse strict inequality first occurs at $j = n^\infty + 1$ (this defines $n^\infty < \infty$).

We now verify, by induction on j, that (H_j) holds for every $j \leq n^\infty + 1$, and that no entry is generated at $j = n^\infty + 1$.

- *Base case ($j = 1$):* the construction generates B_1 by (4), and $\beta_1(\overline{k}) \to \beta_1^\infty$ was shown above.
- *Inductive step ($j \leq n^\infty$):* given (H_j), the two displayed limits above show that the discrete trough and width conditions at B_j converge to their limiting counterparts at β_j^∞, which hold strictly; hence there exists $\overline{N}_j$ such that for all $\overline{k} > \overline{N}_j$ the construction generates B_{j+1}, and (3), via the recursion displayed at the start of the proof, yields $\beta_{j+1}(\overline{k}) \to \beta_{j+1}^\infty$, i.e., ($H_{j+1}$) holds.
- *Termination ($j = n^\infty + 1$):* by the same convergence, the strictly failing condition at $\beta_{n^\infty+1}^\infty$ implies that for all sufficiently large $\overline{k}$ the construction does not generate $B_{n^\infty+1}$. Setting $\overline{N}(\kappa, \delta, \rho) = \max_{j \leq n^\infty+1} \overline{N}_j$, for all $\overline{k} > \overline{N}$ the tuple is exactly $\mathcal{B}(\overline{k}) = \big(0, B_1(\bar{k}(\overline{k}), \ldots, B_{n^\infty}(\overline{k})\big)$.

Since $\overline{k}' = \kappa \underline{k} + 1 \geq \overline{k}$, both cutoffs exceed $\overline{N}$ whenever $\overline{k} > \overline{N}$. Applying the conclusion at $\overline{k}$ and at $\overline{k}'$ gives $n(\overline{k}) = n(\overline{k}') = n^\infty$ and, for every $j \leq n^\infty$, $\beta_{j+1}(\overline{k}) \to \beta_{j+1}^\infty$ and $\beta_{j+1}(\overline{k}') \to \beta_{j+1}^\infty$.

Normalized limit. The zero component scales to 0. Fix a component $1 \leq j \leq n^\infty$. Since $B_j(\overline{k})/\underline{k} = \beta_{j+1}(\overline{k}) \to \beta_{j+1}^\infty$ and $\underline{k}/\overline{k} \to 1$,

$$\frac{\kappa\, B_j(\overline{k})}{\overline{k}} = \kappa\, \frac{\underline{k}}{\overline{k}}\, \beta_{j+1}(\overline{k}) \longrightarrow \kappa\, \beta_{j+1}^\infty.$$

Likewise $B_j(\overline{k}')/\underline{k}' = \beta_{j+1}(\overline{k}') \to \beta_{j+1}^\infty$ and $\underline{k}'/\overline{k} = \kappa \underline{k}/\overline{k} \to \kappa$, so

$$\frac{B_j(\overline{k}')}{\overline{k}} = \frac{\underline{k}'}{\overline{k}}\, \beta_{j+1}(\overline{k}') \longrightarrow \kappa\, \beta_{j+1}^\infty.$$

The two normalized components share the limit $\kappa\,\beta_j^{\infty}$, so their difference is $o(1)$. This completes the proof.

□