Original Manuscript

Journal of Conflict Resolution
2023, Vol. 0(0) 1–29

DOI: 10.1177/00220027231222004
journals.sagepub.com/home/jcr


How Loyalty Trials Shape Allegiance to Political Order

Mirko Reul[1] **and Ravi Bhavnani**[2]

Abstract

"Loyalty trials" are common to a range of conflict settings, with consequences that range from harassment to imprisonment, torture, or death. Yet, they have received little if any attention as a general phenomenon in studies of state repression, civil war, or rebel governance, which focus on particular behaviors that authorities use to put people on trial, such as dissent, defection, and resistance. Using a computational model and data on the German Democratic Republic and the Occupied Palestinian Territories, we focus on the dynamics of "loyalty trials" held to identify enemy collaborators—the interaction between expectations, perceptions, and behavior. We use our framework to explore the conditions under which trials result in widespread defection, as in the German Democratic Republic, or in conformity as illustrated by our study of the Occupied Palestinian Territories. The polarizing nature of loyalty trials and the propensity to over- or under-identify threats to political order have notable implications for democratic and non-democratic societies alike.

Keywords

social conflict, labeling, defection, Occupied Palestinian Territories, German Democratic Republic

[1]Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland.
[2]Department of International Relations and Political Science, Graduate Institute of International and Development Studies, Geneva, Switzerland

Corresponding Author:
Mirko Reul, Faculty of Business and Economics, University of Lausanne, 1015 Lausanne, Switzerland.
Email: mirko.reul@unil.ch

Data Availability Statement included at the end of the article

Introduction

From rebel rulers to nation-state governments, we observe a recurrent pattern as political actors define loyalty expectations for their subjects, delineating deviant behavior that "departs from the normative" and poses a threat to political order (see Raybeck 1991, 23). In the wake of the 2015 Paris attacks, emergency laws authorized raids to identify individuals capable of causing "big harm" (Chassany 2017). The Chinese government placed millions of Uyghurs under surveillance and detained hundreds of thousands "under suspicion of political disloyalty" to prevent separatism (Roberts 2018, 19). And following the Taliban takeover in Afghanistan, individuals who collaborated with the Karzai regime either fled the country or went into hiding to avoid being identified, despite official assurances of amnesty (Gwladys et al. 2021).

Spanning the gamut from ethnic minorities with alleged ties to extremists (Mueller and Stewart 2012) to collaborators accused of being subversive foreign agents (O'Brian 1948), political actors hold loyalty trials to more palpably distinguish conformers from defectors, and label those identified as deviant (see Åkerström 1991, 11-16; Coser 1956). Trials comprise long judicial procedures or instant judgements devoid of due process, as with public denunciations or vigilante killings, typically held in a "zone of anomie in which legal determinations—the very distinction between public and private—are deactivated" (Agamben 2005, 50-51). Underpinning these ostensibly disparate phenomena is the notion of *betrayal*—established, perceived, or fabricated. Whereas the 'true' loyalty of individuals is typically private information, the onus lies on those "labeled" to demonstrate loyalty, failing which they are deemed culpable of defection—defectors then subject to harassment, ostracism, imprisonment, torture or death. As political actors specify the criteria for membership in their communities (see Schlichte and Schneckener 2015, 415; Thiranagama and Kelly 2010, 2) and use trials as an enforcement mechanism, their actions spur widespread social conformity or resistance.

Research on defection has paid scant attention to what constitutes disloyalty, disregarding the interplay between expectations, perceptions, and behavior—the dynamic nature of "labeling defection" (see, for example, work by Kalyvas 2008; Schutte 2017; Sullivan 2016a). Given that each loyalty trial effectively redefines the boundaries between behavior considered acceptable and unacceptable, who accuses whom, who defects or conforms, and under what conditions, is key. The challenge, in this regard, lies in reconstructing the individual experiences of the labelers and the labeled, as well as the associated implications. Our effort to better understand the dynamics of loyalty trials is driven by two key questions: under what conditions do loyalty trials generate conformity or defection? And in undertaking loyalty trials, to what extent do political actors over- or under-identify threats to their political order, prosecuting innocents (Type I error) or failing to prosecute those guilty of defection (Type II error)?

We draw on existing research in criminology and conflict studies to specify mechanisms that link loyalty trials to shifts in allegiance, formalizing our theory by means of an agent-based computational model (ABM). The model, validated by

qualitative evidence from two markedly different conflict settings (Bhavnani, Donnay, and Reul 2020)—the German Democratic Republic (GDR) during the leadership of Erich Honecker (1971-1989) and the Occupied Palestinian Territories (OPT) during the Second Intifada (2000-2004)—permits us to identify the drivers of allegiance for different types of regimes. In the GDR, the misidentification of defectors—quotidian behavior being perceived as disloyalty—generated cascades of defection given unrealistically high expectations. By contrast in the OPT, social incentives for loyalty drove an increase in group allegiance, effectively curtailing defection.

In the following section, we review existing literature that pertains to loyalty trials from various, closely related domains. We then specify the mechanisms that underpin loyalty trials, formalizing these by means of a computational model. Next, we show how our model simulates allegiance shifts between conformity and defection, adjusting model parameters to capture the particularities of the GDR and the OPT. We conclude by discussing the implications of our framework and analysis for loyalty trials held in both democratic and non-democratic settings.

Related Work

Loyalty trials are common to a range of conflict settings, yet, have received little if any explicit attention as a general phenomenon in studies of state repression, civil war, or rebel governance. Instead, studies focus on particular behaviors that authorities use to put people on trial, such as dissent, defection, and resistance, short of developing an account of what we refer to as "loyalty trials" that is comparable across contexts.

Beginning with the literature on state repression, the prevailing paradigm suggests that state actors repress social and political rights, with targets for repression selected on the basis of *de jure* or *de facto* rules (Tilly 2003). As such, repression may be overt or covert (Davenport 2005), preventive or reactive (Dragu and Przeworski 2019), target more open or hidden forms of mobilization (Sullivan 2016b), and serve to deter or increase future challenges (Lichbach 1987, 269). Opposition groups, in turn, adapt their behavior in response to opportunity structures (Tarrow 1994), with research focusing on varying expressions of "voice" or "exit" (Hirschman 1970; 1993).

Sullivan's work (2016b) on state repression in Guatemala is notable in this regard, given his coding of "overt" and "covert" mobilization perceived to constitute a political challenge. However, even here the focus is on exceptional mobilization against political order, disregarding quotidian behavior that poses no ostensible threat to the regime.[1] In a similar vein, work on counter-terrorism focuses on minimizing the over- or under-identification of exceptional defectors by state authorities (Dragu 2017; Polo and Wucherpfennig 2022), disregarding defection below a threshold of violent attacks.

Scholarship on civil war also pays short shrift to the interplay between perceived and private loyalty. Drawing on evidence from the Greek Civil War (1943-1949), Kalyvas (2006) argues that violence by armed groups is 'selective' in areas characterized by incomplete territorial control where defection to rival authorities occurs, yet he remains agnostic about the range of behaviors construed as disloyal as well as the consequences

of misidentification. Much of the literature that builds on Kalyvas' seminal work focuses on the cohesion of armed organizations (e.g. Sinno 2008; Staniland 2012; Pearlman and Cunningham 2012), including the conditions for fighters to desert or defect to rival organizations (Albrecht and Ohl 2016; McLauchlin 2010; Oppenheim et al. 2015; Koehler, Ohl, and Albrecht 2016) and those that underpin the incidence of selective or indiscriminate violence (Kalyvas 2012). Arjona (2016, 174-176) finds that armed groups vying for control of Colombian communities took popular norms into account, killing social deviants in an effort to 'bootstrap' their legitimacy and that local populations, in turn, exercised agency over denunciations to authorities (Arjona 2016). In her work on the Spanish Civil War, (Balcells 2010, 301-302) notes that local councils provided militias with lists of suspected right-wing supporters who were placed on loyalty trials, resulting in imprisonment or execution. By punishing those they could justifiably label as defectors, authorities assigned blame for governance failures to "defecting, criminal or disloyal elements among the fighters or the population" (Schlichte and Schneckener 2015, 419). Notable for its attention to the varied nature of authority-subject relations during civil conflict, including the ability of civilians to resist authorities (e.g. Arjona 2016), this literature also stops short addressing what behaviors and perceptions result in labeling of "threat" or "disloyalty" via loyalty trials.

A handful of case studies do consider how the interplay between loyalty expectations and perceptions of disloyalty result in loyalty trials: coercive models of social control were less likely to elicit denunciations than voluntary models during the Spanish Inquisition and in Romanov Russia (Bergemann 2017); popular perceptions drove the killing of Republican officers during the Spanish Civil War (McLauchlin and Parra-Pérez 2018); local populations in Afghanistan were found less likely to denounce enemy activity to ethnic others (Lyall, Shiraito, and Imai 2015), and minority Arab Americans with personal experiences of repression were more likely to protest in Detroit (Santoro and Azab 2015). In Mosul, a survey experiment on post-conflict perceptions found that civilians who collaborated with the Islamic State were more likely to be forgiven by their peers when service provision was perceived as involuntary (Kao and Revkin 2022). Yet, these rich and variegated studies fall short of formalizing the mechanisms that link defection to loyalty expectations and their associated consequences, what we turn to in the section that follows.

The Micro-dynamics of Loyalty Trials

We suggest that political order is co-produced by authorities who expect loyalty—personal sacrifice meant to enhance group welfare (Levine and Moreland 2002)—and subordinates, who to varying degrees, conform to loyalty expectations. The micro-dynamics of loyalty trials—the interplay between explicit and observable loyalty expectations and perceptions of disloyalty, on the one hand, and an individual's true allegiance, on the other—have significant implications for political order. We begin by discussing these dynamics below, before turning to our formal model.

The identification or labeling of defectors has a profound implications for how the labeled see themselves and are seen by others. In most instances, perceived defection is sufficient to initiate a loyalty trial, based on peer-to-peer accusations or official suspicion and arrest. Those labeled may or may not have violated loyalty expectations, and not all of those who violate expectations are labeled (Becker 1963, 9). To distinguish between 'true' and 'false' labels, loyalty trials consider both the motivations of suspects as well as perceptions of their behavior: loyalty, in this regard, is effectively co-produced by the 'labeler' and the 'labeled' (see Levine and Moreland 2002; Poulsen 2020, 9). A label can be perceived as false on substantive grounds when defection was not intended by the labeled; on procedural or emotional grounds when the conduct of the labeler is disrespectful (see Sherman 1993); or on normative grounds when defection is attributed to conflicting loyalties that are socially acceptable (see Sykes and Matza 1957).

Consequently, loyalty trials typically result in one of the four outcomes depicted in Table 1. When an individual is not labeled, she either conforms (*conformer*, cell I) or defects (*secret defector*, cell II); and when an individual is labeled, she may also conform (*false defector*, cell III) or defect (*defector*, cell IV). As such, conformers exceed loyalty expectations and are identified as loyal, and vice-versa for defectors. Secret defectors violate expectations but are perceived as loyal or tolerated (as described by Scott 1985; Wedeen 1999), whereas individuals who are perceived as disloyal but privately loyal are falsely labeled.

Second, defector labels generally present a claim that the individual may be threatening group goals to the benefit of a rival, thus directly challenging their status as a group member. But the reaction of the labeled depends on the particular circumstances under which the label was assigned. False defectors are expected to view themselves as members of the group and attempt to convince their peers of their innocence, engaging in demonstrably loyal behavior to do so. By contrast, defectors tend to have few opportunities to demonstrate loyalty, and may prefer to be ostracized from the group whose goals or beliefs they no longer identify with. Thus, labeling need not determine

Table 1. Typology of Individual Defection.

	Conforming	Defecting
~ Labeled	**I** "Conformer" True Negative	**II** "Secret Defector" Type II Error
Labeled	**III** "False Defector" Type I Error	**IV** "Defector" True Positive

Source: Adapted from Becker 1963, 20. Note: Conformers are privately loyal but not labeled as defectors. Defectors are both privately disloyal and labeled for behavior that falls short of loyalty expectations. The veracity of defector labels is determined in *loyalty trials*.

allegiance, though it reduces the agency of the labeled, coercing them into demonstrating loyalty or seeking acceptance by rival groups.

Third, defector labels directly politicize individual behavior, influencing how such behavior is perceived by those labeled and their peers, with notable consequences for political order. The constitutive act of labeling serves as a signal to others who exhibit similar characteristics or behavior. For the disloyal, trials alter both the perceived risks of being labeled (see Oliver, Marwell, and Teixeira 1985) and the benefits of collaboration (see Kalyvas 2006). For those tasked with labeling, true positives (*defector*, cell IV) serve to increase suspicion and mistrust, whereas true negatives (*conformer*, cell I) serve to maintain or increase trust. It follows that under certain conditions, loyalty trials may effectively erode challenges to political order or exacerbate them (see Lichbach 1987), as with over- or under-identification—false positives (*false defector*, cell III) or false negatives (*secret defector*, cell II) respectively (Schutte, 2017).

Fourth, private loyalty increases with social and material rewards—as with the approval of behavior that visibly benefits a group (see Marques et al. 1998) or monetary rewards for denouncing rival activity (e.g. Piotrowska 2020)—and decreases by means of social control and sanctions (see Hechter 1987; Heckathorn 1988). Disloyalty too is either positively incentivized by rival authorities—through public declarations of support and political asylum—or negatively under threat of punishment, as with the blackmailing of group members by intelligence organizations. Where the incentives provided by the ingroup exceed those provided by the outgroup, behavior is more likely to shift towards conformity, and vice-versa for defection (see Kalyvas 2008, 1059).

To summarize the discussion thus far, loyalty trials both directly and indirectly shape behavior in conflict settings: directly as a function of expectations, behavior, and perceptions of disloyalty; and indirectly by means of demonstration effects, as individuals observe the trials of others and act on private knowledge about their own behavior. The micro-dynamics of labeling therefore have consequences for both individual and group allegiance: when misidentification is low, group allegiance is likely to be maintained; as misidentification increases, allegiance is likely to shift towards conformity or defection, driven by expectation and the use of selective incentives, such as punishment and reward. Empirically, the micro-dynamics of loyalty trials—the interplay between loyalty expectations, private allegiance and perceptions, on the one hand, and individual reactions to loyalty trials, on the other—leads to socially complex outcomes that are challenging to conceptualize and analyze in a systematic fashion. In the section that follows, we formally specify the attributes, mechanisms and resulting behaviors.

Model Specification

We use an agent-based computational model (ABM) to systematically explore the relationship between loyalty expectations and perceptions of disloyalty on the one hand, and group conformity on the other. Agent-based modeling is a "computational approach that enables a researcher to create, analyze, and experiment with models

composed of agents that interact within an environment" (Gilbert 2008, 1). In an ABM, each agent's behavior shapes the behavior of other agents, as well as the properties of their shared social environment. The social environment, in turn, changes in response to changes in individual and aggregate behavior. Specified computationally, ABM can be run and rerun to assess variations in key model parameters, a task that is difficult to achieve by means of closed-form solutions. ABM therefore constitute one means of studying complex adaptive systems, and have been applied to a host of issue domains including residential segregation (Schelling 1969), political parties and elections (Kollman, Miller, and Page 1998), civil conflict (Epstein 2002) and urban violence (Bhavnani et al. 2014), to name but a few.

We begin by specifying a general model and in a second step, set model parameters to capture the particularities of loyalty trials in two contexts: the GDR and the OPT. Our specification builds on the Riolo, Cohen, and Axelrod (2001) tag-tolerance model, which is in turn motivated by Holland (1995). Readers who wish to skip the technical model description may move directly to the summary of model steps below.

Table 2 provides an overview of key model parameters. Each agent is defined by an i, p, q triplet $\in [0, 1]$, elements of which respectively signify private behavior, publicly perceived behavior, and tolerance for deviant behavior.

Whereas tag values vary across agents and over time, official loyalty expectations are given by $\lambda \in [0, 1]$. As loyalty expectations increase, so does the range of outgroup interactions considered unacceptable and the personal sacrifice required to maintain allegiance. When $\lambda = 1$, any indication of disloyalty is considered defection from the group. In such cases, even the failure to demonstrate group conformity, for example with violent attacks against nominal rivals, can lead to being labeled a defector. Conversely, $\lambda = 0$ signifies that there are no loyalty expectations.

Incentives, provided by a mix of rewards and punishments, are given by $k \in [-1, 1]$. When $k = 0$, incentives provided by the in- and outgroup are balanced, for example when a political authority taxes literature which glorifies its rivals just as much as the

Table 2. Key Model Parameters.

Agent-level	
i_A	A's private behavior
p_A	A's perceived behavior
q_A	A's tolerance for deviant behavior
l_A, d_A	A was labeled, is defecting
Group-level	
$\bar{i}, \bar{p}, \bar{q}$	Mean allegiance, allegiance perceptions, tolerance
$\sigma_{i,p}$, σ_q	Spread in allegiance, tolerance
λ	Loyalty expectations
k	Reward & Punishment

rival is willing to pay for its distribution, or when an ingroup vilifies and ostracizes regime critics but an outgroup is glorifying and welcoming the vilified as political refugees. Conversely, when $k = 1$, behavior in service of the ingroup is more strongly incentivized, whereas when $k = -1$, it is behavior in service of the outgroup that is incentivized more strongly.

We provide a formal description of key model mechanisms below. To interpret the results we focus on group conformity, defined by the difference between private behavior and loyalty expectations:

$$\Delta_\lambda = \sum_{A=1}^{N} i_A - \lambda \tag{1}$$

Additional outcomes, parameter sweeps and model specifications are provided in Appendix B.

Mechanism I: Loyalty Trials

Key Model Steps. We define the relationship between A's public allegiance and perceived deviation from loyalty expectations:

$$\delta_p = \lambda - p_A \tag{2}$$

Loyalty trials are conducted for $T = 10\%$ of agents in each iteration. An agent B has $p = 3$ opportunities to randomly select some other agent A for pairwise interaction.[2] The probability of selecting A over any other agent decreases with k or p_A:[3]

$$p(B \to A) = \frac{e^{k\delta_p}}{\sum_{A=1}^{N} e^{k\delta_p}} \tag{3}$$

When A's defection deviates from loyalty expectations more than B can tolerate, A is labeled a defector by B. Conversely, A is not labeled by B if her defection is tolerable:

$$\begin{aligned} \delta_p > q_B &\to l_A = 1, \\ \delta_p \leq q_B &\to l_A = 0 \end{aligned} \tag{4}$$

Irrespective of perceptions, A is defecting if private behavior violates loyalty expectations:

$$\begin{aligned} i_A < \lambda &\to d_A = 1, \\ i_A \geq \lambda &\to d_A = 0 \end{aligned} \tag{5}$$

A's defector type is then determined by crossing l_A, d_A:

$$\begin{aligned} d_A = 0 \wedge l_A = 0 &\rightarrow A^I : \text{conformer} \\ d_A = 1 \wedge l_A = 0 &\rightarrow A^{II} : \text{secretdefector} \\ d_A = 0 \wedge l_A = 1 &\rightarrow A^{III} : \text{falsedefector} \\ d_A = 1 \wedge l_A = 1 &\rightarrow A^{IV} : \text{defector} \end{aligned} \tag{6}$$

Mechanism II: Allegiance Shifts

We capture the direct effects of loyalty trials with updates to public and private allegiance. After every interaction t with agent B, the public perception of A's behavior is updated as follows:

$$p_{A_{t+1}} = p_{At} - p_{At}\delta_{pt} \tag{7}$$

It follows that perceptions are *additive* and *contagious*—the more (less) frequently A is perceived as a defector by some other agent, the greater (lower) the likelihood she will be perceived as a defector by others. We define the relationship between A's private behavior and deviance from loyalty expectations:

$$\delta_i = \lambda - i_A \tag{8}$$

After P interactions, labeled agents change their behavior based on deviance from loyalty expectations:

$$i_{Ag+1} = i_{Ag} - i_{Ag}\delta_{ig} \tag{9}$$

Thus, defectors with $i_A < \lambda$ decrease allegiance, and false defectors with $i_A > \lambda$ increase allegiance.

Mechanism III: Adaptation

We capture indirect effects of loyalty trials by updating agent characteristics. First, for every true (false) defector, aggregate tolerance decreases (increases):

$$\overline{q}_{g+1} = \overline{q_g} + \left(\sum_{A=1}^{N} A^{III} - \sum_{A=1}^{N} A^{IV} \right) \frac{1}{N} \tag{10}$$

When defectors outnumber false defectors, aggregate tolerance for defection decreases, and vice-versa, with intensity increasing as a function of labeled defectors.

Second, we assign a fitness score to agents that increases with the distance between private and perceived behavior and deviance from loyalty expectations—an effect moderated by punishment and reward—and decreases with labeling. Formally:

$$f_A = \frac{\delta_i^2}{e^{k\delta_i}} - |p_A - i_A| \cdot l_A \tag{11}$$

We provide a more detailed discussion of fitness scores in Appendix A. Fitness scores are assigned to $T = 10\%$ of agents, which are then randomly paired (with replacement) and agents with lower fitness adopt the properties (i, p, q) of their partners. Following Riolo, Cohen, and Axelrod (2001), each agent mutates her new tags and tolerance level with probability $M = 0.1$.[4]

Key Model Steps

Each model run consists of the following steps:

1. *Experimental Condition*: Set initial parameter values (e.g. loyalty expectations and allegiance perceptions).
2. *Simulate* dynamics of loyalty trials repeatedly:
   I. *Loyalty Trials*: Agent B interacts with some other agent A, who is labeled a defector (conformer) if her behavior is perceived as more (less) deviant from loyalty expectations than B can tolerate. Agents are (not) guilty of defection if their private allegiance is (not) violating loyalty expectations of political authorities. Crossing labels with their veracity yields the defector types from Table 1.
   II. *Allegiance Shifts*: The public perception of A's allegiance is decreased (increased) with every label. Labeled defectors update their private allegiance, such that false defectors tend to increase and true defectors tend to decrease their allegiance.
   III. *Adaptation*: The tolerance of all agents increases (decreases) based on the difference between true and false defectors. Agents are selected for play in the next generation based on fitness—the trade-off between benefits of disloyalty (loyalty), defector repression, and accusations of defection—with agent tags and tolerance subject to random mutation.
3. *Results*: Analyze defector type prevalence and group conformity across experimental conditions.

Model Results

General Model

We begin by discussing how group conformity changes in response to loyalty expectations and behavior in the context of our model. The general relationship is depicted in Figure 1, with each cell representing a different experimental setting. We note that small changes in initial behavior (x-axis) and loyalty expectations (y-axis) can lead to significant changes in group conformity (given by the coloring of heatmap cells).

Group conformity decreases with increasing loyalty expectations, and endogenous allegiance shifts are most likely when agents are borderline conforming. Assuming that agents are initially as loyal as perceived, the model produces two straightforward equilibria: conformity for $\bar{i}, \bar{p} \gg \lambda$, and conversely, defection for $\bar{i}, \bar{p} \ll \lambda$.

Figure 2 depicts two typical patterns of *allegiance shifts* by defector types over the course of model runs. In panel (A), loyalty is rewarded and conformity increases as defection decreases. Labeled defectors are rewarded for increasing their loyalty more than secret defectors are for disloyalty, and tolerance for defection increases as 'Type I' outweigh 'Type II' errors. Ultimately, loyalty trials subside with increasing tolerance and group conformity. Conversely in panel (B), disloyalty is rewarded and defection increases as conformity decreases. Opportunities for secret defection outweigh the benefits of loyalty in response to labeling, and 'Type II' outweigh 'Type I' errors. As in Granovetter models of political protest (Granovetter 1978; Kuran 1989), cascades of true defection ensue.[5]

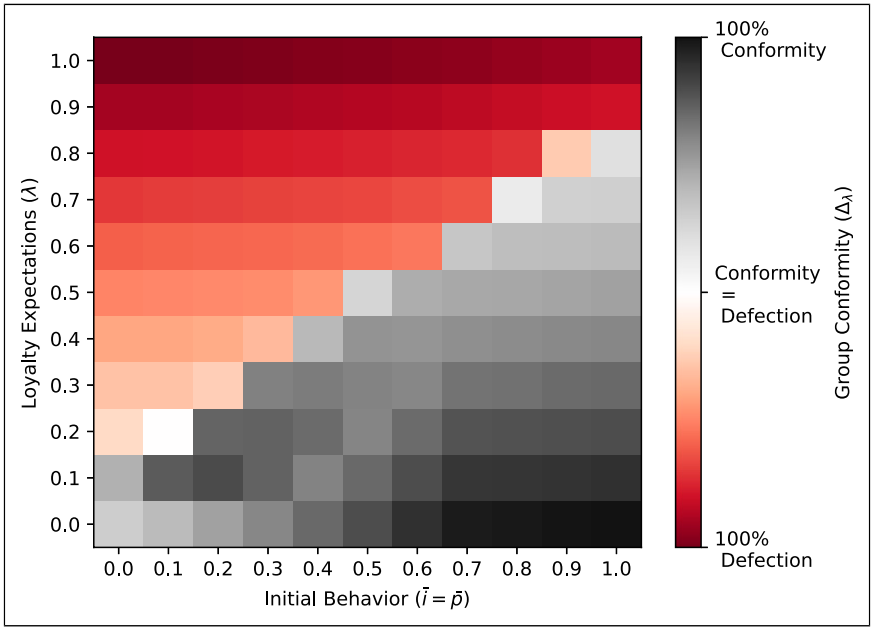

Figure 1. General model: Group conformity.

Note. Each cell depicts group conformity for an experimental condition, averaged across simulations, based on initial loyalty expectations (λ) and allegiance ($\bar{i}, \bar{p}$). For the general model, we assume that loyalty incentives correspond to expectations ($k = \lambda$), though we relax this assumption in applying the model to our empirical cases. Each experimental condition is simulated $S = 30$ times, and each simulation lasts $G = 100$ generations. Simulations are seeded with $N = 1000$ agents, $M = 0.1$ probability of agent mutation, $P = 3$ agent pairings per generation, $T = 10\%$ proportion of agents updating per generation, σ_i, σ_p, $\sigma_q = 0.1$ initial dispersion of agent parameters, and $\bar{q} = 0.1$ initial agent tolerance. Agent parameter values are drawn from the normal distribution. Table 9 and Table 10 in the online appendix provide details on parameters and results.

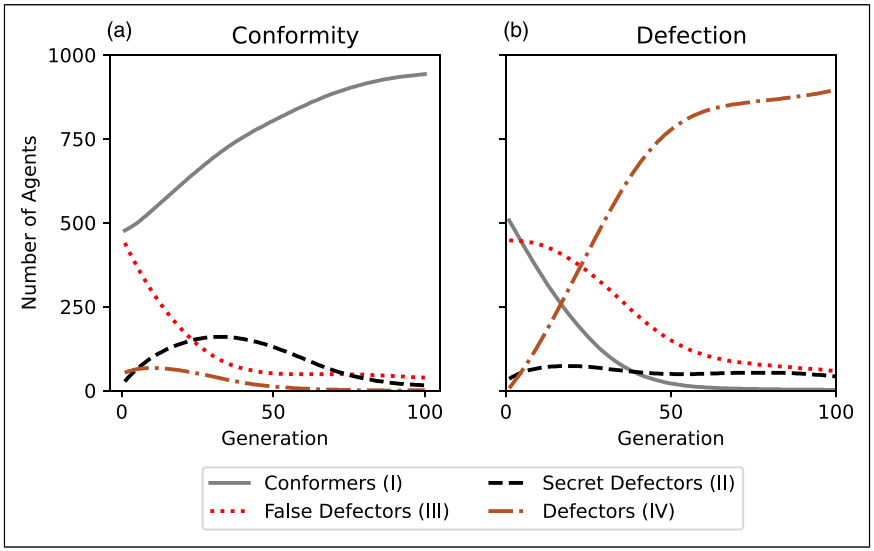

Figure 2. Types of allegiance shifts.

Note. Each panel shows the LOWESS for the prevalence of the four defector types from Table 1 (y-axis), over $G = 100$ generations (x-axis) in a simulation run that is typical for one of the two observed allegiance shifts. (a) Loyalty is rewarded ($k = 1$), (b) Disloyalty is rewarded ($k = -1$). Other parameters are set to a baseline that makes agents equally likely to increase conformity and defection on average: $\lambda = \overline{p} = \overline{i} = 0.5$, σ_i, σ_p, $\sigma_q = 0.1$, $\overline{q} = 0.1$, $G = 100$, $N = 1000$, $M = 0.1$, $p = 3$, $T = 10\%$.

These two patterns, robust to a wide range of auxiliary parameter specifications (see Appendix B), result in highly polarized outcomes (see Montalvo and Reynal-Querol 2005): conformity increases in populations that are rewarded for loyalty (pattern A), whereas defection increases in population rewarded for disloyalty (pattern B). As these conditions are not mutually exclusive, the dynamics of loyalty trials can result in oscillation between conformity ($k = 1$) and defection ($k = -1$).

A key assumption in the general model is that most agents are initially conformers and perceived as such ($\overline{i} = \overline{p}$). Empirically, conflicts deviate from this assumption: defectors may be perceived as conformers and vice-versa. Allegiance outcomes therefore depend on the specific combination of loyalty expectations, defector repression, and the relationship between public and private allegiances, which we turn to in the section below.

Contextualized Model

Conflict in the GDR took the form of state repression until the fall of the Berlin Wall in 1989. Against the backdrop of the Cold War, the SED-led regime feared attempts by the West, the Federal Republic of Germany (FRG) in particular, to undermine its economy and status as an independent state. By contrast, the OPT are characterized by periods of

civil violence, from the first Arab Uprising in 1936 against the British administration to the Second Intifada in 2000 against the Israeli occupation. In this section, we explore how the dynamics of loyalty trials operate in settings characterized by vast contextual differences. Model validation in the GDR is based on existing literature and selected *Stasi* surveillance, purposefully sampled from archives in Berlin. In the OPT, information on loyalty trials was gleaned from existing literature and public databases. To ensure that we interpret this information correctly, we conducted 20 interviews with Israeli and Palestinian experts selected for their familiarity with the topic of collaboration and their ability to transfer knowledge at minimal personal risk (see Appendix C).[6]

Model contextualization comes with numerous challenges: it requires ontological assumptions about the reference "group", the situational context in which defector labels are applied to presumed members, and the minimal expression of a label that constitutes an accusation of disloyalty. Moreover, defection as construed by political authorities is both rare and challenging to observe: when defection exceeds conformity, social order is likely compromised, failing which defection either remains undetected or is detected and punished. The observational challenge is exacerbated by the relational nature of loyalty: expectations, perceptions and tolerance for disloyal behavior are permanently in flux in conflict settings, and their overt expression is rarely documented. It follows that available estimates of defection are unreliable, given that the requisite data is either classified or unverifiable, with even the most diligent government employees prone to conceal Type I and Type II errors in an effort to justify their activities (see Appendix C.1 for details). Given that the available data provides at best a poor approximation of true defection, with little to no indication of false or secret defection, we rely upon a qualitative, most-different case comparison—a critical engagement with historical sources resulting in an interpretive coding of loyalty trials.

With these caveats in mind, we note that the GDR and OPT differ on at least three dimensions that are not endogenous to the model: the volatility of loyalty expectations (which changed more frequently in the OPT relative to the GDR), the number of political authorities tasked with enforcing loyalty expectations at the country-level (a single authority in the GDR, multiple competing authorities in the OPT), and the share of defectors officially labeled by state agents (GDR security agencies had more official control over defectors relative to those in the OPT). Our operationalization of loyalty in Table 3 reflects these substantive differences between the GDR and the OPT.

Despite these differences, we argue that the mechanisms linking loyalty trials to allegiance outcomes work similarly in both settings, with the caveat that some behaviors construed as disloyal in one case are not in the other, such as selling land to the outgroup or emigration. For the purposes of comparison, we limit our discussion to a single authority expecting the same level of loyalty from all ingroup members, but note that the dynamics of loyalty trials may be applicable to smaller units of analysis with appropriate adjustments to model parameters. We associate an increase in loyalty with

Table 3. Contextualization for GDR & OPT Settings.

		Parameters	
Loyalty Level	Disloyal Behaviors	GDR (1971)	OPT (2000)
Security (0.1 − 0.3)	Plan Revolution		
	Land selling		
	Enemy-informing		
Unity (0.4 − 0.6)	Join ingroup opposition		
	Regime-critical protest		
	Refuse authority support	$\overline{p}$	λ
Well-being (0.7 − 0.9)	Illegal emigration	λ	$\bar{i}$
	Work in rival area	$\bar{i}$	$\overline{p}$
	Personal outgroup contact		

Note. We view each of the listed behaviors as violating a level of loyalty in the given range. Loyalty expectations reflect the minimum personal sacrifice that is expected from all group members by a single political authority. Private and perceived loyalty parameters indicate which types of disloyalty group members would on average *not* commit. Parameter values for N = 1000 representative agents reflect relative differences between the two conflict settings, and are drawn from the normal distribution. **GDR**: λ = 0.7, k = −1, $\bar{i}^{95\%} = 0.8$, $\bar{i}^{4\%} = 0.5$, $\bar{i}^{1\%} = 0.2$, σ_i = 0.05, $p_A = i_A - 0.1$, σ_p = 0.1, $\overline{q}$= 0, σ_q = 0.01. P = 1, T = 0.5%, G = 1800. **OPT**: λ = 0.6, k = 1, $\bar{i}^{59\%} = 0.7$ with $p_A^{59\%} = i_A^{59\%+0.1}$, $i^{20\%}$ = 0.5 with $p_A^{20\%} = i_A^{20\%}+0.5$, $i^{20\%}$ = 1.0 with $p_A^{20\%} = i_A^{20\%}-0.5$, $i^{1\%}$ = 0.2 with $p_A^{1\%} = i_A^{1\%}$, σ_i = 0.1, σ_p = 0.2, $\overline{q}$= 0.1, σ_q = 0.05. P = 3, T = 1%, G = 400. This contextualization represents, rather than pinpoints or predicts, individual behavior.

more quotidian behavior—from defending the group's physical security (e.g. refusing enemy-informing to the outgroup), to maintaining its unity and status (e.g. supporting policies unfavorable to the outgroup) and improving the socio-economic well-being and independence of its members (e.g. employment and taxation benefiting the ingroup).[7]

We assess model outcomes in relation to empirical evidence from our cases. In both settings, most group members are privately conforming with loyalty expectations ($\lambda < \bar{i}$). In the GDR, privately loyal individuals were perceived and labeled as defectors from high loyalty expectations ($\lambda - \overline{q} > \overline{p}$). In the OPT, perceived defection from moderate loyalty expectations was tolerated ($\lambda - \overline{q} < \overline{p}$). In the following section, we justify our choice of parameter values for each case, and discuss the results with reference to evidence from archival materials and existing studies.

GDR Allegiance During the East-West Détente

In the GDR, the Central Committee of the Socialist Unity Party (SED) was the sole political authority to enforce loyalty expectations during the Cold War, with a view towards countering the Federal Republic of Germany (FRG). The Ministry for State Security (*MfS* or *Stasi*) drew on an infamously vast surveillance and reporting system to conduct loyalty trials, which generally took the form of denunciations

followed by interrogations and sometimes mock court trials. Punishments ranged from demotions and party reprimands to imprisonment and (until 1987) death sentences (Raschka 2001). We focus on Erich Honecker's tenure as general secretary of the SED between 1971 and 1989, a period with relatively stable loyalty expectations until authorities acquiesced to mass protests and border-crossings in the fall of 1989 (see Opp 1994).

Parameter Settings

Loyalty Expectations. Demands for unification with the FRG and related resistance to the Soviet-backed SED-regime had been violently repressed during the 1950s (Pollack and Rink 1997, 8; Thomson 2018), as border fortifications and closure of the East-West Berlin crossing in 1961 stemmed the flow of emigration to the West (Passens 2012, 114). To justify its relevance and activities, the MfS and its head Erich Mielke coined the term "political-ideological diversion" to construe social deviance driven by Western aggression as defection (Gieseke 2014, 48-59).

Loyalty Incentives: While a minority of privileged SED cadres received benefits for loyalty (e.g. travel authorizations to non-socialist countries, access to Western currency and products), most East Germans had few loyalty incentives, and there was significant protection of defectors: between 1963 and 1989, the FRG paid the GDR to release a total of 33,755 political prisoners into its territory (Borbe 2010, 21), interpreted as "insurance" by would-be defectors in case of arrest (Raschka 2001, 122).

Private Loyalty: The vast majority of East Germans were borderline conforming with high loyalty expectations (see Pollack 1997, 307-308), notwithstanding the small minority of defectors who 'illegally' emigrated to the West, openly criticized the SED-party regime, or actually provided sensitive information to Western organizations.

Perceived Loyalty: Despite the onset of détente in the late 1960s, perceptions of loyalty did not increase (see Gieseke 2014, 59-65): East-West contact and regime-critical statements by Marxist political circles, artists and church members were perceived as betrayal by state security (Rink 1997), and politicized community organizations treated social deviants who glorified life in the West as defectors (see Budde 2014).

Tolerance: Given the prevalence of informants, as well as the rewards and protection granted to informants (Piotrowska 2020), there was a high chance that perceived defection would be labeled. From schools and work places to neighborhoods for state security personnel (Krähnke et al. 2017), deviant behavior was reported to authorities, followed by investigations, interrogations, and formal sanctions.

Interaction: Citizens in the GDR developed a tendency to withdraw from public life (see Pfaff 2001), suggesting that perceptions of disloyalty spread relatively slowly across the population, particularly by learning of individuals targeted covertly by the MfS through private accounts.

By 1971, GDR authorities had high loyalty expectations, despite widespread conformity. They over-identified defectors, but could not match the incentives for disloyalty provided by their Western rivals.

Results. Following the pattern of *defection* in Figure 2 (b), Figure 3 shows how East German allegiance declines as falsely labeled defection increases tolerance and secret defection, resulting in cascades of true defection. To corroborate this shift, we draw on individual cases from the 'Stasi archives' and statistics compiled by historians.

For **conformers**, the détente was an opportunity to engage in privately beneficial and borderline loyal behavior. Most consequential for behavioral change was the easing of travel restrictions and recognition of sovereignty between East and West, with the 1975 Helsinki Accords signalling a normalization of relations (Gieseke 1999, 539; Raschka 2001, 37-44). As a result, those with family connections to the West submitted emigration requests that rose by 70 percent between 1975 and 1976 (Eisenfeld 1999, 385), pushing an intolerant MfS to over-identify defection (Gieseke 2014, 61; Passens 2012, 167-169).

False defectors placed on trial attempted to to prove their loyalty (e.g. Krähnke et al. 2017, 235-252), although their labeling encouraged more widespread defection. Examples include state employees with Western contacts (e.g. BArch, MfS, HA XVIII, 6320)—in this case, an employee with security clearance at the finance ministry who was put on trial for enemy-informing due to his wife's unreported family contacts in the West—migration request denials that were perceived negatively by colleagues (e.g. BArch, MfS, HA XVIII, 37797), and church officials who complied with the MfS but encouraged activism perceived as disloyalty (e.g. BArch, MfS, BV Potsdam, KD KY, Nr. 75, Bd. 1-3).

Secret defectors resulted from adaptation of labeled behavior, following the regime's increase in tolerance to avoid false labeling (see Gieseke 2014, 134). Examples include the use of ambiguous symbols for activism that did not warrant official trials by the state (see Gieseke 2008, 240; e.g. BArch, MfS, HA IX, Nr. 25283, Bl. 34-36; BArch, MfS, HA IX, Nr. 25609, Bl. 9-127), and the concealing of Western contacts in response to disciplinary measures (e.g. BArch, MfS, HA XVIII, Nr. 28434). But given extensive surveillance, secret defection was unsustainable in the long run.

Defectors were defiant of attempts to treat their behavior as disloyal, and received support from their peers. This included persistent emigration requests after labeling (e.g. BArch, MfS, HA XVIII, Nr. 38403), protests against the denunciations of discontent workers (e.g. Halbrock 2015, 144-145), and overt criticism of the GDR in response to labeling. Take the example of two engineers, tried for requesting visits to family in West Berlin and subsequently denied work-related travel authorization to non-socialist countries—a privilege granted to loyal employees. Labeling motivated the couple to submit an emigration request, which was approved after several attempts at dissuasion in their workplace (BArch, HA XVIII 38403). Defection was exacerbated by protection from the FRG, Western human rights organizations, media, and the

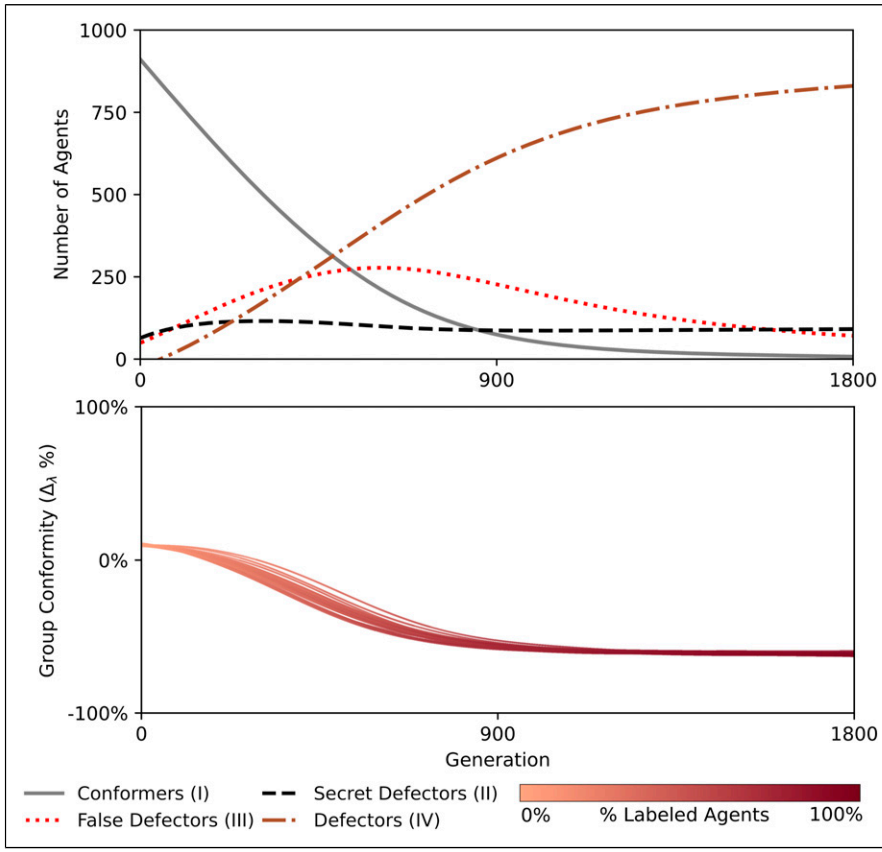

Figure 3. GDR allegiance.
Note. **(Top)** Defector pattern LOWESS over generations, averaged across simulations. **(Bottom)** Group conformity LOWESS for $S = 30$ simulations. Parameter values as listed in Table 3. Counterfactual runs in Appendix C.2 show that results do not change significantly across levels of k: all else equal, even maximum rewards or punishments by GDR authorities could not prevent cascades of defection.

protestant church, which competed with the GDR regime for individual loyalties (Eisenfeld and Eisenfeld 1999, 97-98). More specifically, members of the church were seen as disloyal to the state and disadvantaged, encouraging many to declare allegiance for one or the other, with long-term consequences for secularization in East Germany. (Wohlrab-Sahr, Schmidt-Lux, and Karstein 2008).

OPT Allegiance During the Second Intifada

In the OPT, loyalty expectations were contested by Israel, the Fatah-led Palestinian Authority, Hamas, and affiliated armed organizations (see Pearlman 2011, 150-186).

These expectations varied considerably with the threat posed by the Israeli occupation as well as the control exercised by means of administrative detention, blackmail, and restrictions on movement—all invariably used to recruit informers (Cohen 2010; Sorek 2010; Nerenberg 2016; Berda 2017). For the purpose of comparison, we limit our discussion to a single authority expecting the same level of loyalty from all ingroup members, with arguably less control over the identification of defectors relative to the Stasi (Tartir 2015; B'Tselem 2021b). We focus on the Second Intifada from 2000 until the death of president Yasser Arafat in 2004, a period marked by relatively stable loyalty expectations when Arafat was at the helm (see Tartir 2015). During this time, loyalty trials generally took the form of ad-hoc accusations, followed by instant punishment by armed groups or government military court procedures (Nerenberg 2016, 244-246; Human Rights Watch 2001, 24-27).

Parameter Settings

Loyalty Expectations. The years after the 1993 Oslo Accords had been marked by a normalization of collaboration with Israel. The accords constrained the ruling PA to enforce moderate loyalty expectations in exchange for international support, including provisions to prevent the prosecution of Palestinian collaborators. By the onset of the Second Intifada in 2000, the PA accommodated Israeli pressure to maintain moderate loyalty expectations, though political authorities did not officially expect members to contribute to the well-being of their own group (see Nerenberg 2016, 198).

Loyalty Incentives: The importance that Palestinians attribute to everyday resistance (Ali 2019) and the widespread knowledge of Israeli arrest and recruitment practices (Cohen 2010), contributed to the strong loyalty incentives that allowed Palestinians to resist the occupation, in spite of the high levels of Israeli coercion and material inducements for defection.

Private and Public Allegiances: Most Palestinians privately exceeded moderate loyalty expectations, with perceived exceeding private loyalty. Exceptions included land-dealers and enemy-informants, perceived as defectors who threatened Palestinian security (see Nerenberg 2016, 211). "Non-Statutory" armed groups (*NSAG*) who attacked Israel were deemed defectors from moderate expectations by the PA to preserve its international state- and peace-builder status (see Pearlman 2011, 118-122,154-156; Tartir 2015, 3), while NSAG perceived moderate defection as loyal efforts to resist official PA collaboration (see Nerenberg 2016, 209).

Tolerance: Fluctuations in the enforcement of moderate loyalty expectations suggest that tolerance for such disloyalty was relatively high and heterogenous (see Human Rights Watch 2001; Kelly 2010). In particular, the PA tolerated informants in recognition of informant's status as victims to the Israeli Security Agency (*Shabak*), whereas for some *NSAG* such perceived defection was not tolerable (see Cohen and Dudai 2005).[8]

Interaction: Compared to the GDR, defector suspicions were regularly shared and loyalty trials carried out in public, as Palestinians recognized the implicit nature of

"complicity with the Israeli occupation" in their daily lives. Labeling constituted an expression of fear over being forced into loyalty conflicts by the *Shabak* (Kelly 2010, 183-184).

Overall, the PA expected a moderate level of loyalty from Palestinians, most of whom were either perceived as loyal or whose disloyalty was tolerated. Over-identification occurred where defector perceptions diverged between official PA and unofficial NSAG labeling, and where the fear of forcible recruitment spurred false perceptions of defection (see Table 11 in the Appendices for specific parameter values).

Results. Figure 4 shows how Palestinian allegiance increases as less defectors are labeled, following the pattern of *conformity* in Figure 2(a): falsely labeled defectors have incentives to increase their loyalty and are increasingly tolerated by authorities.

Most **conformers** were not actively involved in the uprising (Pearlman 2011, 163), but the vast majority consistently supported it, and there is evidence of attitudinal shifts towards conformity, as support for political collaboration and personal contacts with Israelis declined by over 10 percent (PSR 2000; 2001; JMCC 1999; 2000).[9] Overt labeling among Palestinians was relatively rare: the allegiance shift was more due to the shared "climate of confrontation" with Israel (Pearlman 2011, 154), and the corresponding social incentives for loyalty.

Defectors were subject to public derogation (e.g. Nerenberg 2016, 237-238; see Abu-Nimer 2011, 97), killings and arrests by NSAG for enemy-informing, or by the PA and Israel mostly for overt mobilization. Between 2000 and 2004, at least 110 Palestinians were put on trial or killed for enemy-informing—most of them during the Israeli incursion of the West Bank in April 2002—24 sentenced to death without sentences carried out, and over 600 detained by the end of 2001 (B'Tselem 2021a; B'Tselem 2021b; Human Rights Watch 2001, 26-27,49-50). In a well-publicized case, the Shin Bet blackmailed a Palestinian into enemy-informing and placed a bomb in his car to kill a senior Hamas member. The accused was vilified by his own legal representatives, sentenced to death after 2 hours of court trial, and summarily executed in public (Williams 2001, 30-32; Human Rights Watch 2001, 45-46; Al-Bitawi 2016, 35).

False defectors had little choice but to repent and demonstrate conformity. Enemy-informants were labeled based on "rumors, suspicions, and popular denunciations" (Human Rights Watch 2001, 23), and their families were stigmatized even when suspicions turned out to be unfounded (e.g. Jalal 2015; Human Rights Watch 2001, 47-48; Williams 2001, 32-36). A PA security chief, for instance, was falsely labeled and demoted for identifying Hamas prisoners to Israel after an attack on his compound, despite reportedly sabotaging the arrest attempt (Yousef 2010, Chapter 22; Kelly 2010, 179).

Secret defectors, whose collaboration with Israel on security issues had previously expected and tolerated by authorities, demonstrated public allegiance to redeem themselves or avoid future labeling (e.g. Cohen 2012, 478-479; Cohen and

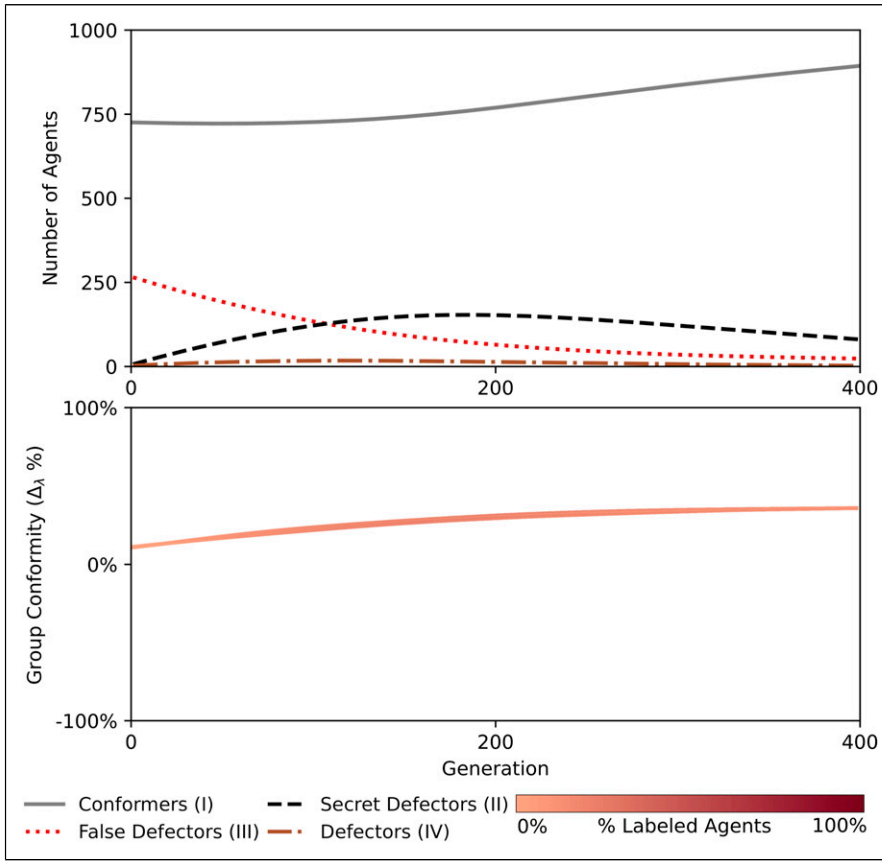

Figure 4. OPT allegiance.
Note. **(Top)** Defector pattern LOWESS over generations, averaged across simulations. **(Bottom)** Group conformity LOWESS for $S = 30$ simulations. Parameter values as listed in Table 3. Note that decreasing loyalty incentives could have led to cascades of defection, but our reading of the case suggests that this was not the case empirically for the OPT overall (see Appendix C.2).

Dudai 2005, 239-240; Pearlman 2011, 154; Berda 2017, 31-32; Kelly 2010, 179), and the few who remained hidden presumably received support from Israel to do so (e.g. Yousef 2010).

Conclusion

Loyalty trials occur across a range of conflict settings characterized by marked differences in regimes, social identification, and the use of selective incentives. Using archival data from the GDR and secondary data from the OPT, our analysis of loyalty trials identifies two polarized outcomes: cascades of defection in the GDR

and a surge of conformity in the OPT. In the GDR, misidentification increased defection, with disloyalty further incentivized by Western organizations externally and by the protestant church internally. In the OPT, by contrast, increased loyalty expectations resulted in greater conformity—loyalty to Palestinian factions that promoted violent resistance. It follows that defection was more likely in the GDR relative to the OPT, given higher expectations and misidentification, and lower incentives for loyalty.

Our analysis of the two cases underscores the measurement problem associated with loyalty trials—the discrepancies between expectations, perceptions and behavior. The case studies also illustrate how both exceptional behavior *and* "quotidian struggles" effectively undermine regime stability (Scott 1985; Wedeen 1999, 87), highlighting the co-production of loyalty by incumbents and rivals alike. In this regard, our framework goes beyond recent scholarship that explains indiscriminate repression with the number or quality of informants (e.g. Steinert 2022, 4-7). Whereas more and better information may reduce the distance between perceived and private loyalties, it also exerts an influence on the tolerance and ability of political actors to label suspects in more subtle ways. Given that conformity and defection are relative to loyalty expectations, with some behaviours construed as disloyal in some contexts but not in others, research on the repression-mobilization nexus would benefit from taking these intricacies into account.

Beyond the particularities of the two cases, our theoretical framework has implications for the study of political order writ large. A key implication concerns the propensity of political actors to over- or under-identify threats to political order, prosecuting innocents (Type I error) or failing to prosecute the guilty (Type II error). Stalin's dictum that every communist was a potential enemy effectively turned Blackstone's ratio (William 1893)—the notion that it is better that ten guilty persons escape than that one innocent suffer—on its head, with implications for some tens of millions of innocent Russians who were killed (Baberowski 2012, 161-172). By contrast, under Communist rule in the GDR, some tens of thousands of delinquent youth and political activists were falsely accused of disloyalty (see Appendix C.1.2), yet a far greater number of 'disloyal' East Germans likely evaded loyalty trials.

We conclude by noting that while treason most commonly ranks among the crimes considered 'worthy' of capital punishment (Thiranagama and Kelly 2010, 1-2), loyalty trials rarely assume center-stage in studies of social conflict. Noteworthy, in this regard, is that loyalty expectations persist well beyond their original manifestations, with attendant implications for 'ethnic defection" (Kalyvas 2008), social trust and cohesion. Former collaborators with the GDR regime are considered untrustworthy some 30 years after unification with West Germany (Zeit 2019), and the PA's history of collaboration with Israel continues to undermine its legitimacy (Tartir 2019). It follows that the interplay between expectations, perceptions, and behavior, as well as the associated perils of over- or under-estimating defection, have appreciable consequences for intra-group polarization and conflict.

Acknowledgments

We thank various research partners in the Occupied Palestinian Territories for sharing their insights on this topic, as well as Christian Carlsen and Friedrich Rother for their help with the archival materials. We also thank Janine Bressmer, Juliette Ganne, Ellen Lust, Laura Nowzohour, Christiana Parriera, Sungmin Rho, Alessandra Romani, David Sylvan, and discussants at ISA 2019, EPSA 2019, SPSA 2020; CYBIS 2020 for helpful comments and suggestions. All remaining errors are our own.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was supported by SNF research grant 188287.

ORCID iDs

Mirko Reul https://orcid.org/0000-0003-2306-0009
Ravi Bhavnani https://orcid.org/0000-0001-6501-5682

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study. The code to reproduce model results is available via the following link: https://github.com/mirkoreul/allegiance-abm.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Kuran (1989) considers that quotidian behavior may be (socially) repressed—arguing that preference falsification generated under the threat of ostracism explains unexpected shifts in popular support for regimes—yet does not account for varying tolerance of defection by political actors, either.
2. The number of interactions and percentage of affected agents per generation is chosen arbitrarily here, as it merely affects the time it takes for the model to converge on a given outcome without affecting the outcome itself (see Appendix B.2).
3. The probability of being selected under $k = 1$ is maximized with $p_A = 0$ and minimized with $p_A = 1$ (perceptibly disloyal agents are tried), while under $k = -1$ it is maximized with $p_A = 1$ and minimized with $p_A = 0$ (perceptibly loyal agents are tried).

4. Mutation adds Gaussian noise of 0, standard deviation 0.01 to tolerance, and draws random private and public tags from the initial distribution with $i_A \sim \mathcal{N}(\mu = \bar{i}, \sigma = \sigma_i), p_A \sim \mathcal{N}(\mu = \bar{p}, \sigma = \sigma_p)$. The chance to mutate is independent for each tag and tolerance.
5. Our model falls short of identifying the 'endpoint' of such cascades, for instance due to conflict termination or regime change. We interpret such endpoints as an exogenous change in loyalty expectations.
6. Collaboration with Israel is a sensitive topic in the OPT, which is why we only contacted key informants who had previously spoken about the subject in public, and only spoke to selected Palestinian authority figures after receiving assurances from such key informants that they are comfortable with discussing it. We therefore did not conduct interviews with Israelis or Palestinians who might be at risk of being identified as enemy-collaborators, were in the past suspected of disloyalty, or in any other way could have been at risk of re-traumatization or reprisals as a result of being interviewed. As an additional safeguard, interviews were not recorded, notes anonymized, and stored exclusively on encrypted drives. Research ethics and the data management plan were approved as part of the project evaluation for funding.
7. Levels of loyalty are treated as transitive, such that lower levels of loyalty imply disloyalty at higher levels. By the same token, higher levels of loyalty expectations encompass lower levels.
8. Informants were unofficially tolerated particularly by Fatah, unless informing led to assassinations (Abdel-Jawad 2001). In those cases, authorities labeled enemy-informants even knowing that defection was coerced by Israeli intelligence, as failure to do so would in turn lead to accusations of betrayal against authorities (see Nerenberg 2016, 210-211).
9. The increase in support for Hamas and Fatah following violent attacks on Israeli targets (Jaeger et al. 2015) is in line with this surge in conformity.

References

Abdel-Jawad, S. 2001. *The Israeli Assassination Policy in the Aqsa Intifada*. Jerusalem: Jerusalem Media & Communication Centre.

Abu-Nimer, M. 2011. "Religious Leaders in the Israeli-Palestinian Conflict: From Violent Incitement to Nonviolent Resistance." In *Nonviolent Resistance in the Second Intifada: Activism and Advocacy*, edited by M. C. Hallward and J. M. Norman, 87-109. Palgrave Macmillan.

Agamben, G. 2005. *State of Exception*. Chicago: University of Chicago Press.

Åkerström, M. 1991. *Betrayal and Betrayers: The Sociology of Treachery*. New Brunswick: Transaction Publishers.

Al-Bitawi, A. H. 2016. *Palestinian Agents and Spies: Israel's Third Eye*. Beirut: Al-Zaytouna Centre for Studies & Consultations.

Albrecht, H., and D. Ohl. 2016. "Exit, Resistance, Loyalty: Military Behavior During Unrest in Authoritarian Regimes." *Perspectives on Politics* 14 (1): 38-52.

Ali, N. 2019. "Active and Transformative Sumud Among Palestinian Activists in Israel." In *Palestine and Rule of Power: Local Dissent vs. International Governance*, edited by A. Tartir and T. Seidel, 71-103. Palgrave Macmillan.

Arjona, A. 2016. *Rebelocracy: Social Order in the Colombian Civil War*. Cambridge: Cambridge University Press.

Baberowski, J. 2012. *Verbrannte Erde: Stalins Herrschaft Der Gewalt*. München: C.H. Beck.

Balcells, L. 2010. "Rivalry and Revenge: Violence against Civilians in Conventional Civil Wars." *International Studies Quarterly* 54 (2): 291-313.

BArch. MfS, BV Potsdam, KD KY, Nr. 75.

BArch. MfS, HA IX, Nr. 25283.

BArch. MfS, HA IX, Nr. 25609.

BArch. MfS, HA XVIII, Nr. 28434.

BArch. MfS, HA XVIII, Nr. 37797.

BArch. MfS, HA XVIII, Nr. 38403.

BArch. MfS, HA XVIII, Nr. 6320.

Becker, H. S. 1963. *Outsiders*. New York and London: Free Press.

Berda, Y. 2017. *Living Emergency: Israel's Permit Regime in the Occupied West Bank*. Stanford: Stanford University Press.

Bergemann, P. 2017. "Denunciation and Social Control." *American Sociological Review* 82 (2): 384-406.

Bhavnani, R., K. Donnay, D. Miodownik, M. Mor, and D. Helbing. 2014. "Group Segregation and Urban Violence." *American Journal of Political Science* 58 (1): 226-245.

Bhavnani, R., K. Donnay, and M. Reul. 2020. "Evidence-Driven Computational Modeling." In *Handbook of Research Methods in Political Science & International Relations*, edited by L. Curini and R. J. Franzese. Sage Publications.

Borbe, A. 2010. *Die Zahl Der Opfer Des SED-Regimes*. Erfurt: Landeszentrale für politische Bildung Thüringen.

B'Tselem. 2021a, July. "Database on Fatalities." Technical report.

B'Tselem. 2021b, September. "Statistics on the Death Penalty in the Palestinian Authority and Under Hamas Control in Gaza." Technical report.

Budde, H. 2014. "Politische Fremdbestimmung durch Gruppen: Stabilisator des SED-Staates." In *Enquete-Kommission Zur Aufarbeitung von Geschichte Und Folgen Der SED-Diktatur in Deutschland*, Vol. 4, 285-291. Deutschland Archiv.

Chassany, A.-S. 2017, October. "France: The Permanent State of Emergency." Financial Times.

Cohen, H. 2010. "The Matrix of Surveillance in Times of National Conflict." In *Surveillance and Control in Israel/Palestine: Population, Territory and Power*, edited by E. Zureik, D. Lyon, and Y. Abu-Laban, 99-112. Routledge.

Cohen, H. 2012. "Society–Military Relations in a State-in-the-Making: Palestinian Security Agencies and the 'Treason Discourse' in the Second Intifada." *Armed Forces & Society* 38 (3): 463-485.

Cohen, H., and R. Dudai. 2005. "Human Rights Dilemmas in Using Informers to Combat Terrorism: The Israeli-Palestinian Case." *Terrorism and Political Violence* 17 (1-2): 229-243.

Coser, L. 1956. *The Functions of Social Conflict*. New York: Free Press.

Davenport, C. 2005. "Understanding Covert Repressive Action: The Case of the U.S. Government against the Republic of New Africa." *Journal of Conflict Resolution* 49 (1): 120-140.

Dragu, T. 2017. "The Moral Hazard of Terrorism Prevention." *The Journal of Politics* 79 (1): 223-236.

Dragu, T., and A. Przeworski. 2019. "Preventive Repression: Two Types of Moral Hazard." *American Political Science Review* 113 (1): 77-87.

Eisenfeld, B. 1999. "Flucht und Ausreise - Macht und Ohnmacht." In *Opposition in Der DDR von Den 70er Jahren Bis Zum Zusammenbruch Der SED-Herrschaft*, edited by E. Kuhrt, 381-424. Wiesbaden: VS Verlag für Sozialwissenschaften.

Eisenfeld, B., and P. Eisenfeld. 1999. "Widerständiges Verhalten in der DDR 1976-1982." In *Opposition in Der DDR von Den 70er Jahren Bis Zum Zusammenbruch Der SED-Herrschaft*, edited by E. Kurt, 83-121. Wiesbaden: VS Verlag für Sozialwissenschaften.

Epstein, J. M. 2002. "Modeling Civil Violence: An Agent-Based Computational Approach". *Proceedings of the National Academy of Sciences* 99 (3): 7243–7250.

Gieseke, J. 1999. "Abweichendes Verhalten in der totalen Institution: Delinquenz und Disziplinierung der hauptamtlichen MfS-Mitarbeiter in der Ära Honecker." In *Justiz Im Dienste Der Parteiherrschaft: Rechtspraxis Und Staatssicherheit in Der DDR*, edited by R. Engelmann and C. Vollnhals, 531-553. Berlin: Links.

Gieseke, J. 2014. *[2001]. The History of the Stasi: East Germany's Secret Police, 1945-1990*. New York: Berghahn Books.

Gieseke, J. 2008. "Bevölkerungsstimmungen in der geschlossenen Gesellschaft. MfS-Berichte an die DDR-Führung in den 1960er-und 1970er-Jahren." *Zeithistorische Forschungen/ Studies in Contemporary History* 5: 236-257.

Gilbert, N. 2008. *Agent-Based Models*. Thousand Oaks: Sage Publications.

Granovetter, M. 1978. "Threshold Models of Collective Behavior." *American Journal of Sociology* 83 (6): 1420-1443.

Gwladys, F., M. Nichols, C. Greenfield, M. Bendeich, and M. Collett-White 2021. "Taliban Are Rounding up Afghans on Blacklist - Private Intel Report". Reuters.

Halbrock, C. 2015. "Denunziation, Meldetätitgkeit und Informationserhebung im Kapillarsystem der SED-Diktatur." In *Hinter Vorgehaltener Hand: Studien Zur Historischen Denunziationsforschung*, Vol. 39, 137-152. Göttingen: Vandenhoeck & Ruprecht.

Hechter, M. 1987. *Principles of Group Solidarity.* Berkeley: University of California Press.

Heckathorn, D. D. 1988, November. "Collective Sanctions and the Creation of Prisoner's Dilemma Norms." *American Journal of Sociology* 94 (3), 535-562.

Hirschman, A. O. 1970. *Exit, Voice and Loyalty.* Cambridge: Cambridge University Press.

Hirschman, A. O. 1993. "Exit, Voice, and the Fate of the German Democratic Republic: An Essay in Conceptual History." *World Politics* 45 (2): 173-202.

Holland, J. 1995. *Hidden Order*. New York, NY: Addison Wesley.

Human Rights Watch. 2001. "Justice Undermined: Balancing Security and Human Rights in the Palestinian Justice System." *Technical Report* 13 (4). (E).

Jaeger, D., E. Klor, S. Miaari, and D. Paserman. 2015. "Can Militants Use Violence to Win Public Support? Evidence from the Second Intifada." *Journal of Conflict Resolution* 59 (3): 528-549.

Jalal, R. A. 2015, February. "Spies' Families Marginalized in Gaza." *Al-Monitor.*

JMCC. 1999, February. "Public Opinion Poll No. 30: On Palestinian - Israeli Peace Index." https://www.jmcc.org/polls.aspx

JMCC. 2000, December. "Public Opinion Poll No. 39, Part Two: Attitudes of the Israeli and Palestinian Publics towards the Peace Process." https://www.jmcc.org/polls.aspx

Kalyvas, S. 2006. *The Logic of Violence in Civil War*. Cambridge and New York: Cambridge University Press.

Kalyvas, S. 2008. "Ethnic Defection in Civil War." *Comparative Political Studies* 41 (8): 1043-1068.

Kalyvas, S. 2012. "Micro-Level Studies of Violence in Civil War: Refining and Extending the Control-Collaboration Model." *Terrorism and Political Violence* 24 (4): 658-668.

Kao, K., and M. R. Revkin. 2022. "Retribution or Reconciliation? Post-Conflict Attitudes Toward Enemy Collaborators." *American Journal of Political Science* 67 (2): 358-373.

Kelly, T. 2010. "In a Treacherous State: The Fear of Collaboration Among West Bank Palestinians." In *Traitors: Suspicion, Intimacy, and the Ethics of State-Building*, edited by S. Thiranagama and T. Kelly, 169-187. Philadelphia: University of Pennsylvania Press.

Koehler, K., D. Ohl, and H. Albrecht. 2016. "From Disaffection to Desertion: How Networks Facilitate Military Insubordination in Civil Conflict." *Comparative Politics* 48 (4): 439-457.

Kollman, K., J. H. Miller, and S. E. Page. 1998. "Political Parties and Electoral Landscapes." *British Journal of Political Science* 28 (1): 139-158.

Krähnke, U., M. Finster, A. Zschirpe, and P. Reimann. 2017. *Im Dienst Der Staatssicherheit: Eine Soziologische Studie Über Die Hauptamtlichen Mitarbeiter Des DDR-Geheimdienstes*. Frankfurt: Campus Verlag.

Kuran, T. 1989. "Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution." *Public Choice* 61 (1): 41-74.

Levine, J. M., and R. L. Moreland. 2002. "Group Reactions to Loyalty and Disloyalty." In *Group Cohesion, Trust and Solidarity*, edited by S. R. Thye and E. J. Lawler, 203-228. Emerald Group Publishing Limited.

Lichbach, M. I. 1987. "Deterrence or Escalation? The Puzzle of Aggregate Studies of Repression and Dissent." *Journal of Conflict Resolution* 31 (2): 266-297.

Lyall, J., Y. Shiraito, and K. Imai. 2015. "Coethnic Bias and Wartime Informing." *The Journal of Politics* 77 (3): 833-848.

Marques, J. M., D. Abrams, D. Paez, and C. Martinez-Taboada. 1998. "The Role of Categorization and In-Group Norms in Judgments of Groups and Their Members." *Journal uf Personality and Social Psychology* 75 (4): 976-988.

McLauchlin, T. 2010. "Loyalty Strategies and Military Defection in Rebellion." *Comparative Politics* 42 (3): 333-350.

McLauchlin, T., and Á. L. Parra-Pérez. 2018. "Disloyalty and Logics of Fratricide in Civil War: Executions of Officers in Republican Spain, 1936-1939." *Comparative Political Studies* 52 (7): 1-31.

Montalvo, J., and M. Reynal-Querol. 2005. "Ethnic Polarization, Potential Conflict, and Civil Wars." *American Economic Review* 95 (3): 796-816.

Mueller, J., and M. Stewart. 2012. "The Terrorism Delusion: America's Overwrought Response to September 11." *International Security* 37 (1): 81-110.

Nerenberg, D. 2016. *Cooperating with the Enemy: Purpose-Driven Boundary Maintenance in Palestine, 1967-2016*. Ph. D. thesis. Washington, D.C: George Washington University.

O'Brian, J. L. 1948. "Loyalty Tests and Guilt by Association." *Bulletin of the Atomic Scientists* 4 (6): 166-172.

Oliver, P., G. Marwell, and R. Teixeira. 1985. "A Theory of the Critical Mass. I. Interdependence, Group Heterogeneity, and the Production of Collective Action." *American Journal of Sociology* 91 (3): 522-556.

Opp, K.-D. 1994. "Repression and Revolutionary Action: East Germany in 1989." *Rationality and Society* 6 (1): 101-138.

Oppenheim, B., A. Steele, J. F. Vargas, and M. Weintraub. 2015. "True Believers, Deserters, and Traitors: Who Leaves Insurgent Groups and Why." *Journal of Conflict Resolution* 59 (5): 794-823.

Passens, K. 2012. *MfS-Untersuchungshaft: Funktionen Und Entwicklung von 1971 Bis 1989*. Berlin: Lukas Verlag.

Pearlman, W. 2011. *Violence, Nonviolence, and the Palestinian National Movement*. Cambridge and New York: Cambridge University Press.

Pearlman, W., and K. G. Cunningham. 2012. "Nonstate Actors, Fragmentation, and Conflict Processes." *Journal of Conflict Resolution* 56 (1): 3-15.

Pfaff, S. 2001. "The Limits of Coercive Surveillance: Social and Penal Control in the German Democratic Republic." *Punishment & Society* 3 (3): 381-407.

Piotrowska, B. M. 2020. "The Price of Collaboration: How Authoritarian States Retain Control." *Comparative Political Studies* 53 (13): 1-27.

Pollack, D. 1997. "Bedingungen der Möglichkeit politischen Protestes in der DDR: Der Volksaufstand von 1953 und die Massendemonstrationen 1989 im Vergleich." In *Zwischen Verweigerung Und Opposition: Politischer Protest in Der DDR 1970-1989*, edited by D. Pollack and D. Rink, 303-331. Frankfurt and New York: Campus Verlag.

Pollack, D., and D. Rink. 1997. "Einleitung." In *Zwischen Verweigerung Und Opposition: Politischer Protest in Der DDR 1970-1989*, edited by D. Pollack and D. Rink, 7-29. Frankfurt and New York: Campus Verlag.

Polo, S. M. T., and J. Wucherpfennig. 2022. "Trojan Horse, Copycat, or Scapegoat? Unpacking the Refugees-Terrorism Nexus." *The Journal of Politics* 84 (1): 33-49.

Poulsen, L. N. S. 2020. "Loyalty in World Politics." *European Journal of International Relations* 26 (4): 1156-1177.

PSR. 2000. Public Opinion Poll #1: 27-29 July 2000. Technical report, Palestinian Center for Policy and Survey Research.

PSR. 2001. Public Opinion Poll #2: 5-9 July 2001. Technical report, Palestinian Center for Policy and Survey Research.

Raschka, J. 2001. *Zwischen Überwachung Und Repression: Politische Verfolgung in Der DDR 1971 Bis 1989*. Wiesbaden: Springer-Verlag.

Raybeck, D. 1991. "Deviance, Labelling Theory and the Concept of Scale." *Anthropologica* 33 (1/2): 17-36.

Rink, D. 1997. "Ausreiser, Kirchengruppen, Kulturopposition und Reformer: Zu Differenzen und Gemeinsamkeiten in Opposition und Widerstand in der DDR in den 70er und 80er

Jahren." In *Zwischen Verweigerung Und Opposition: Politischer Protest in Der DDR 1970-1989*, edited by D. Pollack and D. Rink, 54-77. Frankfurt and New York: Campus Verlag.

Riolo, R., M. Cohen, and R. Axelrod. 2001. "Evolution of Cooperation without Reciprocity." *Nature* 414 (6862): 441-443.

Roberts, S. R. 2018. "The Biopolitics of China's 'War on Terror' and the Exclusion of the Uyghurs." *Critical Asian Studies* 50 2): 232-258.

Santoro, W. A., and M. Azab. 2015. "Arab American Protest in the Terror Decade: Macro- and Micro-level Response to Post-9/11 Repression." *Social Problems* 62 (2): 219-240.

Schelling, T. 1969. "Models of Segregation." *American Economic Review* 59 (2): 488-493.

Schlichte, K., and U. Schneckener. 2015. "Armed Groups and the Politics of Legitimacy." *Civil Wars* 17 (4): 409-424.

Schutte, S. 2017, September. "Violence and Civilian Loyalties: Evidence from Afghanistan." *Journal of Conflict Resolution* 61 (8): 1595-1625.

Scott, J. C. 1985. *Weapons of the Weak: Everyday Forms of Peasant Resistance*. New Haven and London: Yale University Press.

Sherman, L. W. 1993. "Defiance, Deterrence, and Irrelevance: A Theory of the Criminal Sanction." *Journal of Research in Crime and Delinquency* 30 (4): 445-473.

Sinno, A. H. 2008. *Organizations at War in Afghanistan and beyond*. Ithaca: Cornell University Press.

Sorek, T. 2010. "The Changing Patterns of Disciplining Palestinian National Memory in Israel." In *Surveillance and Control in Israel/Palestine: Population, Territory and Power*, edited by E. Zureik, D. Lyon, and Y. Abu-Laban, 113-129. Routledge.

Staniland, P. 2012. "Between a Rock and a Hard Place: Insurgent Fratricide, Ethnic Defection, and the Rise of Pro-state Paramilitaries." *Journal of Conflict Resolution* 56 (1): 16-40.

Steinert, C. V. 2022. "The Impact of Domestic Surveillance on Political Imprisonment: Evidence from the German Democratic Republic." *Journal of Conflict Resolution* 67 (1): 38–65.

Sullivan, C. 2016a. "Political Repression and the Destruction of Dissident Organizations." *World Politics* 68 (04): 645-676.

Sullivan, C. 2016b. "Undermining Resistance: Mobilization, Repression, and the Enforcement of Political Order." *Journal of Conflict Resolution* 60 (7): 1163-1190.

Sykes, G. M., and D. Matza. 1957. "Techniques of Neutralization: A Theory of Delinquency." *American Sociological Review* 22 (6): 664.

Tarrow, S. 1994. *Power in Movement: Social Movements and Contentious Politics*. Cambridge and New York: Cambridge University Press.

Tartir, A. 2015. "The Evolution and Reform of Palestinian Security Forces 1993–2013." *Stability: International Journal of Security & Development* 4 (1): 1-20.

Tartir, A. 2019. "Criminalizing Resistance: Security Sector Reform and Palestinian Authoritarianism." In *Palestine and Rule of Power: Local Dissent vs. International Governance*, edited by A. Tartir and T. Seidel, 205-226. Palgrave Macmillan.

Thiranagama, S., and T. Kelly. 2010. "Introduction: Specters of Treason." In *Traitors: Suspicion, Intimacy, and the Ethics of State-Building*, edited by S. Thiranagama and T. Kelly, 1-23. Philadelphia: University of Pennsylvania Press.

Thomson, H. 2018. "Grievances, Mobilization, and Mass Opposition to Authoritarian Regimes: A Subnational Analysis of East Germany's 1953 Abbreviated Revolution." *Comparative Political Studies* 51 (12): 1594-1627.

Tilly, C. 2003. *The Politics of Collective Violence*. Cambridge: Cambridge University Press.

Wedeen, L. 1999. *Ambiguities of Domination: Politics, Rhetoric, and Symbols in Contemporary Syria*. Chicago: The University of Chicago Press.

William, B. 1893. *Commentaries on the Laws of England*. Philadelphia: J. B. Lippincott.

Williams, D. 2001. "Collaborators: Recent Cases in the Palestinian Territories." In *The Phenomenon of Collaborators in Palestine*, 29-39. Jerusalem: PASSIA.

Wohlrab-Sahr, M., T. Schmidt-Lux, and U. Karstein. 2008. "Secularization as Conflict." *Social Compass* 55 (2): 127-139.

Yousef, M. H. 2010. *Son of Hamas: A Gripping Account of Terror, Betrayal, Political Intrigue, and Unthinkable Choices*. Carol Stream: Tyndale House Publishers.

Zeit. 2019, May. "Staatssicherheit: Öffentlicher Dienst wird bis 2030 auf Stasi-Tätigkeit überprüft." *Die Zeit*.