THE GRADUATE INSTITUTE GENEVA

INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT

GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

Enemies Within: Labeling Defectors to Rival Authorities

THESIS
submitted at the Graduate Institute
in fulfillment of the requirements of the
PhD in Political Science/International Relations

by

Mirko Reul

Thesis N° 1448

Geneva
2022

Enemies Within: Labeling Defectors to Rival Authorities

INSTITUT DE HAUTES ETUDES INTERNATIONALES ET DU DEVELOPPEMENT
GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

Enemies Within: Labeling Defectors to Rival Authorities

THESIS
submitted at the Graduate Institute
in fulfillment of the requirements of the
PhD in Political Science/International Relations

by

Mirko Reul

Thesis N° 1448

Geneva
2022

Mirko REUL

Sur le préavis de David SYLVAN, professeur émérite à l'Institut et co-directeur de thèse, de Ravinder BHAVNANI, professeur à l'Institut et co-directeur de thèse, de Sung Min RHO, professeure assistante à l'Institut et membre interne du jury, et de Ellen LUST, Professor, Governance and Local Development Institute, Department of Political Science, University of Gothenburg, Sweden et experte externe, la directrice de l'Institut de hautes études internationales et du développement autorise l'impression de la présente thèse sans exprimer par là d'opinion sur son contenu.

Le dépôt officiel du manuscrit, en 5 exemplaires, doit avoir lieu au plus tard le 15 décembre 2022.

Genève, le 15 novembre 2022

Marie-Laure Salles
Directrice

Thèse N° 1448

INSTITUT DE HAUTES ETUDES INTERNATIONALES ET DU DEVELOPPEMENT
GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

Abstract

How do allegiances shift from incumbent rulers to rival authorities? Political actors, from rebel groups to state authorities, take extreme measures to ensure the loyalty of their subjects. Those who are labeled as defectors or 'traitors' may be ostracized, imprisoned, tortured or killed. But the expected consequences of such punishments remain disputed by an established body of scholarship on state repression, civil wars, and criminal behavior. Either the labeling of people as defectors deters undesirable behavior, leading to widespread conformity with rules set by authorities. Or it intensifies defection from political orders, as the labeled shift support to rival authorities who support their behavior. This project relies on a mixed-methods approach to investigate popular allegiance in the former German Democratic Republic (GDR) and the Occupied Palestinian Territories (OPT), drawing on archival research, a lab experiment, semi-structured interviews, and a computational model. Overall, groups in conflict construe quotidian behavior as indicative for exceptional disloyalty. Labeling individuals as defectors questions their group membership, pushing the labeled and their peers to change behavior in order to either prove their loyalty or intensify their defection. The project contributes to our understanding of conflict by viewing the stability of political order through the quotidian behaviors of 'ordinary' individuals, and the relationship between international rivalries and domestic repression. And it introduces a novel perspective on the defiance of authoritarian rule in the GDR, as well as on 'collaboration' with Israel in Palestine today.

Acknowledgements

This project would not have been possible without the advice and support of many friends, research partners and colleagues. I am especially grateful to Mai Albzour, who helped me understand key issues surrounding Palestinian collaboration with Israel, introduced me to experts on the subject, and over the years proved herself to be an invaluable discussion partner and friend who never tired of answering my many questions. Relatedly, I want to thank Ahmed, Basil, and Nadeen for open discussions about their personal experiences and research, as well as the anonymous research partners in Israel and Palestine who shared their personal and at times highly sensitive experiences and insights for this project. My research on the former German Democratic Republic was made possible by archivists and historians at the *Stasi Archives*. I especially thank Christian Carlsen and Friedrich Rother for their open-minded support of my unconventional approach, explanations of archival materials, and countless hours spent on censoring sensitive information. The third paper in this manuscript would not have been possible without Noah Bacine, Tom Batistoni, and their colleagues at the CESS Oxford, who implemented the lab experiment under a lot of time pressure. I especially thank Noah Bacine, who through numerous discussions ensured that the design is relatable to existing literature, and who patiently taught me the practicalities of implementing experiments in the lab.

Many of the ideas in this dissertation only came to fruition thanks to the constant stream of support from my supervisors and colleagues at the Graduate Institute Geneva. I especially thank my co-supervisors, David Sylvan and Ravi Bhavnani, whose support and trust enabled me to pursue this project, and whose encouragement to go deeper in opposite directions made this dissertation interesting. I am equally grateful to Ellen and Sungmin for their excellent feedback on a complicated manuscript that they had not seen before, and for their thoughtful comments during my defense. I also thank Karsten Donnay for his critical teachings in model design and implementation, as well as Paul Huth, Thomas Gidney, and Matthew Bamber for their support in obtaining the research grant that funded this project.

It took a village of people to get me to start this PhD and to live through the experience. First among them is Julia, without who I would never have arrived at this PhD, and who was an incredible source of comfort, happiness and inspiration. I am also grateful for the friendships that made life at the Institute enjoyable and often bearable, and especially to my friends in Geneva who helped me escape the web of frustrations that make administering and writing a PhD difficult. I am endlessly grateful to the Unicorns, Alim, Eliza, Janine, Jonathan, Juliette, and Paroma: your company made this journey such an exciting race to a misty chalet that I could forget this project and learn something valuable. Lastly, I thank my family for making this ending possible from the beginning, and especially you *achii* for the countless hours spent in alternate realities.

Contents

List of Figures

List of Tables

PREFACE

Popular allegiance is a salient political issue whenever societies feel existentially threatened. In the United States for instance, accusations of disloyalty against the political left were rampant during the 'Red Scare' following the Russian Revolution of 1917, 'McCarthyism' with the onset of the Cold War, and the subequent 'Counterintelligence Program' originally designed to repress Communism (Cunningham 2004; O'Brian 1948). Similar 'loyalty panics' targeted Socialists in Switzerland during World War II (Zimmermann 2015), Jews and Muslims during the Spanish Inquisitions (Bergemann 2017), 'counter-revolutionaries' during *la Terreur* in France (Lucas 1996), and the prosperous peasants and bourgeoisie in the Bolshevik Soviet Union (Fitzpatrick 1996). There are vast differences between these settings, yet in each case, minorities found their loyalty questioned, and though their labeling appeared justified to authorities and group members alike at the time, far more people were denounced guilty by association than were actually threatening the group.

The unjust and often deadly outcomes of such panics continues to inspire public condemnation and critical scholarship (e.g. Berda 2017; Dragu 2017; Mueller and Stewart 2012; Roberts 2018; Stieglitz 2001). Akin to the 'moral panics' described by Cohen (2011), they follow a seemingly inevitable pattern of public anxiety over political deviance, an over-estimation of the threat, and an escalating use of repressive instruments to address it. All the while, polarization ensues to the extent that some political authorities appease an anxious body politic while others are vilified for defending the disloyal. And when defectors are put on 'loyalty trials' by one group, they are welcomed by their rivals: during the Cold War while Communists were surveilled and tried for defection in the West, so were Western sympathizers in the Eastern block, and political authorities on each side provided people with incentives to defect from the other.

The three papers in this dissertation are motivated by what to me appears as a contradiction in political projects that seek to protect communities from harm, and cause harm to the communities they seek to protect. Abuses of state power in these projects are as apparent to distant observers as they are deemed necessary by the decision-makers and moral entrepreneurs who apply them; and the ensuing human rights violations are horrendous for victims yet invisible or acceptable to perpetrators at the time. And since these conflicting perceptions occur across different international and political orders, I do not seek to explain the variable structural and institutional determinants that make loyalty panics more or less likely, but to understand their emergence and consequences across these conditions. My approach therefore grants less importance to particular structures and political institutions, and more to social organizations and prototypical group members that shape political behavior. The overarching goal is to arrive at a "theory of the middle range" that improves our understanding of popular allegiance across diverse settings (Merton 1968).

The scope of the the theory is at least limited to *social conflict* that is grounded in hostility between groups (Simmel 1955, 37), where each group feels threatened by members who perceptibly endanger it to the benefit of the opposing rival. At minimum, there must be attempts to enforce general rules on all ingroup members, the violation of which threatens the well-being of the entire group, and instills a belief that violators may be acting on behalf of an outgroup. Thus while I draw on studies of deviance

from social norms, I am specifically concerned with deviance that threatens *political* projects. That is because allegiance relates to particular norms around behavior and outgroup affiliation that are salient only when a group is perceptibly threatened. Neither the rival group nor the threat it poses have to 'exist' in any sense of the term, but it has to be 'real' in the imagination of ingroup members. At its core, the theory draws on the relationship between *labeling* and *loyalty*. The two concepts are not usually combined: labeling is mainly of interest to sociologists who seek to understand how societies construe behavior as socially deviant, and to criminologists who study how labeling individuals as deviant makes them more or less likely to commit crimes in the future. Loyalty is arguably one of those constructs that is more often casually invoked than explicitly researched in social sciences, with the most seminal exception being the study of 'Exit, Voice and Loyalty' by Hirschman (1970). Throughout what follows, I am interested in how people are *labeled* when they are perceived as *disloyal* to a group, and how this in turn shapes adherence in the sense that Hirschman proposed, and in other ways that he did not propose.

To that end, each paper looks at allegiance shifts from a slightly different ontological perspective, and drawing on very different methods and empirical data. The first paper, co-authored with Ravi Bhavnani, introduces the overarching theoretical framework, and specifies linkages between allegiance shifts at the micro- and expectations for loyalty at the macro-level in an agent-based computational model. The model is then empirically contextualized, using primary and secondary data from the former German Democratic Republic during the *détente* and the Occupied Palestinian Territories during the Second *Intifada*. Primary data here refers to my reading of archival materials from the *Stasi* archives for the East German case, and some twenty semi-structured interviews with Palestinians and Israelis that were mostly used to develop a model consistent with real-world dynamics. This papers glosses over some of the nuances at the individual level to make an argument about the larger forces that influence how popular allegiance shifts between political communities. International rivalries between groups shape loyalty expectations within groups, and while political authorities exercise some measure of control over loyalty trials and their consequences, they are constrained by popular perceptions of what constitutes disloyal behavior. Of note, we show that these dynamics are at play in two very different empirical settings, which comes with the usual trade-offs between external and internal validity. This paper is perhaps most relevant to mainstream research on international and domestic conflict, state repression, rebellion and regime change.

The second paper presents a preliminary analysis of archival data on 319 individuals who were surveilled for defection by the *Stasi* in the German Democratic Republic. I explore the determinants of individual allegiance shifts through case studies of suspected enemy-informants, dissidents, economic 'saboteurs' of the economy, and 'illegal' emigrants. This paper takes a 'meso-level' view on political deviance as constructed between official authorities and group members, and a more grounded approach to empirical evidence on defection from groups. At the core of the argument is a more nuanced conceptualization of loyalty as behavior, perception, and imagination. Notably, this paper provides evidence on some of the micro-level mechanisms that are proposed in the first paper, as it discusses in more detail how individuals are not only tried by authorities but label each other for disloyalty, and how reactions to labeling differ even under seemingly similar conditions. Though far from an ethnographic study of loyalty in an active conflict setting, this paper might alleviate some of the skepticism that critical scholars would naturally develop while reading the first paper.

The third paper presents preliminary results from a laboratory experiment, where two groups enter a modified Tullock contest with the possibility to punish other group members and defect from the group. I explore the causal effects of two factors that may influence the effects of labeling disloyalty: (1) communication between group members and (2) unequal abilities to demonstrate loyalty. The study is run with participants from the pool of the Center for Experimental Social Sciences at Oxford University, most of whom are students. This paper addresses a key methodological shortcoming of the

first two, where the labeling of disloyalty is only observed through the accounts of others. For obvious reasons, it is difficult to systematically observe when one person labels another as disloyal, especially in active conflict settings. Just as problematic would be the claim that labeling in the lab is equivalent to labeling in real-world conflicts. For instance, social identification had to be induced with communication, the hostility between groups with a zero-sum competition, and the effects of labeling with monetary punishment. However, to the extent that key aspects of 'loyalty conflicts' are introduced in the lab, micro-level behavior which is not usually observable can be studied systematically, suggesting how labeling *could* plausibly affect loyalty. As such, this paper is most interesting to political scientists and economists interested in the identification of causal effects at the cost of ecological and external validity.

The diverse data construction projects behind each of these papers warrant some discussion. At its outset, the research plan for this project was to devote equal time to archival research on East Germany and interviews in Palestine, and in particular to collect sufficient data in Palestine to write a second comparative paper on the two contexts, rather than focusing on East Germany. While travel restrictions are always a possibility even for white foreigners in Palestine, I had not anticipated that they would occur for a pandemic, shortly after my first fieldwork in November 2019, nor did I expect that I would have to wait over two years before I could re-start it. By the time I returned in February 2022, some of my contacts who had first-hand experience with the highly sensitive topic of Palestinian-Israeli collaboration had understandably moved on to other careers or lost interest in the subject. It is worth noting that Palestinians in particular have good reason to mistrust foreign researchers, seeing as they are repressed in an active conflict on which much has been written that does not match their experiences.[1] The evidence I could gather, even with the help of local research partners, was too little too late for inclusion in this dissertation, but will lead to a separate publication.

In a similar vein, I was unable to use the laboratory until pandemic-related restrictions were lifted, which was particularly problematic because the design was too different from existing studies to be implemented without several pilot sessions. As a result, much of the data will only be collected after the submission of this manuscript, and readers will have to content themselves with results from the pilot sessions that were run under difficult conditions between August and September 2022. To conclude the apologetic remarks, while the *Stasi* archives provide exceptional support to researchers compared to other archives,[2] they are overloaded with requests even outside the context of a pandemic, and it was difficult to get face-time with the documents even when the partial lifting of restrictions permitted access again. As a result, I only obtained most of the requisite data in 2022, and had to postpone my plans for a more comprehensive statistical analysis until such a time as I can systematically code the vast corpus of documents I managed to sample. That said, the first paper would not have been much different with or without the pandemic, and the other two provide a fair amount of evidence on the labeling of political deviance as disloyalty.

Overall, the project identifies the criteria that authorities use to label their subjects as defectors, the effects that labeling has on individual shifts toward conformity or defection, and the consequences of these shifts for political orders in East Germany during the Cold War and, to a lesser extent, in Palestine. The following chapters present each paper as a stand-alone contribution. I then draw some of the key findings together, and discuss the general applicability of the findings to other 'loyalty panics'.

[1] As one research partner remarked when I inquired about having an open discussion regarding 'collaboration' among university students: "Palestinians do not like to talk about this issue here, it is very different from, say the situation in Vichy France after World War II, we are under occupation and have been for a very long time" (Interview on 13.11.2019).

[2] For instance, archivists find and provide materials based on topics or search queries; while dedicated case workers advise and censor copies of documents for researchers.

PAPER 1
How Loyalty Trials Shape Allegiance to Political Order*

Mirko Reul†[1] and Ravi Bhavnani[1]

[1]Graduate Institute of International and Development Studies, Department of Political Science/International Relations, Geneva, Switzerland.

Abstract

The notion of defection in service of the enemy features prominently in studies of conflict. Yet, far less attention has been paid to behavior political actors construe as defection—the extent to which collaborators are over- or under-identified—prosecuting innocents or failing to prosecute the guilty. This is all the more noteworthy given consequences that range from harassment, to imprisonment, torture, or death. This paper focuses on the dynamics of "loyalty trials" held to identify enemy-collaborators. We argue that the co-production of loyalty—the interaction between expectations, perceptions, and behavior—increases conformity as expected or generates unintended cascades of defection. Using a computational model and data collected from archives and interviews in the GDR and the OPT, we find that defection increases when quotidian behavior is perceived as disloyal, and that conformity increases as disloyalty is increasingly tolerated. The polarizing nature of loyalty trials have notable implications for political order.

*We thank 20 anonymous research partners for sharing their insights about the Occupied Palestinian Territories, 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, Sungmin Rho, Alessandra Romani, David Sylvan, and discussants at ISA 2019, EPSA 2019, SPSA 2020, and CYBIS 2020 for their helpful comments and suggestions. All remaining errors are our own. This project was supported by SNF research grant 188287.

†Corresponding author. E-Mail: mirko.reul@graduateinstitute.ch.

1. Introduction

Loyalty trials span the gamut from the trials of 'socialists' accused of being subversive foreign agents (O'Brian 1948) to the persecution of ethnic minorities for alleged ties to extremists (Mueller and Stewart 2012). In the wake of the 2015 Paris attacks, emergency laws authorized raids to identify individuals capable of causing "big harm" (Chassany 2017). Muslims were consequently harassed, physically assaulted, and denounced for wearing headscarves, frequenting mosques, or praying in public (Faytre 2020; Onishi and Méheut 2020). 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; Hoshur et al. 2018). 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).

From rebel rulers to nation-state governments, a recurrent motif is observed, 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). By doing so, political actors specify the criteria for membership in the communities they govern (see Schlichte and Schneckener 2015, 415; Thiranagama and Kelly 2010, 2), with those labeled being subject to harassment, ostracism, imprisonment, torture or death. In conjunction with official expectations, loyalty trials are shaped by social perceptions of defection. And where official expectations and perceptions diverge, defectors may receive condemnation or social support, spurring widespread conformity with or resistance to official loyalty expectations.

Research on defection has paid scant attention to what constitutes disloyal behavior, as well as the dynamic nature of "labeling defection", disregarding the interplay between loyalty expectations, perceptions, and behavior (see, for example, work by Kalyvas 2008; Schutte 2017; Sullivan 2016a). Given that each loyalty trial defines or 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 their 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 the mechanisms that link loyalty trials to defection, formalizing our theory by means of an agent-based computational model (ABM). The model—validated by empirical evidence from two markedly different conflict settings, 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 explore sequences of interactions and identify the conditions under which different types of regimes over- or under-identify threats to their political order. In the GDR, labeled defectors were perceived as conformers by East Germans, resulting in cascades of defection. By contrast, loyalty trials in the OPT were not fully under the control of the Palestinian Authority (PA) and widely perceived as justified, leading defectors to cease collaboration with Israel.

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 in Section 3, formalizing these by means of a computational model in Section 4. In Section 5 we show how our model simulates allegiance shifts between conformity and defection, and then adjust model parameters to capture the particularities of the GDR and the OPT. We conclude with a discussion of policy measures to safeguard

against the polarizing effects of loyalty trials.

2. Related Work

Loyalty trials are procedures conducted by an ingroup to determine whether individuals are acting in the service of a rival outgroup. They are common to a range of conflict settings, yet, have received considerably little attention in studies of state repression, civil war, or rebel governance. Beginning with the literature on state reporession, the prevailing paradigm suggests that state actors repress social and political rights (Cope et al. 2018) with targets for repression selected on the basis *de jure* or *de facto* rules (Tilly 2003). As such, repression may be overt or covert (Davenport 2005; Sullivan and Davenport 2018), 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 (2016) on state repression in Guatemala is notable in this regard, given his coding of overt and covert behavior perceived to constitute a political challenge. However, this literature focuses on exceptional challenges to political order as an outcome of interest, disregarding changes in quotidian behavior that are not intended to challenge the regime.

Scholarship on civil war also glosses over the interplay between perceived allegiance and political loyalty. Drawing on evidence from the Greek Civil War (1943-1949), Kalyvas (2006) argues that violence by armed groups was 'selective' in areas characterized by incomplete territorial control where defection to rival authorities occurs. Whereas selective violence does *not* attribute "guilt by association"—the mere presence of local agents being sufficient to cultivate this perception (Kalyvas 2006, 190)—Kalyvas remains agnostic about the range of behaviors construed as disloyal as well as the consequences of misidentification. And while he suggests that autocrats need not resort to selective targeting in the absence of rivals (Kalyvas 2006, 143), we argue that even stable autocracies have rivals that engender defection. Much of the literature that builds on Kalyvas' seminal work has focused on the cohesion of armed organizations (Pearlman and Cunningham 2012; Sinno 2008; Staniland 2012), including the conditions for fighters to desert or defect to rival organizations (Albrecht and Ohl 2016; Gates 2017; Koehler et al. 2016; McLauchlin 2010) and those that underpin the incidence of selective or indiscriminate violence (Kalyvas 2012). Yet, this work too pays scant attention to the dynamics of loyalty trials.

Turning to scholarship on rebel governance, 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, resulting in warnings and increased compliance by those suspected of defection. 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, then, 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; Mampilly 2011; Weinstein 2007), this literature also stops short of considering who is labeled a threat to authority, as well as reactions on the part of those labeled.

A handful of case studies do consider how the interplay between loyalty expectations and perceptions determines whether people are put on loyalty trials, and if so, who: 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 et al. 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.

3. The Micro-Dynamics of Loyalty Trials

We suggest that political order is co-produced by authorities who demand loyalty—personal sacrifice meant to enhance group welfare (Levine and Moreland 2002)—and subordinates who to varying degrees conform to **loyalty expectations**.[1] To identify deviance from expectations, and thus more palpably distinguish conformity from defection (Åkerström 1991, 11-16; Coser 1956) **loyalty trials** are held, typically in a "zone of anomie in which legal determinations—and above all the very distinction between public and private—are deactivated" (Agamben 2005, 50-51). 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, have significant implications for political order. We begin by discussing these dynamics below, before turning to our formal model.

To begin with, the identification or **labeling** of defectors has a profound implications for how the labeled see themselves and are seen by others (see Åkerström 1991; Becker 1963; Farrington and Murray 2014). It follows that a suspicion of defection is sufficient to initiate a loyalty trial, whether unofficially via peer-to-peer accusations or officially by means of arrest and interrogation. 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 the motivation of suspects as much as perceptions of their behavior by others: 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; or on normative grounds when defection is attributed to conflicting loyalties that are socially acceptable (Sherman 1993; Sykes and Matza 1957; Tyler and Huo 2002).

Consequently, loyalty trials yield one of four outcomes shown in Table 1.1. When an individual is not labeled, she either conforms (*true conformer*, cell I) or defects (*secret defector*, cell II). In a similar vein, when an individual is labeled, she either conforms (*false defector*, cell III) or defects (*true defector*, cell IV). As such, true conformers exceed loyalty expectations and are identified as loyal, and vice-versa for true defectors. Secret defectors violate expectations but are perceived as loyal or tolerated (see Eck et al. 2020; Scott 1985), and conversely, 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 of the label that shape their experience. False defectors are expected to view themselves as members of the group and attempt to convince their peers of their innocence, at times demonstratively engaging in loyal behavior to do so. By contrast, true defectors

[1] Political order is here understood as "a regular, predictable, and interconnected pattern of institutional and ideological arrangements that structures political life" (Lieberman 2002, 702).

Table 1.1. Typology of Individual Defection

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

Source: Adapted from Becker 1963, 20. Note: True conformers are privately loyal but not labeled as defectors. True 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*.

tend to have few opportunities to demonstrate loyalty, and may even prefer to be ostracized from the group when they no longer identify with the goals or beliefs of other members. Thus labeling need not determine allegiance (see Becker 1963; Matza 2010), though it reduces the agency of the labeled such that they are coerced into demonstrating loyalty or seeking acceptance by rival outgroups.

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 both alter the perceived risks of being labeled (see Oliver et al. 1985) and the benefits of collaboration (see Kalyvas 2006). For those tasked with labeling, the discovery of defection (*true defector*, cell IV) increases suspicion and mistrust, whereas widespread conformity (*true conformer*, cell I) serves to increase trust. As such, loyalty trials can, under certain conditions, diminish challenges to political order as well as exacerbate them (see Davenport and Inman 2012; Lichbach 1987), such as with the over-(*false defector*, cell III)) Mueller and Stewart 2012; Schutte 2017) or under-identification (*secret defector*, cell II) of defectors.

Fourth, private loyalty is positively increased by social and material rewards—including approval by other members for behavior that visibly benefits the group (see Abrams et al. 2018; Hutchison et al. 2011), and monetary payment for denouncing rival activity by authorities (e.g. Piotrowska 2020)—and negatively via social control and sanctioning regimes (see Hechter 1987; Heckathorn 1988).[2] In a similar vein, disloyalty is either positively incentivized by rival authorities, for example through public declarations of support and political asylum, or negatively coerced under threat of punishment, for example through 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 shifts towards 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, perceptions, and behavior; 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 has consequences then 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. To reiterate, the micro-dynamics of loyalty

[2]Note that in some cases, authorities prefer that insufficiently committed group members defect, and may facilitate their exclusion rather than reward or repeatedly punish them to elicit conformity (see Iannaccone 1992; Hirschman 1970, 60; Horz and Marbach 2022, 5-6).

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.

4. Model Specification

We use an agent-based computational model to systematically explore the relationship between loyalty expectations and perceptions on the one hand, and group conformity on the other. We begin by specifying a general model that satisfies our theoretical discussion, 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 et al. (2001) tag-tolerance model, which is in turn motivated by Holland (1995).

Table 1.2. 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
$\sigma_{i,p}, \sigma_q$	Spread in allegiance, tolerance
λ	Loyalty expectations
k	Reward & Punishment

Table 1.2 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, the range of outgroup interactions considered unacceptable and the personal sacrifice required to maintain allegiance increase. 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 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:

$$\Delta_\lambda = \sum_{A=1}^{N} i_A - \lambda \tag{1.1}$$

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

4.1. Mechanism I: Loyalty Trials

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

$$\delta_p = \lambda - p_A \tag{1.2}$$

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.[3] The probability of selecting A over any other agent decreases with k or p_A:[4]

$$p(B \to A) = \frac{e^{k\delta_p}}{\sum_{A=1}^{N} e^{k\delta_p}} \tag{1.3}$$

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:

$$\begin{aligned} \delta_p > q_B \to l_A = 1, \\ \delta_p \leq q_B \to l_A = 0 \end{aligned} \tag{1.4}$$

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

$$\begin{aligned} i_A < \lambda \to d_A = 1, \\ i_A \geq \lambda \to d_A = 0 \end{aligned} \tag{1.5}$$

[3] 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 A.II.2).

[4] 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).

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

$$\begin{aligned} d_A = 0 \wedge l_A = 0 \rightarrow A^I : \textbf{true conformer} \\ d_A = 1 \wedge l_A = 0 \rightarrow A^{II} : \textbf{secret defector} \\ d_A = 0 \wedge l_A = 1 \rightarrow A^{III} : \textbf{false defector} \\ d_A = 1 \wedge l_A = 1 \rightarrow A^{IV} : \textbf{true defector} \end{aligned} \tag{1.6}$$

4.2. 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:

$$p_{A_{t+1}} = p_{At} - p_{At}\delta_{pt} \tag{1.7}$$

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:

$$\delta_i = \lambda - i_A \tag{1.8}$$

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

$$i_{Ag+1} = i_{Ag} - i_{Ag}\delta_{ig} \tag{1.9}$$

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

4.3. Mechanism III: Adaptation

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

$$\bar{q}_{g+1} = \bar{q}_g + (\sum_{A=1}^{N} A^{III} - \sum_{A=1}^{N} A^{IV})\frac{1}{N} \tag{1.10}$$

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

We capture indirect effects of loyalty trials with updates to agent characteristics based on fitness scores,

given by:

$$f_A = \frac{\delta_i^2}{e^{k\delta_i}} - |p_A - i_A| \cdot l_A \tag{1.11}$$

It follows that agent fitness increases with deviance from loyalty expectations, an effect that is moderated by punishment and reward, and decreases with labeling, an effect that is moderated by the distance between private and perceived behavior. We provide a more detailed discussion of fitness scores in Appendix A.I. Table 1.3 illustrates the payoffs for each defector type assuming that expectations are extremely high: **Secret defectors** receive a payoff of 0.4, given that they reap the benefits of outgroup interaction without being labeled. **True conformers** receive a payoff of 0, assuming that loyalty is valued less by ingroup authorities than is disloyalty by rival authorities. **True defectors** receive a payoff close to -0.5, given that they are labeled for private disloyalty. **False defectors** receive a payoff of -1, the equivalent of a 'sucker's' payoff, given that they are privately loyal but perceived to be disloyal. 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 et al., each agent mutates her new tags and tolerance level with probability $M = 0.1$[5].

Table 1.3. Payoffs for Defector Types

	Conforming	**Defecting**
	I	**II**
Not labeled	"True Conformer"	"Secret Defector"
	0	**0.4**
	III	**IV**
Labeled	"False Defector"	"True Defector"
	-1	**-0.5**

Note: Payoffs rounded to nearest tenth, based on extreme loyalty expectations ($\lambda = 1$), matching loyalty incentives ($k = \lambda$), and maximum distance between A's private and public allegiance, such that $p_A - i_A = 1$ for secret defectors, 0.9 for true defectors, 0 for true conformers, and 1 for false defectors. As a result, $A^{II} > A^{I} > A^{IV} > A^{III}$.

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. A 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 1.1.
   - 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.

[5]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.

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.

5. Results

5.1. General Model

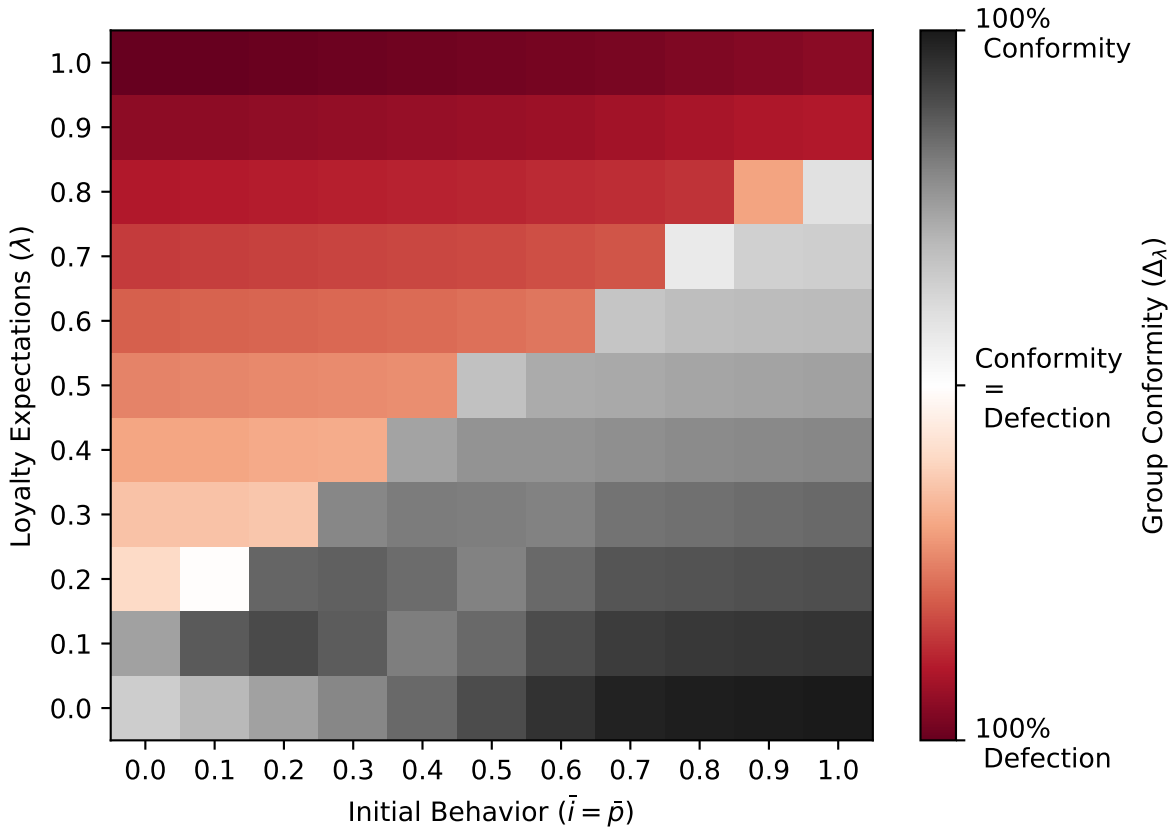
We begin by discussing how group conformity changes in response to loyalty expectations and behavior. The general relationship is depicted in Figure 1.1, with each cell representing a different experimental setting. We note that small changes in initial behavior (x-axis) and loyalty expectations (y-axis) can lead to significant changes in group conformity (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: true conformity for $\bar{i}, \bar{p} >> \lambda$, and conversely, true defection for $\bar{i}, \bar{p} << \lambda$. But when some agents are conforming while others are not, the dynamics of loyalty trials disrupt these equilibria.

Figure 1.2 depicts two typical patterns of *allegiance shifts* by defector types over the course of model runs. In panel (A), true conformity increases while 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), true defection increases while conformity decreases. Opportunities for secret defection outweigh the benefits of compliance in response to labeling, and as tolerance increases, 'Type II' outweigh 'Type I' errors. As in Granovetter models of political protest (Granovetter 1978; Kuran 1989), cascades of true defection ensue. Note however that shifts toward defection do not occur in response to labeling without opportunities for covert disloyalty, and that our model simulates the conditions for cascades to occur, but stops short of identifying their endpoint.

These two patterns, robust to a wide range of auxiliary parameter specifications (see Appendix A.II), result in highly polarized outcomes (see Esteban and Ray 2008; 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 reward-seeking for conformity and defection. Insofar as the intensity of these dynamics is given by the frequency of agent interaction over a time, regimes that produce fragmented spaces of social interaction delay, rather than prevent, irreversible allegiance shifts (e.g. Pfaff 2001). Empirically, allegiance outcomes 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.

A key assumption in the general model is that most agents are initially conformers and perceived as such ($\bar{i} = \bar{p}$). Empirically, conflicts deviate from this assumption: defectors may be perceived as conformers and vice-versa. We contextualize the general model to analyze these conditions with reference to the loyalty expectations of regimes in the GDR and the OPT, highlighting how the dynamics of loyalty trials operate in both settings despite vast contextual differences.

Figure 1.1. General Model: Group Conformity



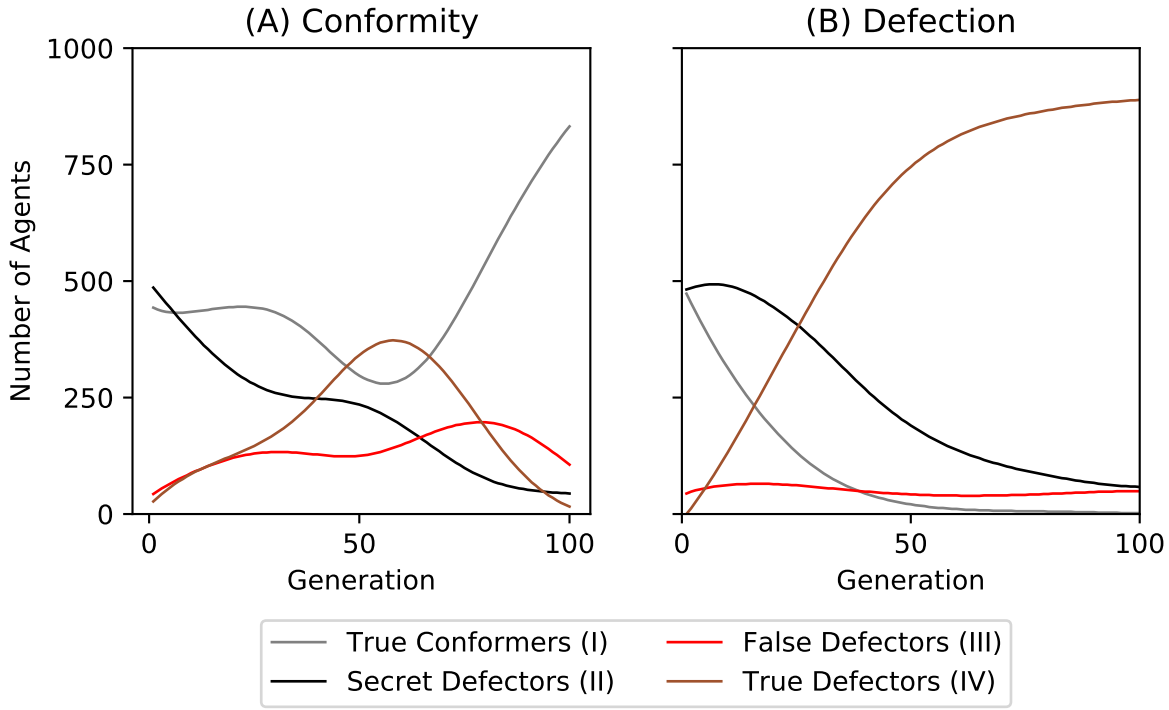
Note: Each cell shows the final group conformity for an experimental condition, averaged across simulations, given initial loyalty expectations (λ) and allegiance ($\bar{i}, \bar{p}$). Each experimental condition is simulated $S = 30$ times, and each simulation lasts $G = 100$ generations. Simulations 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, $\sigma_i, \sigma_p, \sigma_q = 0.1$ initial dispersion of agent parameters, and $\bar{q} = 0.1$ initial agent tolerance. 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. Agent parameter values are drawn from the normal distribution. Table A.4 and Table A.5 provide details on parameters and results.

5.2. Contextualizing the Model

Model contextualization comes with several challenges: it requires ontological assumptions about the reference "group", the situational contexts 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 it is construed by political authorities is both rare and challenging to observe: the incidence of defection exceeds conformity only while a group fragments, and is otherwise worthwhile hiding for defectors and authorities alike. This observational challenge is exacerbated by the relational nature of loyalty: expectations, perceptions and tolerance for disloyal behavior are permanently in flux in ongoing conflict setting, and their overt expression is rarely documented. Available estimates of defection are unreliable, given that the requisite data is either classified or unverifiable. Even the most diligent government employees are prone to misrepresent Type I and Type II errors in an effort to justify their activities, making such data unsuitable for model contextualization (see Appendix A.III.1 for details).

Thus in place of validating the model statistically, we content ourselves with a qualitative, most-different

Figure 1.2. Types of Allegiance Shifts



Note: Each panel shows the LOWESS for the prevalence of the four defector types from Table 1.1 (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 (see Table A.5): $\lambda = \bar{p} = \bar{i} = 0.5$, $\sigma_i, \sigma_p, \sigma_q = 0.1$, $\bar{q} = 0.1$, $G = 100$, $N = 1000$, $M = 0.1$, $P = 3$, $T = 10\%$.

case comparison. While a qualitative contextualization is less reliable than a quantitative enumeration, it provides a more truthful representation of real-world dynamics in the GDR and the OPT. The specification of the model is partially based on a review of *Stasi* surveillance on 319 individuals in the GDR, as well as 20 interviews with Israelis and Palestinians in 2019 and 2022 for the OPT (see Appendix A.III).

The GDR and OPT differ on at least three dimensions that are not endogenous to the model: the volatility of loyalty expectations (changing more frequently in the OPT than in the GDR), the number of political authorities who attempt to enforce different expectations at the country-level (one in the GDR, multiple competing authorities in the OPT), and the share of defectors who were labeled officially by state agents (GDR security agencies had more official control over defectors than those in the OPT). Despite these differences, we argue that the mechanisms linking loyalty trials to allegiance outcomes work similarly in both settings. 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.

Table 1.4 illustrates how our operationalization of loyalty is designed to reflect substantive differences between the GDR and the OPT. We associate increases in loyalty with increasingly quotidian behavior

Table 1.4. 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	$\bar{p}$	λ
Well-Being $(0.7 - 0.9)$	Illegal Emigration	λ	$\bar{\bar{i}}$
	Work in Rival Area	$\bar{\bar{i}}$	$\bar{p}$
	Personal Outgroup Contact		

Note: We view each of the listed behaviors as violating a level of loyalty in the given range. 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. For each conflict setting, we set parameters based on qualitative evidence, such that in the GDR $\bar{p} < \lambda < \bar{\bar{i}}$ and in the OPT $\lambda < \bar{\bar{i}} < \bar{p}$. Loyalty expectations reflect the minimum personal sacrifice that is expected from all group members by a single political authority, as we do not distinguish between subgroups with different expectations. Private and perceived loyalty parameters indicate which types of disloyalty group members would on average *not* commit. The full numerical representation of parameters is given in Table A.6. This contextualization represents, rather than pinpoints or predicts, individual behavior.

and decreasing rival activity, from defending the group's physical security (e.g. refusing enemy-informing to the outgroup), over maintaining its unity and status (e.g. supporting policies unfavorable to the outgroup), to improving the socio-economic well-being and independence of its members (e.g. employment and taxation benefiting the ingroup).[6] This ordering of behaviors in terms of loyalty requires validation in specific conflict settings. Some behaviors are not construed as disloyal in one case but are in the other: selling land to the outgroup was not construed as disloyalty in the GDR, and neither were revolutionary plots or emigration in the OPT. But we can categorize disloyalty by the threat it poses to a group—its security, unity and status, and socio-economic well-being.[7] We seed an initial loyalty distribution based on existing studies, and discuss how the allegiance shifts that the model produces line up with evidence at the micro- and macro-levels. In both settings, most group members are privately conforming with loyalty expectations ($\lambda < \bar{\bar{i}}$). But whereas in the GDR privately loyal individuals were labeled as defectors from high loyalty expectations ($\lambda - \bar{q} > \bar{p}$), perceived defection from moderate loyalty expectations was tolerated in the OPT ($\lambda - \bar{q} < \bar{p}$). In the following, we justify the specific parameter values and discuss the results for each setting with reference to existing studies and archival materials.

[6]We also assume that higher levels of loyalty subsume lower levels, such that expectations of socio-economic sacrifice would equally construe behavior threatening the physical security or unity of the group as unacceptable.

[7]Each of the listed disloyal behaviors encompasses a range of activities that are perceived to indicate disloyalty, and similar indicators may indicate different types of betrayal . For example in the OPT, homosexual and extramarital relations are associated with enemy-informing particularly since the First Intifada (1989-1993), as Israeli intelligence is known to blackmail this socially marginalized group of Palestinians (Be'er and Abdel-Jawad 1994; O'Conner and Simeonov 2013; see al-Bitawi 2016, 63). By comparison in the GDR, sexual deviance was generally associated with activism benefitting the West, as security agents construed a link between social deviance and Western ideology. But within the "greedy institution" of the Stasi, extra-marital relations were associated with Western recruitment efforts (Krähnke et al. 2017, 172), and the perception that sexual deviance is intolerable was reinforced by true defectors such as Werner Stiller, who spied for Western intelligence and emigrated to the FRG with the help of his secret girlfriend (Glocke 2002; Gieseke 1999, 539).

5.3. 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 control that authorities exercised over the identification of defectors was relatively high, as the Ministry for State Security (*MfS* or *Stasi*) drew on an infamously vast surveillance and reporting system to conduct loyalty trials, with punishments ranging from demotions and party reprimands to imprisonment and (until 1987) death sentences (Kowalczuk 2013; Müller-Enbergs 2008; 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 Lohmann 1994; 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), and the 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, the MfS construed internal dissent and economic inefficiencies as defection, with its head Erich Mielke coining the term "political-ideological diversion" to construe an affiliation between deviant individuals and Western aggression (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), a fact that was known to would-be defectors who took it as an "insurance" 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 (Gieseke 2003; Rink 1997), and politicized community organizations exacerbated over-identification by treating social deviants who glorify life in the West as defectors (see Budde 2014).

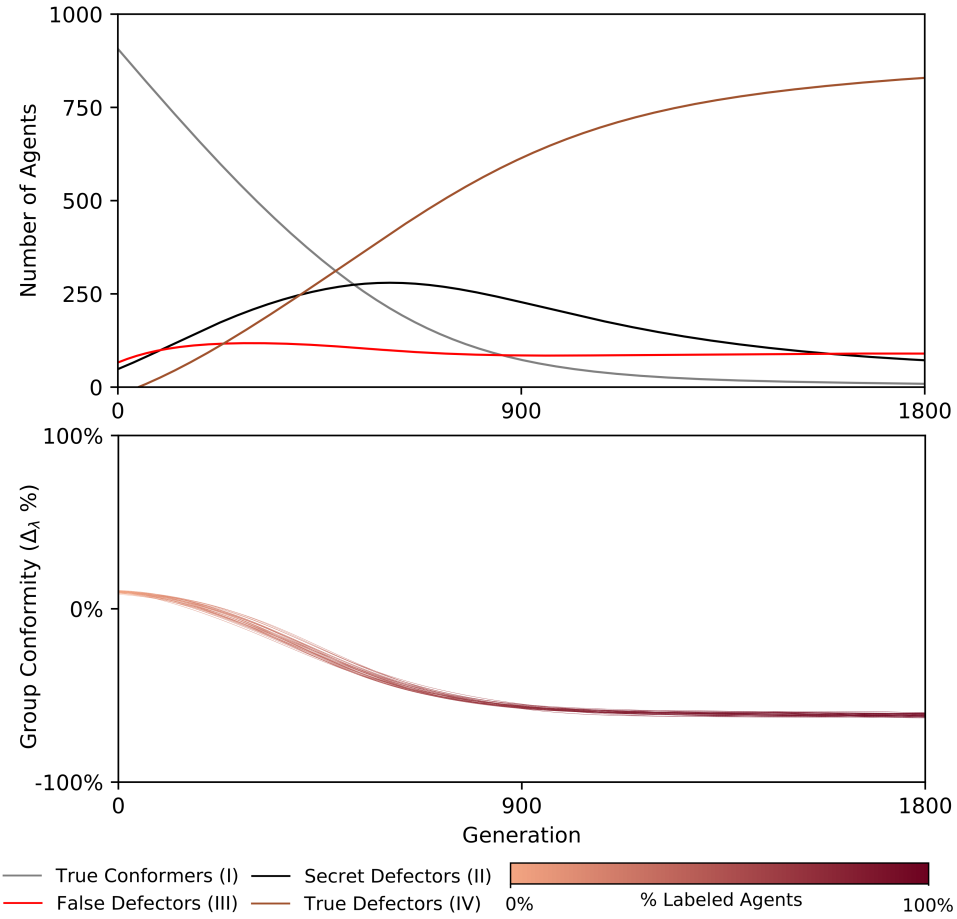
Tolerance: Given the prevalence of informants, as well as authority rewards and protection granted to informants (Piotrowska 2020), there was a high chance that perceived defection would be labeled. From schools and work places to neighborhoods for state security personnel (Hoffmann 2012; Krähnke et al. 2017), deviant behavior was reported to authorities, followed by investigations and interrogations, if not formal sanctions.

Interaction: Citizens in the GDR at the time had developed a tendency to withdraw from public life (see Pfaff 2001), suggesting that perceptions of disloyalty spread relatively slowly across the population. By the same token, given preemptive, individual and covert targeting by the MfS, a relatively small proportion of society learned of loyalty trials through private contacts.

Overall by 1971, GDR authorities had high loyalty expectations for mostly conforming citizens. They over-identified defectors, most of them privately and preemptively, but could not match incentives for disloyalty provided by their Western rivals.

Results

Figure 1.3. GDR Allegiance



Note: **(Top)** Defector pattern LOWESS over generations, averaged across simulations. **(Bottom)** Group conformity LOWESS for $S = 30$ simulations. Counterfactual runs in Appendix A.III.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. Parameter values for $N = 1000$ representative agents reflect relative differences between the two conflict settings, and are drawn from the normal distribution: $\lambda = 0.7$, $k = -1$, $\bar{i}^{95\%} = 0.8$, $\bar{i}^{4\%} = 0.5$, $\bar{i}^{1\%} = 0.2$, $\sigma_i = 0.05$, $p_A = i_A - 0.1$, $\sigma_p = 0.1$, $\bar{q} = 0$, $\sigma_q = 0.01$. Due to relatively infrequent public interactions regarding defectors, $P = 1, T = 0.5\%$. Compared to the OPT, infrequent interactions are offset by a larger number of generations, corresponding to the relatively longer time frame: $G = 1800$.

Figure 1.3 shows how East German allegiance declines as defectors are increasingly labeled, following the *defection* pattern in Figure 1.2: falsely labeled defection indirectly increases tolerance and secret defection, which in turn lead to cascades of true defection in defiance of labeling. We corroborate our argument about the mechanisms leading to this shift at the micro- and macro-levels, drawing 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 were the easing of travel restrictions and recognition

of sovereignty between East and West, and the 1975 Helsinki Accords that signaled a normalization of relations to GDR citizens (Gieseke 1999, 539; Raschka 2001, 37-44). Particularly those with family connections to the West submitted emigration requests, which rose by 70% between 1975 and 1976 (Eisenfeld 1999, 385), pushing an intolerant MfS to over-identify defection (Gieseke 2014, 61; Passens 2012, 167-169).

Falsely labeled defectors complied with authorities to prove their loyalty particularly in conforming social circles (e.g. Krähnke et al. 2017, 235-252; Hoffmann 2012), but their labeling encouraged adaptation towards defection. Examples include state employees who were pushed to cease Western contacts or switch jobs (e.g. BArch, MfS, HA XVIII, 6320), emigration request denials that were perceived negatively among work colleagues (e.g. BArch, MfS, HA XVIII, 37797), and church officials who complied with the MfS but still encouraged activism perceived as disloyalty (e.g. BArch, MfS, BV Potsdam, KD KY, Nr. 75, Bd. 1-3).

Secret defection resulted from adaptation of labeled behavior, following the regimes 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 (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.

True defectors were defiant of attempts to treat their behavior as disloyal, and increasingly supported by their peers in their defection. Examples include persistent emigration requests after labeling (e.g. BArch, MfS, HA XVIII, Nr. 38403), overt backlash against denunciations of discontent workers (e.g. Halbrock 2015, 144-145), and overt criticism of the GDR in response to labeling. Defection was exacerbated by the protection of defectors by the FRG, Western human rights organizations, media, and the protestant church (Eisenfeld and Eisenfeld 1999, 97-98). In particular, the church-state rivalry produced similar loyalty trials to those that forced a minority of GDR citizens to declare allegiance for one nation-state over the other, with long-term consequences for secularization (Wohlrab-Sahr et al. 2008).

5.4. OPT Allegiance During the Second Intifada

In the OPT, political authority is contested along with territorial control between Israel, the Fateh-led PA, Hamas, and affiliated armed organizations that at times enforce loyalty expectations autonomously (see Pearlman 2011). Palestinian loyalty expectations vary considerably with the threat the Israeli occupation poses to national self-determination, and the social control exercised by Israeli institutions (Albzour 2017; Handel and Dayan 2017; Nerenberg 2016; Zureik 2010), including the use of administrative detention, blackmail, and movement restrictions to control and recruit informers (Berda 2017; Cohen 2010; Sorek 2010). The many organizations comprising the Palestinian Security Services have arguably less formal control over the identification of defectors than the Stasi did, as evidenced by the pletora of armed groups who at times label defectors independently (B'Tselem 2021b; Be'er and Abdel-Jawad 1994; Tartir 2015).

We discuss the Second Intifada from 2000 until the death of president Yasser Arafat in 2004, again marked by relatively stable loyalty expectations while he was at the center of the pletora of PA security agencies officially responsible for enforcing them (see Tartir 2015). Compared to the GDR, the fitting of parameters to the asymmetric civil conflict in the OPT is challenging, given that loyalty expectations were contested and private and public allegiances relatively diverse (see Nerenberg 2016, 215-248).

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 Palestinians who are collaborating with Israel. By the onset of the Second Intifada in 2000, the PA was accommodating Israeli pressure to maintain moderate loyalty expectations, but political authorities did not officially expect socio-economic loyalty (see Nerenberg 2016, 198). [8]

Loyalty Incentives: The importance that Palestinians attribute to *Sumud* as a form of everyday resistance (Ali 2019), and the widespread knowledge of Israeli arrest and recruitment practices given their long history (Cohen 2008; 2010), contribute to the strong loyalty incentives that allow Palestinians to resist the occupation, in spite of the high levels of Israeli coercion and material rewards offered to defectors.

Private and Public Allegiances: Most Palestinians privately exceeded moderate loyalty expectations, and their perceived allegiance exceeded private allegiance, while rare land-dealers and enemy-informants were perceived as defectors who threaten the security of Palestinians (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 (al-Bitawi 2016; Cohen and Dudai 2005; DCI Palestine 2012; Hass 2019).[9]

Interaction: Compared to the GDR, defector suspicions were regularly shared and salient loyalty trials carried out in public, as Palestinians recognized the implicit "complicity with the Israeli occupation" in their everyday lives, and labeling constituted an expression of fear over being forced into loyalty conflicts, particularly by the *Shabak* (Kelly 2010, 183-184).

Overall, the PA enforced moderate loyalty expectations on mostly conforming Palestinians, most of whom were either perceived as loyal or tolerated. Over-identification occurred where defector perceptions diverged between official PA and unofficial NSAG labeling, while the spread of public rumors on enemy-collaborators was driven by a fear of precarious situations that coerce disloyal behavior (see Table A.6 for specific parameter values).

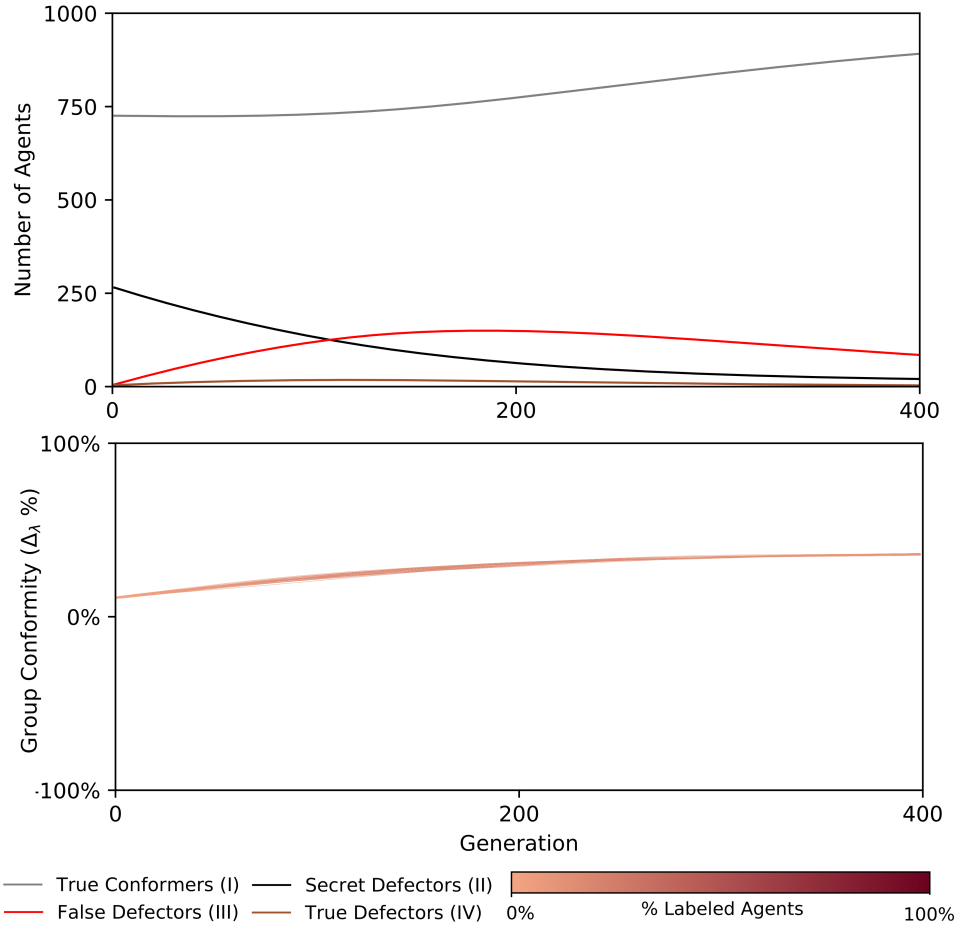
Results

Figure 1.4 shows how Palestinian allegiance increases as less defectors are labeled, following the *conformity pattern* in Figure 1.2 (A): falsely labeled defectors have incentives to increase their loyalty and are increasingly tolerated by authorities.

[8] This reflected the general perceptions among Palestinians that some have no choice but to work in Israel or settlements to provide for their families, even though economic collaboration ultimately threatens Palestinian liberation (see Ophir 2020; Roy 1987).

[9] Informants were unofficially tolerated particularly by Fateh, 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).

Figure 1.4. OPT Allegiance



Note: **(Top)** Defector pattern LOWESS over generations, averaged across simulations. **(Bottom)** Group conformity LOWESS for $S = 30$ simulations. 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 A.III.2). Parameter values for $N = 1000$ representative agents reflect relative differences between the two conflict settings, and are drawn from the normal distribution: $\lambda = 0.6$, $k = 1$, $\bar{i}^{59\%} = 0.7$ 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 $p_A^{1\%} = i_A^{1\%}$, $\sigma_i = 0.1$, $\sigma_p = 0.2$, $\bar{q} = 0.1$, $\sigma_q = 0.05$. Due to relatively frequent public interactions regarding defectors, $P = 3, T = 1\%$. Compared to the GDR, the frequency of agent interaction is offset by the relatively shorter time frame: $G = 400$.

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% (JMCC 1999; 2000; PSR 2000; 2001). 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.

True defection was difficult to sustain given 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, 110 Palestinians were killed for collaboration, most of them during the Israeli incursion of the West Bank in April 2002,[10] 24 sentenced to death without

[10]This lends credence to the notion that rival threats drive loyalty trials, an observation that holds across conflict

sentences carried out, and over 600 detained by the end of 2001 (B'Tselem 2021a; 2021b; Human Rights Watch 2001, 26-27,49-50). Those on trial were vilified, including by legal representatives (e.g. Williams 2001, 30-32; Human Rights Watch 2001, 45-46; al-Bitawi 2016, 35).

False defectors had little choice but to repent and try to prove their conformity. This was the case for labeled enemy-informants, who were accused based on "rumors, suspicions, and popular denunciations" (Human Rights Watch 2001, 23), and whose families are stigmatized even if labels turn out to be false (e.g. Jalal 2015; Human Rights Watch 2001, 47-48; Williams 2001, 32-36), and for PA security collaborators who were labeled for conceding to the Israel agenda (e.g. Kelly 2010, 179).

Secret defectors, whose collaboration with Israel on security issues had previously been not only tolerated but expected by authorities, demonstrated public allegiance to redeem themselves or avoid future labeling (e.g. Cohen 2012, 478-479; Cohen and Dudai 2005, 239-240; Pearlman 2011, 154; Berda 2017, 31-32; Kelly 2010, 179), and those few who remained hidden presumably received support from Israel to do so (e.g. Yousef 2010, see al-Bitawi 2016).

6. Discussion

Loyalty trials occur across a range of conflict settings, despite differences in regimes, repression, and social identification. Anchored by 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 on the outside, and by the protestant church internally. In the OPT, by contrast, increased loyalty expectations and group solidarity resulted in greater conformity. It follows that defection and allegiance shifts were more likely in the GDR relative to the OPT, given higher expectations and misidentification, and lower incentives for loyalty.

Beyond the particularities of the two cases, our theoretical framework has implications for political order writ large. First, loyalty trials polarize conformers and defectors alike, a process that is largely beyond the control of political actors. Second, increasing loyalty expectations and decreasing tolerance in democracies runs the risk of triggering cascades of false labeling, what we identify as a Type-I error. Research on the repression-mobilization nexus would benefit from taking loyalty expectations into account, as they vary even among democratic regimes and may impose powerful constraints on collective action. In a related vein, the propensity of regimes to under-identify defection negatively affects regime legitimacy and survival. Discrepancies between expected and perceived loyalties are challenging to measure, and point to the "quotidian struggles" that make seemingly stable regimes vulnerable (Scott 1989; Wedeen 1999, 87). Third, our account suggests that repression and external support for dissent are effectively co-produced, going beyond recent scholarship that reduces the identification problem to the number of informants (e.g. Steinert 2022).

While loyalty trials rarely assume center-stage in studies of social conflict, treason most commonly ranks among the crimes considered 'worthy' of capital punishment (Thiranagama and Kelly 2010, 1-2). Noteworthy, in this regard, is that loyalty expectations persist well beyond their original manifestations, with attendant implications for social cohesion, trust, and "ethnic defection" (Kalyvas 2008; Staniland 2012). Former collaborators with the GDR regime are considered untrustworthy some 30 years after unification with West Germany (Spiegel 1993; Zeit 2019), and PA collaboration with Israel continues to undermine its legitimacy (Tartir 2019). Four years after the alleged 2016 coup d'etat attempt in Turkey,

settings in the OPT: today loyalty expectations regarding informing to Israel in the West Bank are low compared to the Hamas-administered Gaza Strip, given PA collaboration with Israel and the focus of Israeli air strikes on Gaza (Murphy 2018; Salah 2019; Xinhua 2019).

state employees were still being arrested on suspicions of affiliation with the Gulen network (Gulen 2020). AP News claimed that Konasheher county in Xinjiang province has the highest density of prisoners per 100'000 residents in the world, linking arrests to terrorism charges (Wu and Kang 2022), a trend that goes back at least to the 1990s (Roberts 2018). And persistent rivalries between nation-states frequently trigger loyalty trials that are deemed unjustified by human rights organizations (e.g. Nechepurenko 2019; Shukla 2019). At the heart of these trials lies the interplay between expectations, perceptions, and behavior, the associated perils of over- or under-estimating defection, and resulting consequences for intra-group polarization and conflict.

PAPER 2
Does Labeling Defectors Stop Betrayal? Evidence from the Stasi Archives*

Mirko Reul[1]

[1]Graduate Institute of International and Development Studies, Department of Political Science/International Relations, Geneva, Switzerland.

Abstract

How does repression affect political deviance? Governments use preventive repression to curb deviant behavior that perceptibly threatens political order and their subjects, ranging from violent attacks and enemy-informing to contentious action and emigration. Research on conflict tends to view political deviance through the lens of the state-challenger model: states criminalize or tolerate deviance, and accuse individuals of disloyalty, who respond with performative compliance, repertoires of defiance, or 'exit'. This paper presents preliminary evidence that the state-challenger model omits unofficial, social determinants of both repression and deviance. The analysis is based on an original sample of *Stasi* surveillance on 319 suspected deviants in the GDR (1961-1989), including operation reports, unofficial denunciations, interrogation protocols, and suspect statements. I suggest that deviance is *labeled* as disloyalty based on three 'facets' of the concept: private behavior, public perceptions, and social constructs of deviant identities. The labeled negotiate their group membership with authorities when the label conflicts with their social identity, and shift allegiance between loyalty and disloyalty mostly in response to *unofficial* labels by their peers. Ironically, repressing political deviance is least consequential for individual behavior when applied to the truly disloyal.

*I especially thank Christian Carlsen and Friedrich Rother at the German Federal Archives for their exceptional assistance with the archival materials, as well as Christian Halbrock and David Sylvan for helpful comments and suggestions. Translations from German and emphases in quotations are my own. This project was supported by SNF research grant 188287.

1. Introduction

The labeling of deviance in social conflict settings changes political behavior. In 1973 East Germany, a border guard soldier in his twenties was increasingly strained between his military service which he began to dislike, and work on the farm of his parents-in-law. Other soldiers denounced conflicts with his wife, pointing to his infidelity and financial debts, and his superiors reprimanded him for alcoholism on duty. Caught between his obligation of military service and personal problems, the 'Deviant Defector' collected classified information on security installations at the border, with plans to emigrate to the West to solve his financial problems and later affect the emigration of his family. Nine months later, another talk aimed at disciplining his social deviance and a remark by his superior make him fear that his activities had been detected. That same same evening, he crossed the border into the Federal Republic of Germany (*FRG*), explaining his motives in a letter to his wife.

After his desertion, the former soldier was asked to share what information he has by Western intelligence services, and accepted their offer of financial support and unification with his family, in exchange for a formal commitment to the organization. Seemingly deviating from his original plan, he received additional training for two months and was instructed to return to the German Democratic Republic (*GDR*), repent but deceive state security about his recruitment, and engage in industrial espionage. However, he admitted to his recruitment under interrogation, and was sentenced to 12 years of prison for informing to an enemy. Finally, his demeanor in prison convinced state security of his loyalty: he was released early and reacquired his GDR citizenship (BArch, MfS, GH, 17/79).

Deviant Defector represents an exceptional class of *allegiance shifts* between loyalty and disloyalty, compared to other deviants who were seen as disloyal by the Ministry for State Security (*MfS*, *Stasi*) in the GDR: A state employee who had been fired for regime-critical remarks was later found repentant and recruited as an informant (BArch, MfS, HA II, 38995, p. 202-215); while another preferred to be demoted over ceasing his Western contacts in an attempt to save his marriage (BArch, MfS, HA XVIII, 6320). Around 1980, a 'circle' of academics was reprimanded for their largely philosophical attempts at reforming the Socialist political order, yet only those who defied this accusation of disloyalty lost their academic positions (BArch, MfS, AOP, 16183/81). And in 1978, a self-declared 'politically neutral' citizen spent months in interrogations, providing details to state security about his alleged special training by Western intelligence for a plot to kidnap a senior member of the Socialist Unity Party (*SED*), only to reveal to his captors how he made it all up to affect his deportation to the West (BArch, MfS, GH, 337/79).

In each case, political actors construe quotidian behavior as disloyalty, to prevent exceptional deviance that ostensibly threatens the group they represent. But while some of those who were *labeled* as disloyal seemingly ceased deviant behavior, others refused or even intensified it. Under what conditions does the *labeling* of individuals as disloyal lead to shifts between loyalty and disloyalty, as opposed to merely confirming prior allegiance?

Some of the behaviors described above could be sorted into established categories of political deviance—such as 'voice' or 'exit' (Hirschman 1970)—yet behavior change between these extremes is not captured, and the accounts above are mostly missing the political motivation that is attributed to deviance by authorities. In Section 2, I review research on political deviance across a spectrum of intrastate conflict settings, arguing that loyalty can be understood as observed, perceived and imagined political deviance. Section 3 describes a methodology to measure these facets of loyalty, based on archival data on 319 East Germans who were surveilled by the *Stasi* after the closure of the East-West border. Section 4 applies this approach to a sample of 15 individuals: I analyze the conditions for authorities and group members to justifiably label deviants, and in turn how the labeled change politicized behavior. I find that labeled individuals shift allegiance towards loyalty when their behavior was deemed unacceptable by peers with

whom they sought to belong. Ironically, this was mostly the case for loyal suspects whose deviance was incidental, such that their labeling was not perceived as justified. I conclude with a discussion of general implications for the study of political deviance in conflict.

2. Labeling Political Deviance in Conflict

One lens to study political deviance in social conflict is the state-challenger model, whereby political actors prescribe, tolerate and repress social and political rights (Cope et al. 2018; Davenport 2007), while their challengers deviate by infraction of these rules, drawing on civil disobedience or contentious action (McAdam et al. 2001; Tarrow 1994; Tilly 1978; 2003). The tactics used by nation-state governments are directed at their challengers or their relatives and associates, ranging from somatic violence to nonviolent forms of co-optation, indoctrination and infiltration of dissident networks (Hassan et al. 2022). In turn, challenger activities in their role as resistance members constitute political deviance, from armed struggle over petitions to robberies and political education (Davenport 2005; Sullivan and Davenport 2017). In this view, deviance is represented by individual insurgents or oppositional organizations who violate rules set by a governing authority, and what makes it political is their intention to challenge political order.

The questions that are addressed in this field consequently focus on motivated challenges to authority and their repression. For challengers, this includes their willingness to mobilize (Cederman et al. 2013; Gurr 1970; Olson 1965; Shesterinina 2016; Wood 2003), their capacity to organize collectively given resource and information constraints (Kuran 1989; Oliver and Marwell 1988; Oliver et al. 1985), and their institutional "bargaining power" vis-a-vis nascent rulers (Arjona 2016). In turn, the efficacy of state repression is impacted by the availability of information given territorial control (Kalyvas 2006; Steinert 2022; Svolik 2012), the conceivable strategies to coup-proofing security forces such that they do not defect to rivals (Dworschak 2020; Koehler et al. 2016), the resources to co-opt and fragment political opposition (Gandhi and Przeworski 2007; Kalyvas 2008), as well as the balancing of authoritarian power-taking and power-sharing against the risk of rebellion (Svolik 2012). Where state and challenger are studied in unison, they constitute independent actors who adapt their covert and overt tactics to manage grievances and resist repression, respectively (Lichbach 1987; Ritter 2014; Sullivan and Davenport 2018).

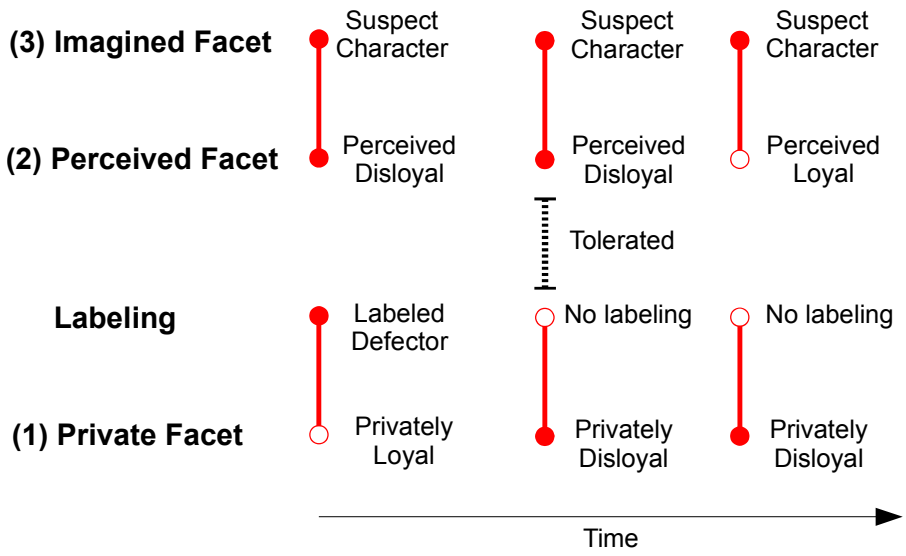
In treating political authorities and their challengers as distinct objects of study, the model suffers from two shortcomings: it ignores political behavior that authorities do not label as deviant, while taking at face-value the political motivation of behavior that is so labeled (see Kalyvas 2003); and it can only gauge behavior change by the designated challenger's mobilization or control over government institutions. These pitfalls have long been recognized in interpretive case studies of resistance and authoritarian rule. In his seminal ethnographic study of everyday resistance among Malaysian peasant farmers, Scott (1985, 289-302) argues that their "ordinary weapons", such as "false compliance" or "sabotage", escaped identification by authorities, were not collectively organized acts, stemmed from personal motives for marginal gains rather than systemic grievances for revolutionary change, and were directed at local superordinate classes rather than at central elites, notwithstanding an awareness for larger political concerns.

Adding to this the role of a central government, Wedeen (1999, 67-68) shows how in Syria since the 1970s, the regime expected that its "citizens provide external evidence of their allegiance to a cult" around president Hafez al-Asad, while Syrian's acted "as if" they believed in their own symbolic displays of loyalty. In this view, political deviance is represented by slowly developing "transgressions" that are recognizably "oppositional" practices, such as joke-telling or comedy skits which ridicule the ruler, yet are tolerated by the regime in the shared understanding that even false displays of loyalty sustain power

relations (Wedeen 1999, 89,92). This cultural lens arguably provides a more meaningful and accurate depiction of political deviance and its absence than the state-challenger model does most of the time (see Davenport 2022; Jamal 2022). But while it explains the absence of violent coercion against deviants, it does not address the 'exceptional' situations where state agents rely on coercion to repress deviance that is not tolerated (see Agamben 2005).

As illustrated in Figure 2.1, I draw on three 'facets' of *political loyalty* to bridge some of the conceptual distance between highly quotidian and exceptional manifestations of political deviance: (1) private, (2) perceived, and (3) imagined disloyalty.

Figure 2.1. Facets of Political Loyalty



Note: Stylized depiction of the three facets of loyalty and labeling over time. At the first observed point in time, a group member is not disloyal (1), but (2) perceived and labeled as if they were, while (3) their ascribed identity fits that of a typical deviant as construed by authorities. In the second observed time step the same individual is (1) privately disloyal. But they are (2) not labeled even though they are still perceived as disloyal, due to their disloyalty being tolerated. In the final time step, the individual is falsely perceived as loyal, even though they are privately disloyal and they match the profile of a traitor. Overall, this case would be interpreted as an allegiance shift from loyalty to disloyalty after labeling.

First, the private facet was introduced by Hirschman (1970, 77-79) as the "special attachment to an organization" that makes individuals more likely to 'voice' their grievances for the uncertain prospect of improving an organization or political order, rather than choosing to 'exit' and thereby improving their choice with relative certainty. For instance, the concept is used to explain a divide between the disloyal population that would 'exit', and the sufficiently loyal population that may remain with the group and resort to 'voice' instead (Hirschman 1993; Pfaff and Kim 2003). This use of the term is akin to what some social psychologists might define as the "adherence to a social unit to which one belongs, as well as its goals, symbols, and beliefs" (James and Cropanzano 1994, 179-180). In this social identity approach, the intrinsic or "dispositional" motivation to make sacrifices for a group is driven by the importance of that group for individual self-perception (James and Cropanzano 1994; Tajfel and Turner 1986), which may then be put to the test by expectations from competing political actors and the incentives they offer for disloyalty. As such, private loyalty is behavior that is intrinsically motivated by social identification. But this conceptualization does not capture how behavior that is not intended as loyalty may be socially constructed "as if" it is, or vice-versa (see Lauderdale 2015; Wedeen 1999).

Second, loyalty is *perceived* personal sacrifice that enhances group welfare at the expense of a rival group

(see Zdaniuk and Levine 2001; Levine and Moreland 2002; Poulsen 2020, 9; Åkerström 1991). In this view loyalty is relational, such that the same behavior may signify loyalty or disloyalty depending on who is observing it. Disloyal behavior may be seen as quotidian (e.g. listening to regime-critical music) or exceptional (e.g. sabotage), tolerable or unacceptable (e.g. due to empathy with the deviant or its absence), and controlled officially by a state security agent or unofficially by another superordinate group member (see Simmel 1964), such as a political party leader, military superior, employer, parent, or teacher. And when a group member does not tolerate perceived deviance in a particular instance, they label the suspect as disloyal. It follows that not all of those who are privately disloyal or even perceived as such are labeled, and not all of those who are labeled are privately disloyal.

Third, in line with the labeling approach to social deviance (Becker 1963; Goode 2015), political deviance is socially construed as a threat to the group.[1] Political authorities define and propagate abstract characterizations of deviants or 'traitor profiles', based on observed or imagined threats to the group, and thereby delineate those who may be suspected of disloyalty from those who are above suspicion (see Davenport 2005; Hewitt 2010; Thiranagama and Kelly 2010). An individual is disloyal to the extent that their ascribed identity matches that profile. For example, deviants may be imagined as belonging to an ethnic group (e.g. Uyghurs in contemporary China), or imagined to interact regularly with threatening outgroups and other deviants (e.g. East Germans with regular visitors from West Germans during the Cold War). Individuals whose ascribed identity fits such images are expected to be disloyal more than those who do not, and may have to prove their loyalty to avoid being perceived as disloyal, even though privately they performed no disloyal actions. By the same token, those whose ascribed identity does not at all match a traitor profile are above suspicion, and may find it easier to accrue "deviance credit" that precludes perceptions of disloyalty and labeling (see Abrams et al. 2018).

Finally, labels communicate an individual's affiliation with an outgroup, from peer-to-peer accusations and slander on a wall to official interrogations, court trials and sentences by political authorities. As for criminal delinquency (e.g. Braithwaite 1989; Farrington and Murray 2014; Sherman 1993; Tyler and Huo 2002), who is labeled as disloyal and under what conditions is key to understanding the effects of the label. When authorities officially label deviants, they reaffirm their traitor profile vis-a-vis other group members, even if they do not affect changes in private loyalty. By the same token, unofficial labeling by 'ordinary' group members is tantamount to 'participatory repression' (see Bergemann 2017), and conversely, not perceiving or labeling deviants constitutes a challenge to authorities. The labeled may be glorified or vilified (Åkerström 1991), respectively elevating their social status in opposition to authorities (see Sherman 1993; Tyler and Huo 2002), or ostracizing them from the group. They might be shamed into repention in compliance with a label (see Braithwaite 1989), or fail to do so when stigmatization prompts them to re-consider and renounce their group membership (see Matza 2010). Therefore, a key factor in determining the consequences of a label pertains to the opportunities it leaves for the accused to prove their loyalty to ingroup members. Conceptually, I distinguish between three effects of labeling on private loyalty:

1. Shift allegiances from loyalty to disloyalty and vice-versa, e.g. when the label alters the importance of a group for the self-perception of the labeled, or unilaterally withdraws their group membership rights.
2. Reinforce prior behavior, e.g. when perceived loyalty merely reaffirms private loyalty.
3. No effect, e.g. when imagined disloyalty does not align with perceptions of behavior.

To summarize, private, perceived and imagined facets of loyalty are invoked to justify the official and

[1] See Raybeck (1991) for a related discussion of "hard" deviance that threatens social order more generally.

unofficial labeling of political deviants. Which facets are invoked impacts the effects of the label on private behavior, either reinforcing prior loyalties or eliciting allegiance shifts. In the next section, I turn to the methodological challenges of measuring facets of loyalty and labeling.

3. Methods and Data

My discussion is based on a preliminary coding of records on 319 surveillance victims from the *Stasi archives* between 1961 and 1989. The purposeful three-stage sample was drawn from 44km of records at the main MfS headquarters in Berlin, and is designed to maximize files on surveillance cases where individuals were labeled and observed afterwards. First, the database was queried to obtain a list of cases that were either completed (archived by the *Stasi*) or involved a suspicion of defection. Second, I pre-selected 453 files on GDR citizens suspected of defection based on archivist file descriptions (e.g. excluding travel authorizations and foreigner surveillance). And third, I selected 6308 relevant pages across individual cases for systematic coding (e.g. excluding duplicate information), based on a review of the classified material at the archives in Berlin between 2019 and 2022. See Appendix B.I for details on sampling and data construction procedures.

Table 2.1 presents a nominal classification of behavior that motivated the *initial* reporting of suspicions. 37% of individuals in the sample were surveilled for quotidian Western contacts or delinquency (matching the Stasi's traitor profile), but their deviance was initially not exceptional, and required disambiguation by the state security case worker. Examples include company managers with friendly relations to Western business partners, and adolescent soldiers who listened to Gothic music associated with Western culture. 38% of the suspects were surveilled but not officially labeled, even though such labeling was over-sampled. Yet, many of those who were not targeted by state officials could be unofficially perceived and labeled by other group members. Private, perceived and imagined disloyalty change as security operations evolve. Individuals may initially be surveilled because their family connections in the West invoke suspicions of illegal emigration; and the informants who are subsequently prompted for information perceive them as disloyal for regime-critical statements; but after interrogating the suspect security agents may disagree with such perceptions, finding instead that the suspects private disloyalty was tolerable in the first place. A central assumption in this paper is that any accusation of disloyalty constitutes a label, and I suggest that each label may affect private loyalty as long as the labeled are aware of the accusation.

I use a constructivist, grounded approach to the archival data (see Glaser and Strauss 1967; Sebastian 2019): rather than taking the documents that are produced by *Stasi* case workers at face value, I treat them as *accounts* that serve to justify their behavior, given appropriate rules in the context of the "social world" in which they and the suspects they surveil operate (see Garfinkel 1967; Livingston 1987). In a sociological study that draws on biographical interviews with 72 Stasi employees, Krähnke et al. (2017, 211) characterized them as "intrinsically motivated and routinely disciplining each other, [and] felt strongly connected to the organization that expected total dedication and unconditional loyalty [from them]". Security officials were pushed to align their self-image with the "Checkist habitus of egalitarian Elitism" (Krähnke et al. 2017, 231). Deviance, ranging from loquacity and extramarital relations to Western contacts of family members, was controlled by superiors and through self-discipline, thus enabling ideological integration with the governing *SED*, as well as the social isolation of employees from the rest of the population, while membership was rewarded with the status of a "comrade of the first category", access to prestigious positions, and material benefits (Krähnke et al. 2017, 231-235). The social world of the MfS thus did not tolerate perceived disloyalty, and labeling within the MfS arguably increased private loyalties more than disloyalty, notwithstanding generational conflicts between the 'old guard' who embodied the Checkist mentality, and younger generations born in the 1950s and

60s who had to be socialized and disciplined into it (Krähnke et al. 2017, 201,240-241).

Table 2.1. Sample Description

Political Deviance	**Labeled**	**Not Labeled**	**Died**	**Sum**
Enemy-Informing/Sabotage	44	31	2	77
Exit	83	40	0	123
Voice	23	10	0	33
Western Contact/Delinquency	47	38	1	86
Obs. Individuals	197	119	3	319

Note: Behavior classification based on initial state security suspicions, given in 'opening' reports that initiate surveillance operations. Label counts are based on a preliminary classification of documents, not including unofficial labeling by actors other than security officials. Three suspects died during the investigation before they were officially labeled.

.

Imagined disloyalty is consistently observable through the archives, to the extent that MfS employees shared a traitor profile that hardly changed during the studied time period. Perhaps more than most East Germans, state security officials observed deviants who conspired with Western supporters against the SED regime (see Grimmer 2003; Glaeser 2011, 485), and justified their loyalty to the organization with that threat (Krähnke et al. 2017, 234). MfS reporting of political deviance fit into the epistemic logic that Western enemies were undermining Socialism 'from within', whether the deviants were aware of their relationship to these destabilizing forces or not. The "discursive culture of Stasi (and [state] party) prevented any systematic elaboration of critical insights", even when it was apparent to officials that the private loyalties of East Germans were misperceived (Glaeser 2011, 486). This practice was aided by a standardized pseudo-scientific grammar and fixed procedural routines which the employees were trained to follow (see Ludwig 2008; Pappert 2008).

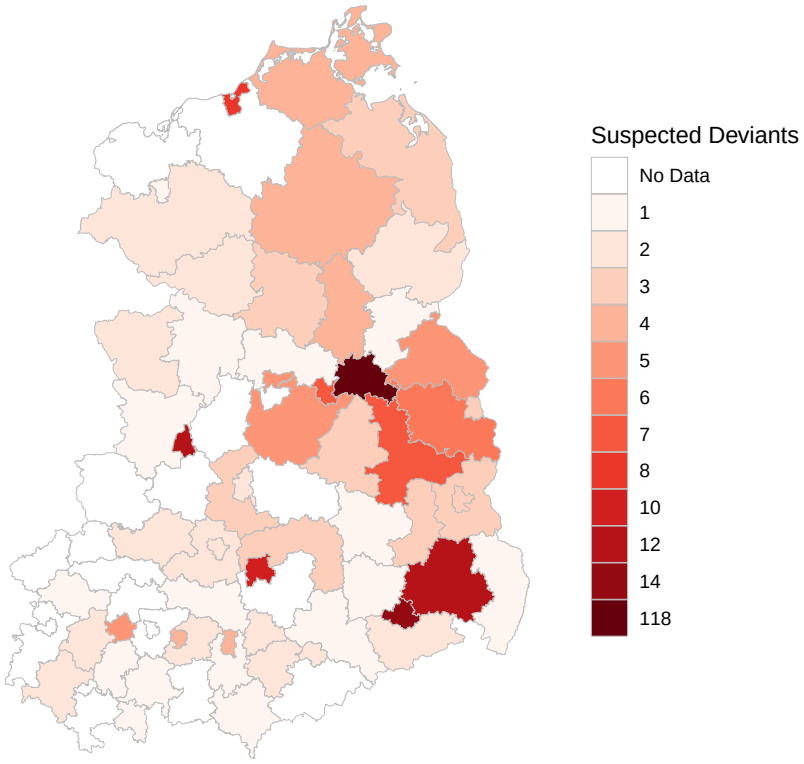
However, one caveat of relying entirely on the *Stasi* headquarter archives is that reliable suspect- and third-party accounts are underrepresented (see Halbrock 2015, 75; Richter 2015), and this limits the extent to which their private loyalty can be observed. Yet some document types (e.g. letter copies) are more reliable than others (e.g. informant reports), and it is usually possible to distinguish between an MfS officials' loyalty perception and private behavior that occurred with certainty. Notably, the Stasi selectively reported on how the neighbors, colleagues or relatives of a suspect perceived their loyalty. A second caveat relates to data coverage: an individual may have been surveilled once in the 1960s and again in the 1980s, but not all in the interim; and some Stasi records were destroyed in the wake of the revolution. For 236 of the sampled individuals, archivists in Berlin provided excerpts of the MfS database on surveilled individuals, which I use to check for missing observations, particularly to identify recurring disloyalty.

Grounding my analysis in the routines of the Stasi allows me to identify the context and sequencing of labeling and loyalty. First, I sort each document into a stage of surveillance operations for which it was produced. Stasi officials account for different activities in these stages, which influence how they perceive and label disloyalty:

- 'Opening': accounts for beginning an "operative person control" or an "operative procedure", giving reasons to investigate a suspected deviant.

- 'Intermediary': accounts for continuing operations to determine the loyalty of suspects, giving reasons to label them and recommend appropriate control measures.

Figure 2.2. Spatial Distribution of Cases



Note: Residence of surveilled East Germans in the sample from the *Stasi Archives*, aggregated to the ADM2-level (current borders). The sample from the former main headquarters of the *Stasi* is heavily biased towards defection in East Berlin (n=118), seeing as other areas such as Dresden (n=14), Magdeburg (n=12), and Leipzig (n=10) host separate archives that correspond to the regional administrations of the MfS, which could not be feasibly included in the sampling strategy.

.

- 'Closure': accounts for concluding an operation, which meant that the file was archived and the individual no longer the focus of official surveillance operations.

Second, I identify indexical expressions for labeling and loyalty in the documents (Garfinkel 1967), based on whose loyalty is accounted for, by whom, and how. Consider the following excerpt from the *closure* report on *Deviant Defector* case from Section 1,[2] which includes the coding of:

- magenta: imagined disloyalty (individualism being stereotypical for defectors),
- orange: perceived loyalty (implicit in family rejection of defector image),
- magenta: perceived disloyalty (rejection of opportunity to make personal sacrifice), and
- red: private disloyalty (neglecting and leaving military service).

[2]Note that I use descriptive labels to protect the anonymity of the individuals whose original files I reviewed, such as *Deviant Defector*.

> "After his marriage to the witness on 15.5.1971 and moving in with her parents in [...], due to financial and material considerations in his free time more and more working in the individual livestock farming of his parents-in-law and following various hobbies, the accused, after his reassignment as platoon leader, got into a conflict with the fulfillment of related higher official duties. As a consequence, and due to the petit bourgeois and Western influence by his parents-in-law he fulfilled his duties with increasing listlessness and considered, despite his promotion to master sergeant at the end of 1971, an early discharge from the border guard and taking up a job in agriculture." (BArch, MfS, GH, 17/79, vol. 4, p. 332; see Figure B.4 for the original excerpt)

First, *Deviant Defector* and his in-laws are matched with the traitor profile of the *Stasi*: the "Western influence" ascribed to the parents-in-law indicates a proclivity to consume Western cultural goods, and "petit bourgeois" a deviation from the ideologically preferred working class, presumably due to their agricultural business being relatively autonomous from state-owned production ("individual"). By comparison, the increasing work hours that *Deviant Defector* experienced after his promotion did not require justification in the closure report. Rather, they presented a foregone opportunity to reap status rewards by increasing his loyalty. As such, he fits the profile of a deviant whose disloyalty is due to Western influences. Second, the suspect's family and military superiors had different perceptions of *Deviant Defector*'s loyalty: the latter perceived a lack of military discipline as disloyalty, while the former supposedly facilitated it. Though there is no explicit account of the family's perception, their alleged 'influence' implies that they did not perceive a neglect of military service as disloyalty. And finally, *Deviant Defector*'s level of commitment to military service, as well as his preparation of enemy-informing and eventually desertion indicate his private disloyalty. However, he only started preparing his desertion after being labeled.

From a methodological viewpoint, drawing on three facets of loyalty permits a more nuanced assessment of an account's validity and reliability, compared to merely analyzing a single facet. As such, the approach sacrifices external for internal validity, compared to analyses that take political deviance as reported in aggregate statistics by political authorities at face value (e.g. Steinert 2022; see Balcells and Sullivan 2018; Becker 2017). First, since the private loyalty of suspects is frequently uncertain, state security officials reference perceived and imagined disloyalty to justify their actions. Perceptions are reported retroactively, based on prior investigations and reporting, to present consistent accounts. By the same token, other group members are associated with either the deviant suspect or the state as a function of their defiance or compliance with attempts to control political deviants. MfS employees invoked the traitor image that they were supposed to internalize in their reporting, and it is at times possible to discern the acceptance of this image among other group members.

Second, the framework permits an analysis of competing accounts. Sometimes these reflect inconsistencies between MfS reporting at different procedural stages. In the *Deviant Defector* case for instance, earlier reports give more elaborate explanations for the increased pressure on the suspect's border guard unit to qualify disloyalty perceptions, while the official sentencing document (which post-dates the section cited above) omits those pressures entirely to declare the suspect's personal problems as "self-inflicted". But the account that is given in MfS reporting could also differ from suspect accounts, as given in a letter or (with some loss in reliability) interrogation statement. In the case of *Deviant Defector*, some of the informants who provided the requisite information may have been unreliable, as he notes in a letter to his wife sent from West Germany:

> "Believe me, that the thing that they told you [about infidelities prior to the defection], is not true. I was dancing a lot, but I never had intimate relations with other women except

> for you. I knew they would tell you many fairy tales. The [informant] is the greatest liar there is. I gave him back his microphones a year ago [...] Now he has lost them himself, don't give him a penny, there is no way for him to prove that I have his microphones."[3] (BArch, MfS, GH, 17/79, vol. 2, p. 17)

In the following section, I describe the traitor profile of the MfS and the cases I selected for analysis. I then discuss how the interplay between imagined, perceived, and private loyalty affects official and unofficial labeling, and in turn, how these labels impacted the private loyalty of suspects.

4. Labeling Defectors in the GDR (1961-1989)

With post-Stalinism and the closure of borders to West Germany in 1961, state security relied increasingly on covert, preventive repression (Pfaff 2001; Pollack and Rink 1997). Their defector image was based on the notion of Western 'political-ideological diversion': the receptivity of the population to political positions that contradicted the Marxist-Leninist political order embraced by the SED party state. The expansion of the state security apparatus was justified by framing economic and social issues as diversion tactics by the enemy (Gieseke 2014, 51-59). Thus, a genuine political deviant was exposed to diversion and therefore suspected of disloyalty (third facet), failed to resist its lure by forming a 'negative-inimical attitude' that perceptually reduced their personal sacrifice for the GDR (second facet), and eventually expressed that attitude in behavior that threatened the security, unity, or socio-economic well-being of GDR citizens (first facet).

Exceptional forms of political deviance that were criminalized included regularly selling information to Western intelligence, sabotaging state-owned company production or scientific development, participating in collective mobilization for 'agitation' against the state, and emigrating to the West without government authorization (see Raschka 2001). The need to prevent such deviance meant that security particularly targeted those who were valuable to Western intelligence (e.g. state employees) and those who were vulnerable to Western influence. Exposure to such influence occurred through Western contacts, media and cultural products, including the consumption of Western television programs, newspapers, literature and music (Rembold 2003). As a result, labeling particularly targeted church members, people who expressed emigration intentions, and sexual deviants (e.g. homosexuals) who were marginalized in the East (Glaeser 2011, 487-488).

The politicization of deviance in relation to Western influence meant that defector labels were, in principle, reserved for those deviants who *intended* to harm the political order of the GDR in a way that benefited its rivals. For instance, Western contacts were tolerated in the absence of ideological diversion; obstruction at manufacturing companies was tolerated when the culprit had no discernible motivation to sabotage the economy; and regime-criticism was tolerable while it could not be instrumentalized for Western propaganda. Thus, even those who were perceived as harboring 'negative-inimical attitudes' and engaged in disloyal behavior could challenge defector labels when they construed Western contacts as innocent, misdemeanor at work as incidental, or regime-critique as conform with state party ideology. Even 'illegal emigration' could be tolerable to authorities when a defector threatened political order more from within than they could from without (see Passens 2012). But by the same token, state security could construe loyal suspects as defectors: regime-critical conversations or visits of Western family were imagined to entail enemy-informing, and unsatisfying work performance could be perceived as deliberate. It follows that the relationship between labeling and loyalty is dynamic and co-constituted between the labeling agent and the labeled. For analytical purposes, I view it as a negotiation process

[3] Of course even in considering this as a suspect's account, the suspect, too, is merely accounting for deviant behavior prior to his defection. But he is doing so differently than the Stasi.

between authorities and labeled suspects over the latter's group membership that maps onto the three procedural stages of *Stasi* surveillance operations:

1. Initial label: suspects engage in perceived disloyalty, authorities register an investigation for them
2. Label confrontation: authorities label and suspects react to being labeled
3. Settlement: the suspect's loyalty or disloyalty is either unilaterally designated by authorities or mutually accepted

In the following Section 4.1, I discuss under what conditions deviance was labeled and how labels were received by group members. And in Section 4.2 I turn to the effects that these conditional labels had on 15 individual suspects. As shown in Table 2.2, these cases were selected to represent two negotiation outcomes across four different behavioral categories: those who settled on remaining loyal to the GDR in the eyes of the Stasi, and those who continued or intensified their disloyalty. *Repentant Spy* and *Interrupted Fugitives* were both labeled for enemy-informing, but where the former was a state agent who shifted from disloyalty to loyalty, the latter were related to a West German professor and criminally prosecuted for more disloyal behaviors than they committed. *Exits Denied* were pressured to remain in the GDR by the *Stasi*, while *Underground Exit* maintained his disloyalty and emigrated. *Academic Reformers* and *Persistent Activists* includes multiple individuals who responded differently to being labeled for regime-critical activities: among the former there were some who proved their loyalty, while the latter repeatedly engaged in disloyalty. Finally, *Abusive Escapist* and *Brother's Keeper* are both soldiers whose social deviance the *Stasi* associated with potential disloyalty: one redeemed himself after threatening to kill his wife and flee to the West, the other refused to view his brother's illegal emigration as disloyalty.

Table 2.2. Paired Comparison

	Loyal	**Disloyal**
Enemy-Informing	*Repentant Spy* (#123, MfS HA IX 24419)	*Interrupted Fugitives* (#1-#2, MfS GH 107/80)
Emigration ('Exit')	*Exits Denied* (#136-#137, MfS HA VII 305)	*Underground Exit* (#239, MfS HA XVIII 25480)
Regime-Criticism ('Voice')	*Academic Reformers* (#64-#68, MfS AOP 16183/81)	*Persistent Activists* (#147-#148, MfS HA XX 350)
Delinquency	*Abusive Escapist* (#5, MfS HA I 14995)	*Brother's Keeper* (#14, MfS HA I 14995)

Note: Behavior classification of 15 individuals based on MfS suspicions and judgement, ranging from most exceptional (enemy-informing) to least exceptional (delinquency). All suspects were labeled, but some were ultimately found privately loyal while others were found to be disloyal.

4.1. Acceptable Labeling of Deviance

State security officials labeled deviance as disloyalty when (1) they could express confidence in the perceptions of informants, (2) legally prosecute private disloyalty, and (3) they could ascribe the suspect an identity that matches imagined disloyalty. In turn for labeled suspects and 'label-free' audiences alike, labels were appropriate when the group member (1) recognized the authority of the labeler, (2) found the modalities of the label tolerable, and (3) considered the behavior for which the label was applied exceptional.

State Security: Confidence in Reported Perceptions

Assuming exceptional political disloyalty was not observed directly by state officials,[4] case workers usually relied on informants. The absence of disloyalty perceptions is not sufficient to clear suspects, but their presence justifies continued operations. For instance, after an investigation in the neighborhood of the suspect 'Underground Exit', it was found that "his political attitude could not be concretely determined, as he did not express his views on this question to the tenants" (BArch, MfS, HA XVIII, 25480, p. 3). By contrast, in the opening report of the *Deviant Defector* case the informant's disloyalty perceptions are sufficiently reliable: "[...] is an unofficial informant experienced in investigative work, who until now has always reported truthfully and objectively" (BArch, MfS, GH, 17/79, vol. 1, p. 82).[5]

Confidence in the disloyalty of suspects justifies the use of operative measures to 'reclaim' their loyalty to the GDR. This was common for emigration cases such as that of the *Exits Denied* couple, whose repeated applications for emigration presented reliable proof of negative-enemical attitudes to Stasi officials, including "connections to enemy organizations" and "plans to use publication outlets in the FRG" to "discriminate against [GDR] state organs"(BArch, MfS, HA VII, 305, p. 120). Operative measures include means to disambiguate or obtain additional evidence of disloyal behavior (e.g. wiretaps, observation, provocation of exceptional deviance), which is commonly expressed in the 'goals' section of opening reports:

> "The operative handling of both individuals aims to obtain [...] proof for the suspicion of imminent punishable actions [...]; proof that both of them, in their realization of subversive activities, are establishing contact to organizations, centers or persons in the FRG who are fighting the GDR or who are organizing negative-inimical activities [...], or that they are acting on instructions of such institutions [...]." (BArch, MfS, HA VII, 305, p. 101)

Trustworthy informants who reported on exceptional disloyalty thus instilled confidence in the appropriateness of operations to preempt or terminate political deviance. However, given that state security

[4] Exceptional disloyalty that was reported through official sources was considered reliable, e.g. when other Socialist security agencies pointed to a spy, or the People's Police (*Volkspolizei*) preempted an illegal border-crossing. Nevertheless, involvement of the People's Police in state security affairs was regarded with suspicion or even open critique, as exemplified by the following statement regarding the case of *Exits Denied*: "the criminal investigation department withdrew the identity cards of the couple [Exits Denied] and told them that as of 19.09.1978 they would be excluded from identity- and visa-free travel for a year. This measure was not coordinated with either the local MfS department or the department for internal affairs [...], and led to further recidivism in the reactions of [Exits Denied]" (BArch, MfS, HA VII, 305, p. 110).

[5] Given the scarcity and utility of 'reliable' informants for the justification of operations, their supervision was at times contested between different state security departments, which could affect how political deviance was labeled. For example in 1982, a case worker in the department for domestic espionage (HA II) complained about the department for national economy surveillance (HA XVIII) in a letter to his department head: "[...] noting that the HA II already received the unofficial informant 'P.' [...], they [HA XVIII] assert a claim on the couple 'R.' [for recruitment as informants]. Apparently for tactical reasons, the HA XVIII over-emphasizes open questions regarding substance abuse and the disorderly intimate life [of the surveilled suspect] to prevent the conclusion of the current operational phase, and to affect a potential distancing of the HA II from handling the operation henceforth." (BArch, MfS, HA II, 32130, p. 70).

had to give the appearance of adhering to criminal laws in the GDR (see Engelmann and Joestel 2016, 162-163; Booß 2018), they could not use unofficial denunciations to justify official defector labels:

> "The most important evidence is unofficial and was compiled by [the] unofficial informant [...] information was repeatedly confirmed through use of operative technologies [...]. It is clear that the officially usable evidence is still insufficient." (BArch, MfS, AOP, 16183/81, vol. 1, p. 186)

State Security: Prosecuting Private Loyalty

Stasi agents resolved conflicts between perceived disloyalty and the inappropriateness to label it officially by *unofficially* labeling political deviance. Operative measures against suspects drew on third-party audiences to label them, the notion being that their peers could exert a 'positive influence' that leads to demonstrations of loyalty. Such labels were unofficial because state officials did not confront the suspect, but rather instructed collaborators at workplaces and sometimes family members to do so. The role of the unofficial source was acknowledged by case workers, e.g. a desired behavioral change for *Exits Denied* was partly attributed to "[...] targeted influencing by the informant [...] of the wife" (BArch, MfS, HA VII, 305, p. 113-114)

In practice, victims of Stasi repression experienced these labels when they were reprimanded, demoted or derogated by superiors, friends, or family members; but also when they were denied spots in higher education despite a competitive background; or denied permission to travel outside the GDR (Raschka 2001, 29-30; Passens 2012, 167). These 'disciplinary measures' revealed to suspects that they had been labeled, given a shared understanding that they are used to punish disloyalty. As a result they might suspect state security involvement, but they could not obtain official confirmation for it. For example after unofficially labeling her through disciplinary measures (BArch, MfS, HA XX, 350, p. 44-46), the closure report for *Persistent Activist* notes:

> "In multiple petitions to the Ministry of the Interior and the Ministry for State Security, she objected against the alleged[sic!] inhibitions of her professional development and the travel ban against her. This was used to have multiple discussions with her, with the goal to discipline her and preempt politically-negative or hostile activities." (BArch, MfS, HA XX, 350, p. 103)

Notably, labeling can involve positive incentives and 'reintegrative measures', including monetary aid, improved accommodation, or promotions (on the phased combination of such measures for emigration applicants, see Passens 2012). These measures may have addressed the grievances of the suspect, but for state security their purpose was to elicit a demonstration of loyalty:

> "Unofficially it was arranged that the parents of [Exits Denied] intensively influence their withdrawal of the emigration requests, above all through offering financial assistance during their visits." (BArch, MfS, HA VII, 305, p. 110)

Unofficial labeling reduced ambiguity in loyalty perceptions, even when the suspect continued disloyalty without changing behavior (e.g., maintained or resubmitted an emigration request). Continued political deviance afterwards implied to the *Stasi* that disloyalty is 'politically-ideologically' motivated rather than incidental, seeing as the suspect refuses to demonstrate loyalty in spite of social ties and material rewards for doing so.

To state security, legally admissible evidence was key to concluding operations on political deviants who were not amenable to change their behavior through unofficial means. The most truthful claim to disloyal behavior could be made through observations of exceptional deviance by state officials, particularly if it involved Western media, humanitarian organizations, or intelligence services. Such cases received significant attention by multiple MfS employees, as they justify the activities of the organization.[6] A confession by the suspect was followed by criminal procedures, but even when there was officially usable information about wrongdoing, state security did not necessarily consider it appropriate to officially label a suspect, contingent on the reaction they anticipated from other suspects and Western rivals. For instance to contain the emigration movement, overtly "provocative demeanor" was officially criminalized and noted in reports, but rarely officially labeled (Passens 2012, 169,183).

State security agents conducted preventive interrogations of political deviants without initiating legal proceedings. Instrumental motivations for these interrogations included the prevention of imminent political deviance, recruitment of suspects as unofficial collaborators, and coercion of confessions that could be used for official prosecution (Raschka 2001, 30-35). Interrogations did not necessarily attach a visible label to political deviants, as suspects at times attempted to avoid the stigma. An informant reported about the reaction of a suspect to his interrogation:

> "[...] he had asked me, that was the first thing he said before he told me about it, that I should not reveal the fact of his interrogation to anybody, even my wife, as he wants to avoid any emergence of rumors [...]" (BArch, MfS, AOP, 16183/81, vol. 5, p. 68-69)

The line between unofficial and official labeling is blurred when it comes to preventive interrogations: the label that the interrogation attaches to the deviant, if any, carries political authority, but it is neither as public nor as permanent as an official criminalization. The audiences that could learn of interrogations include the same family members and colleagues that the Stasi would use to issue unofficial labels as well. For instance, a different suspect in the *Academic Reformers* case reportedly found that his colleagues were aware of the interrogations (BArch, MfS, AOP, 16183/81, vol. 5, p. 144). At minimum, irrespective of their visibility, interrogations are a certain signal to suspects that they are risking an official defector label, and state security used the threat of criminal prosecution where they sensed that suspects might be afraid of incarceration. This very much depended on the support that the suspect might receive from audiences of the label though, and the expectation that official prosecution might invite 'backlash' was sufficient to deem it inappropriate:

> "[...] at the time it appeared that the use of criminal prosecution against [Exits Denied] would have been inexpedient, particularly given [her] pregnancy. Criminal prosecution against her could have triggered an enemy smear campaign [...] The interrogation was conducted by department IX/2 [for official investigations] following a conspiratorial police escort [...]. The interrogation also led to the unsettling of [Exits Denied] about the still possible occurrence of legal consequences for him and his wife" (BArch, MfS, HA VII, 305, p. 112)

Official labels lead to more widespread (though not necessarily more meaningful) stigmatization of defectors, and persisted longer compared to preventive interrogations and purely unofficial labels.

[6] For instance the 'Interrupted Fugitives' who were caught attempting to leave the GDR could be construed as enemy-informants and saboteurs in the process of the investigation, which was documented across 96 volumes, each comprised of hundreds of pages. By comparison the *Underground Exit* investigation spanned 30 pages, as the MfS could merely document the 'negative-enimical attitude' and Western contacts of the suspect, but found no evidence of exceptional political deviance.

Though the public was typically excluded from court trials, sentences could be publicly announced to deter others from political disloyalty, and former political prisoners were treated as defectors by the public even after their release (Raschka 2001, 85,124-127). Moreover, state security continued to surveil political prisoners until they demonstrably re-integrated into society, mostly through prison reports, mail control, and the continued use of informants:

> "The prisoner expressed a desire to take residence with his wife [...]. This wish [...] is supported by prison leadership, as the personal connection between the prisoner and his wife during his incarceration allowed for a positive organization of the educational process. After his release from prison, [Deviant Defector] wants to take a job in an agricultural business. In sum it can be assessed that the educational process of the prisoner has developed positively [...]. It can be expected that he will conform with the laws of our state in the future." (BArch, MfS, GH, 17/79, vol. 11, p. 37)

State Security: Volition and Closing Operations

Before they could cease attempts at affecting change in behavior, state security had to reason that a suspect no longer intends to engage in political deviance. For a deviant to demonstrate loyalty requires that they comply with expectations of state security officials, which generally includes admissions of guilt, public declarations of repentance, and denunciations of other deviants.[7] Repentance was formally expressed in 'statements' signed by the suspect as proof of repentance and future loyalty. All suspects in the *Academic Reformers* case, for instance, had to write such a statement as a public record of the intentions behind their disloyalty, in this case for co-workers and fellow SED-party members to recognize the wrongdoers among them (BArch, MfS, AOP, 16183/81, vol. 6).

Such symbolic demonstrations of compliance were necessary, but not sufficient to satisfy state security agents. Their surveillance had to permit the interpretation that the suspect is 'distancing' from political deviance and deviants, regularly interacting with loyal citizens, and adequately performing at their workplace. By contrast, 'persistent' negative-enemical attitudes could lead to repeated labeling. Particularly activists might be intimidated ahead of public events to preclude their mobilization, and any contact to political deviants could lead to a re-opening of investigations, as demonstrated by *Persistent Activist*:

> "After her release in June 1980 it was observed that she again made contact with politically negative circles in the capital of the GDR and interacts with them [...] These are predominantly Polish citizens that live in Westberlin and belong to an anti-socialist circle with active connections to similar circles in the Polish People's Republic" (BArch, MfS, HA XX, 350, p. 53)

Official labels entailed the most severe sanctions, which could include imprisonment and (until 1987) the death penalty. But the most intimidating official labels were not necessarily the most consequential for allegiance shifts. From a labeling perspective, an observable shift from disloyalty towards loyalty would be contingent on the suspect's acceptance of the label's veracity, which may be independent of, or even inversely related to, the sanctioning modalities attached to the application of the label. This leads to the question of how subordinate group members view the veracity of labels.

[7] Recruitment for unofficial informing was desirable for the MfS, not only to identify and close additional operations in the future, but also as a demonstration of loyalty by the former suspect, and a control mechanism to the extent that informing implied regular reporting to a case officer.

Group Members: Authority
The appropriateness of a label depends on the labeled's recognition of the labeler's superordination, given their formal roles in the context of the interaction where labeling takes place (Beetham 2002; Simmel 1964). By default, GDR citizens are subordinate to those in positions of state power, but the extent to which this subordination is recognized depends on the social status hierarchy between labeler and labeled, as well as on the trust that is placed in the morality of the former (see Tyler and Huo 2002. One can expect more subordination to high-ranking state security officials among deviants in their own ranks than by an average citizen who has not been raised into the class of the socialist elite (see Gieseke 1999; Krähnke et al. 2017), and a junior soldier might be less confident in asserting the innocence of their behavior to a security official than a doctor who knows they are being coveted by the state for their expertise.

In the *Stasi* records, authority recognition was most frequently observed through interrogations, where characterizations of the suspect's authority recognition ranged from "unsettled" or "understanding and self-critical" to "contentious" or "hateful against the MfS" (BArch, MfS, HA VII, 305, p. 165; MfS, HA I, 14995, p. 138; MfS, GH, 16/77, vol. 4, p. 183). As reported by an informant, one of the *Academic Reformers* portrayed his MfS interrogator more as a discussion partner, "selected based on the psychological characteristics of the individuals", and suggested that "he rarely had such a professional discussion about the purpose of philosophy as during the interrogations, and apparently this would only be possible in two institutions, either in illegal circles or at the Ministry for State Security" (BArch, MfS, AOP, 16183/81, vol. 5, p. 57). Formal authority may therefore not elicit compliance or even prompt defiance, as in this paraphrasing of an emigration request: "He guarantees peace to nobody unless he is let go [...] leaves it to the intelligence of certain authorities which consequences should ensue if he is forced to stay [...] 'I am demanding immediate exit from your state, in which I have no place as an enemy' " (BArch, MfS, HA XX/AKG, 6215, p. 2-3).

Group Members: Label Modality
Even if group members attributed authority to the labeler, the modalities of the label might have been inappropriate. Modality may refer to the repertoire of 'violence' that is being applied (see Gutiérrez-Sanín and Wood 2017), including overt surveillance that merely threaten subjects with a label, derogation, denunciation, interrogation, reprimand, arrest and imprisonment. The acceptance of the labeling process is contingent on a perceptibly 'fair' treatment of the labeled, which not only increases long-term trust in authorities (see above), but also directly increases compliance in a particular instance (see Tyler and Huo 2002; Wood 2003). In the GDR next to the inappropriateness of legal proceedings and prison conditions, unofficial labeling could be deemed inappropriate. For example in one case, when a worker was denounced to the People's Police for posting "we strike for higher loans" on a board, it was the informer who was 'counter'-labeled as a "fink" and a member of the local party leadership "objected strongly against the *manner* of the denunciation" (BStU, MfS, BV Rostock, AKG, 559, p. 76-77, cited in Halbrock 2015, 144-145).

In other settings, suspects might public assert their acceptance of official conduct, as when an informant claims to have heard from *Academic Reformer* that "the manner of the interrogation was very correct, there are very qualified people working there who do an excellent job". A necessary condition for such an airy description may have been that the suspect was not worried about the negative consequences of the label, given reassurances by audiences:

> "[...] a spontaneous solidarity had developed at his workplace [...] people are offering potential support, e.g. next to moral also material, such that they say if you need a lawyer, we will of course collect money and you do not need to worry about that etc [...] This

> spontaneous solidarity apparently strongly impresses *Academic Reformer* and gives him without doubt backing and psychological security [...]" (BArch, MfS, AOP, 16183/81, vol. 5, p. 57-59)

Perhaps anticipating that there was little acceptance for the mistreatment of alleged deviants, security officials tended to report that suspects were brought in for interrogations or arrested 'without any resistance'. In turn, motivated disloyalty was not deterred by such arrests, which could be seen as a means to affect emigration, given that the FRG paid for the release of political prisoners into West Germany (Raschka 2001, 121-122). This could reach the point where the inappropriateness of security practices makes them ineffective among those who experienced repression. To account for the persistent misleading of state security about his plans to kidnap a state official, a suspect mentioned in Section 1 wanted "to demonstrate to the investigators and later the court how quickly a man can be suspected of criminal actions merely because he is interested in particular problems [regarding the operation of Western intelligence]" (BArch, MfS, GH, 337/79, vol. 8, p.45).

Group Members: Normality

Finally, even a label that is procedurally inappropriate can be acceptable given that it is applied for 'abnormal' or exceptional behavior. When the traitor profile of authorities is shared by group members, even the labeled might accept it. For instance, an MfS employee who got demoted because his brother had submitted a request to emigrate found that associative label appropriate: "I was no longer a role model. Told that to others myself [laughs]. I myself demanded it from others." (cited in Krähnke et al. 2017, 46).

By contrast when the traitor profile is not shared by group members, labeling may be ignored or can lead to alienation. For example, in September 1961 an SED campaign to prevent GDR citizens from watching Western television failed: though loyal members of the Free German Youth (*FDJ*) publicly labeled those who installed suitable antennas on their rooftops, the campaign at best increased the concealment of reception and was abandoned (Geserick 1989, cited in Brücher 2000, 51; see Spiegel 1961). By the same token, when security officials prompt suspects about such indicators during interrogations, the accused could present their behavior as innocent:

> "Question: Where do you get your political information?
> Answer: I have great interest in world political events, and therefore inform myself through subscription of the 'Union', the broadcast of the German Democratic Republic as well as the 'Deutschlandfunk' [West German TV] [...] one only gets closest to the truth by following both commentaries and forms ones own position." (BArch, MfS, GH, 107/80, vol. 2, p. 15).

Overall, suspects accepted their label if they recognized the authority of the labeler, the way the label was applied as appropriate, and the behavior that they were labeled for as exceptional. In turn, third-party audiences may have applied unofficial labels when they found political deviance threatening (exceptional), feared to be labeled in the same manner (conditional on the modality of the label), or sought to protect the suspect from official criminalization (respecting authority). The efficacy of official and unofficial labels is then contingent on the sequencing of political deviance and its official and unofficial labeling. In the following, I draw on selected cases of labeled deviants to discuss how the acceptance of labeling affected negotiations over behavioral outcomes. While their criteria to judge the acceptance of labels differ, all participants of the labeling process attempt to have their perceptions recognized, and settle on a shared understanding that firmly positions the labeled among sufficiently loyal ingroups or disloyal outgroups.

4.2. Authority-Suspect Interactions

Repentant Spy and Interrupted Fugitives

Initial suspicion. In October 1982, an unofficial informant reports to the domestic espionage department of the MfS: an employee of the Western federal domestic intelligence agency (*BfV*) in Cologne requested that the informant delivers a message to 'Repentant Spy' (#123), a 58-year old employee at the GDR's cultural association whose main task was the recruitment of informants in the FRG. The message asks *Spy* to make contact with the employee, stating that his "friends in Cologne are expecting *Spy* in the FRG within two months", that he is "teasing" the *BfV* a little and "has to be brought in line a bit" (p. 17). This suggests to the MfS that the two are "acquainted", and that *Spy* was at least suspected to a recruitment attempt and is at worst about to leave the GDR for his Western employees, prompting an operation for enemy-informing. The MfS is *confident* in the perceived disloyalty of *Spy*, seeing as the denunciation matches MfS findings that *Spy* has West connections, is *vulnerable* to recruitment (used to cheat on his wife with partners in the West),[8] and that the *BfV*'s "great interest" in *Spy* is plausible given his high security clearance level (p. 18-20).

Similarly in August 1975, state security is confident that the 'Interrupted Fugitives' (#1, #2), respectively working as a physicist and as a medical doctor, are political deviants about to leave the GDR with their children: surveillance showed their exceptional attempt and failure to meet with the smuggler that would take them across the border; they are still in contact with #2's family of political deviants, her parents and sister had previously left the GDR, and her father was a known regime-critic himself. In contrast to *Repentant Spy* however, the couple's exceptional deviance was observed directly. The suspects account for their private disloyalty by another label prior to their arrest, when their daughter got denied a position in high school despite "good grades": "The main reason for my wife and I to move to the FRG is the separation of my wife from her parents [...] and that our children have few education opportunities in the GDR, seeing as they are children of the intelligentsia and [...] worker's children have preferred access to higher education spots, which are sparse" (vol. 2, p. 10).

Label confrontation. The MfS could not officially prove the disloyalty of *Spy* in the absence of exceptional deviance: *Spy* did not react to state security operations that tested his private loyalty, which involved the delivery of the message to see if he would make contact to the *BfV*. Though he did not, disloyalty perceptions were maintained as he "more and more reduce his contact to people in his area highlighting his bad health" (p. 18), which fits the traitor profile of someone planning to emigrate illegally. The head of department approves a preventive interrogation in April 1983, where *Spy* admits to his recruitment in 1979, but declares that he did not *intend* to be disloyal: he was afraid that West German intelligence would reveal his Fascist past, his illegal currency exchanges during his stay in Cologne, as well as an affair with his secretary (p. 43), and he accepted DM5,000 for "few" information that is "essentially public knowledge, really had no confidential character" (p. 33). *Spy* is repentant about failing his "trial by fire" and "cowardly" hoping that he could "evade a dangerous enemy" rather than coming forward voluntarily.

By contrast, the interrogation of *Interrupted Fugitives* was aimed at "proving to the accused" that their political deviance was driven by rival influences (p. 154). The father of #2, a physicist like her husband, was alleged to have connections to FRG intelligence and to have recruited #1 in 1965. The *Stasi* perceived family visits as a disguise for the transmission of "significant" information about microelectronics by #1, family gifts as payment for the "sabotaging" of the GDR's research, and family exchanges as a conspiracy, jointly organized by the *Interrupted Fugitives*. However, the account given of *Fugitive*'s reaction to their arrest suggests that they viewed their exit as normal and made no effort to

[8] To state security, this indicates blackmail by the *BfV* or that *Spy*'s intends to leave his wife and emigrate to the West, though the case worker notes that "on the outside both [*Spy* and his wife] have a harmonic marriage".

demonstrate loyalty, but made no mention of enemy-informing either: #1 expects that his "children will have better education opportunities in the FRG", admits and defends his consumption of Western television programs, criticizes several GDR policies, and emphasizes that they "still intend to get to the FRG" (vol. 2. p. 11-18).[9]

Settlement. After *Spy* signed multiple declarations that he is "not an enemy", feels "deeply rooted in Socialism", and is "committed to put right the damange" he had done (p. 323), the MfS decides to drop his prosecution, and conclude the operation in July 1983. Criminal procedures would be justified, but "the enemy might use *Spy*'s imprisonment for agitation against the GDR", and as his conduct suggests to the MfS that he will be loyal in the future, it would be acceptable to arrange for his early retirement (p. 46-47, 346-347).

By comparison, the *Fugitives* could not claim any deviance credit that would create a discrepancy between imagined and perceived disloyalty, nor did they assert their private loyalty. In 1977 both were sentenced to prison for attempted exit, enemy-informing and sabotage, which were presented as intentionally motivated by their disloyal upbringing and "hatred" of the GDR (p. 133). Once they admitted their intention to leave the GDR, their backgrounds were retrofit to paint a consistent picture of motivated political deviance, as their personal motives were inconsistent with the traitor profile of authorities. While state security records give little evidence of disloyalty after the imprisonment, prison surveillance points to suspicions of secret communication (vol. 87), suggesting continued private disloyalty.

Overall, authorities were confident in the guilt of *Spy*, but though he was initially perceived as disloyal and labeled officially, the MfS believed his disloyalty to be inadvertent, which did not fit their traitor profile. In turn, *Spy* accepted that his behavior was exceptional when confronted with the label, and demonstrated his loyalty. By contrast, *Fugitives* did not truly engage in enemy-informing, but their private exit intentions were aligned with the perception of disloyalty for that behavior, and their background fit the image of the 'spying saboteur' so well that the MfS decided to label them for that as well. Labeling reaffirmed the traitor profile of state security, but private loyalty was seemingly unaffected by it.

Exits Denied and Underground Exit

Initial suspicion. The couple *Exits Denied* were unofficially labeled for social deviance at their work- and living places prior to voicing their intention to exit (p. 123-127). The MfS perceives #137's (a preschool teacher) as loyal until the onset of her relationship with #136 (an assistant at a photographic laboratory), attributing her "negative-inimical attitude" to "his influence" (p. 127). The couple first appeared on the radar of the MfS in 1977 when the father of #136 denounced #137 for intending to leave the GDR with his daughter. The initial MfS account of confidence in the denunciation is missing from the file, but it was sufficient to warrant a preventive interrogation.

Similarly, *Underground Exit* was unofficially labeled for social deviance at the state-owned textile company where he worked in 1982 (p. 60), which according to an informant was followed by "alcohol consumption that [later] subsided and was obviously only connected to the reprimand" (p. 44). In contrast to *Exits Denied*, the MfS had already surveilled him since 1977 for his 'political underground' activities, with indicators for perceived disloyalty including "many Western connections" and membership in the student community of the protestant church (p. 1, 25). But despite his apparent grievances against

[9] In a phone conversation with a researcher in 2012, #2 stated that the allegations of enemy-informing and sabotage were invented by the MfS, and that the imminent motivation for the exit was the daughter's nonadmission (citation censored; referencing the source to this statement would violate the anonymity that I must grant the Stasi victim in this case).

the GDR, including "no motivation to work", "negative statements about comrade Erich Honecker", and attendance of a "goodbye party" for prospective emigrants (p. 7), they are insufficient for the *Stasi* to express confidence in their perception of disloyalty at the time: he "takes no clear political stance" (p. 1).

Label confrontation. The seemingly imminent exceptional political deviance of *Exits Denied* warranted an official, preventive interrogation by the MfS department on political underground activities (*HA XX*), during which the couple was coerced to distance themselves from crossing the border illegally. Following confrontation with the official label, the couple did not desist but submitted an emigration request in 1977, and #137 complained to the local government council that they were "treated unfairly" by the MfS during the interrogation (p. 120-121). #137 similarly viewed the interrogation as an unacceptable attempt by the MfS to prevent their marriage, began to "neglect her work" and "make negative statements" about the GDR, and lost her job after marrying #136 half a year later (p. 127).[10] When the emigration request was rejected, they submitted a letter of protest with language that is commonly used by emigration applicants to receive approval for their exit, threatening publicity and leaving the GDR illegally if their request is not granted (p. 107, 121-122),[11] and perceiving the dismissal of #137 as an "occupational ban" (p. 106). MfS reporting attributed these behaviors to the political-ideological diversion of the FRG, and alleged that the couple maintains West contacts with the intention to affect their emigration. In line with this perceived disloyalty, the objective was to officially prove the couple's agitation against the state and illegality of their behavior, as well as identifying potential 'political underground activities' and contacts (p. 128).

Similarly following unofficial labeling at work, *Underground Exit* submitted a request to emigrate to the FRG in February 1983, but did so by stating his intention to marry a citizen of the FRG. In November 1983, an informant reported that *Underground Exit* increasingly believed in the legality of his request,[12] yet still does not make negative-inimical statements in his neighborhood, while his "alcohol consumption declined again" and his "work intensity improved significantly" (p. 44), providing no indication of disloyalty.[13]

Settlement. After *Exits Denied* maintained their emigration requests and associated threats in repeated official 'talks' with state organs for several months, the couple withdrew it in June 1978, stating that their parent's employment status was threatened by the MfS should they refuse to comply, and "demand" the lifting of "alleged" occupational bans against them and their parents in exchange for dropping their attempts to publish their story in the West (p. 109, 162). Meanwhile, the MfS established that the two (unsuccessfully) concealed their contact to a distant relative in the FRG who only verbally supports their emigration (p. 107-109), and found that their attempts to contact state organs and publication outlets did not yield a supportive response (p. 115). Moreover, seeing as an informant reported that

[10]Notably, the MfS states that on 01.09.1977 the work contract of #137 "was terminated [...] in mutual agreement" (p. 127), but it is plausible that her superiors at work felt the need to demonstrate loyalty to the state and facilitated her dismissal. Moreover "after a disciplinary procedure", she was dismissed without notice from her position as a preschool teacher in January 1978.

[11]Here including: "renouncing their GDR citizenship", "sympathizing with Amnesty International", "accusing the GDR of violating the Helsinki Accords [invoking the freedom of movement prescriptions therein]", threatening to "publish their case in FRG media", and attaching their GDR identification cards to the letter of protest (p. 106, 121-122).

[12]This confirmed the *Stasi*'s suspicion that *Underground Exit* might react 'negatively' to the publication of the 'Madrid agreement' in September 1983, whereby such emigration through marriage could be interpreted as permissible (p. 44; see Hanisch 2012, 289).

[13]The increasing compliance of the suspect is disregarded in a later MfS assessment from March 1983, whereby he "has a bad work attitude since November 1983", which may be due to the case worker's access to additional information. Another possibility is that the department head who justified the approval of the exit in the same context overlooked the informant report, to ensure consistency with later accounts of repeated misdemeanor at work, which justified the conclusion that "all possibilities for influencing [him] have been exhausted" (p. 52).

Exits Denied intend to re-issue their request in anticipation of increasing tolerance for legal emigration, the security official does not believe that they actually intend to engage in exceptionally public protest or border-crossings (p. 109-110), which is shortly thereafter confirmed by the couple's re-issuing of their emigration request (p. 163). Rather than prosecuting the couple, the MfS intimidates #136 by threatening legal consequences in a separate interrogation, and instructs informants and parents to unofficially label #137 (p. 112-114).[14] In response, and possibly due to #137 giving birth to a child in the GDR and "apparently no longer supporting all the intentions of her husband", *Exits Denied* declared that they "made a mistake" with their request (p. 167), and the MfS concludes the operation by reinstating the two in desired occupations, providing a day care place for their child, and promising a larger flat.[15] Satisfied by informant reports about the couples repenting attitude, and the observation that they voted in communal elections, the case worker archives the operation, noting that an informant will gradually cease personal contact with #137 after ensuring *Exits Denied*'s permanent loyalty (p. 118).

Given a lack of access to the socially deviant friend circle of *Underground Exit*, as well as a regression of work performance, alcoholism, and a suspicion that he "is obviously homosexual and the 'mariage' is just a pretext", the MfS concludes in March 1984 that "there are no political-operative objections against an emigration" (p. 46, 52).[16] The department for political underground activities finds that *Underground Exit* "interacts with people who deal with problems of totalitarianism", was officially labeled during a visit in Prague where he held the FRG passport of a friend, and has contact to other emigration applicants at his workplace (p. 47). His superiors conclude that he presents an "unbearable burden to the work collective" who "support the implementation of disciplinary measures". After several more months of disciplinary measures, *Underground Exit* is fired and emigrates to West Berlin in 1985 (p. 47, 60-63).

Ultimately, *Underground Exit* is privately disloyal, yet he did not fit the traitor profile of the MfS in the absence of political statements or exceptional deviance. Labeling at his workplace may not have caused his political deviance, but there is evidence that it reinforced his social deviance to the point where he applied for emigration to the West. And by the time he was officially labeled in Prague, the MfS had already decided to approve his emigration request, precluding the need for further investigations or labeling due to a match with their traitor profile. By contrast, *Exits Denied* were similarly adamant about leaving the GDR, but used threats of exceptional political deviance to achieve their goal, and therefore fit the traitor profile of the MfS from the start. The labeling of the couple led to a shift towards loyalty, but only when official labeling was complemented by unofficial labeling, and conditions in their private lives that changed their sense of belonging to the GDR, independent of state security efforts.

Academic Reformers and Persistent Activists

Initial suspicion. A group of five *Academic Reformers* (#64-#68) were (aspiring) social science scholars specializing in Marxism/Leninism when in 1974, their activities attracted the attention of the MfS. An unofficial informant reports that two of them (#64, #65) attempted to recruit her for an

[14] #136 had been denounced in his neighborhood for having beaten his previous wife and doing the same with #137 (p. 125), and the MfS claimed to have used his reported "attempt to conceal his weaknesses through his oppositional attitude—including vis-a-vis his wife" in their intimidation (p. 113).

[15] Notably, #136 considered changing his job since "after his colleagues heard about the withdrawal of the emigration request, they behaved in a reserved manner to the point of making negative statements", but could not get the (photographic laboratory) position he desired since he failed the required security background checks.

[16] Seemingly unaware of this settlement five months later, his superiors at work still lament a bad performance despite multiple talks about "disciplinary violations", perceiving an "increasingly clear negative political attitude", and suggesting that his friends are encouraging *Underground Exit* to such behavior to achieve his exit, while implying that the marriage to the FRG citizen would be fictitious.

"oppositional circle", with "revisionist" aims that include changes to the democratic participation of the working class in Socialism, the relationship between political party and government, and its information politics (p. 32). The case worker does not observe any private disloyalty, but in 1975 justifies an operative procedure with perceptions of disloyalty that are presumed to conceal exceptional political deviance. As reported by the informant, though they are members of the SED, the *Academic Reformers* are pessimistic about changing it from within, and instead seek to develop, through regular seminars and literature studies, "a thorough theoretical preparation regarding [...] state and democracy problems", as well as ways to "mobilize the working class" and make contact to oppositional groups in the GDR and abroad, all the while "taking precautions [...] to avoid arrest" (p. 33-34). This perception is validated by the *Stasi*'s confidence in the informant who has "long been active for the HVA VI [department recruiting for foreign espionage] and is considered reliable" and "is a university student [...] herself". Possibly due to the high confidence in the informant and the general difficulty of placing informants in opposition movements, the case worker contents himself with narrating her reports, e.g.: "*According to the IM*, the danger of the group stems from its really intelligent leader (calls himself 'ZK'), who due to his conviction is willing to make personal sacrifices (arrest) and due to his envisioned alliance politics could quickly become a basin for different oppositional groups" (vol 1., p. 98).[17]

The 'opening' account for *Persistent Activists* is missing, but an intermediary report recalls the original suspicion. They were similarly suspected of exceptional 'voice' based on an informant report from the foreign espionage department in 1977 that they have contacts to "oppositional forces" (artists/writers) in Poland and West Berlin. They received and propagated 'illegal' literature, hosted meetings of other 'opposition figures', and "on the initiative of #148" planned to collect signatures for a petition in protest against the denaturalization of several writers from the GDR, though "a publication in the West was rejected by everyone" (p. 8). The MfS gives seperate accounts of how these perceptions of #147's and #148's loyalty fit their traitor profile. #148 is a 'friend' of #147 and had been "using his position in the FDJ to propagate hostile views against the GDR", and is the one with "extensive connections to oppositional people" in Socialist countries (p. 22). #147, in turn, had been a loyal unofficial informant from 1971 to 1977, aiding the arrest and conviction of six people, "among them a spy". However since 1972, she had defied MfS attempts to "extract her from the ideological influence of politically-negative people", to the point where she rejected collaboration with the MfS since she "could not reconcile with herself the reporting of GDR-critical people who have 'honest intentions' for Socialism". Subsequently, she had been unofficially reprimanded for her 'undisciplined' work as a fashion artist in 1979 (p. 3, 23). The MfS is therefore confident in the disloyalty of Persistent Activists, and in line with their traitor profile asserts that their contacts represent "hostile organizations who operate against the GDR".

Label confrontation. Some of the *Academic Reformers* had been labeled unofficially at their workplace before the MfS considered doing the same, and there is suggestive evidence that this reinforced private disloyalty.[18] The MfS however had problems with collecting "officially usable evidence" for political deviance as "the most important operative results were obtained by the unofficial informant" (vol. 1, p. 186). And while some of *Reformer's* activism passed the threshold of exceptional deviance (publications abroad that were considered agitation against the state), an official label was not appropriate as the

[17]Based on how the case worker described the informant's contributions relative to his own, she might have made a better security official, seeing as she not only collected but also aligned the intelligence with the organization's traitor profile for him.

[18]The aforementioned informant reports that one of them had two articles rejected by the university magazine due to "ideological errors. Because of that *Academic Reformer* demanded that the group should not be passive, but has to go into the offensive" (vol. 1, p. 182). Similarly, #66 was reprimanded for membership in a working group that "only dealt with revisionist literature" (vol. 1, p. 178). And a later statement by one of the reformers explains that "Slowly I became aware of my impotence in the face of opportunism and careerism [...] as I was more and more attacked not with arguments but suspicions, I began to formulate my views not as questions but positions. From this time dated my friendship with members of the group." (vol. 5, p. 216).

MfS could not link exceptional activities to particular individuals.[19] The around five security officials who pitched in on the case eventually matched the obscure international connections of *Reformer* #67 to their traitor profile,[20] and highlighted the magnitude of *Reformers* perceived disloyalty by citing their elaborate, yet-to-be realized plans.

The MfS chose to 'decompose' *Academic Reformers* with individually tailored interrogations between November 1977 and January 1978. #65 and #68 admitted to their "inner conflict" and the "breach of trust" they committed particularly over their revolutionary aspirations in the circle, and demonstrated loyalty by agreeing to report on the involvement of group members in specific activities. Interrogators dismissed the importance of #66 and her political deviance, mostly interrogating her about her husband #67: she is found to express similarly "critical" political positions as him, and demonstrated ignorance about the political deviance of the group ("of course he [#67] and I have friend circles, but these are not panel discussions").[21] By contrast, #64 and #67 accepted their labels yet defied the notion that it be treated as disloyalty vis-a-vis MfS officials and third parties, explaining that limited publication and discussion opportunities in the GDR were to blame for their behavior (vol. 20, p. 49-55). When asked how to proceed after the interrogations by an informant, #64 replied that they have to "wait for the whole theater with interrogations etc. to be over, after that there are many possibilities", and expects that they will no longer work academically with their political views, with plans to do company-work for a year (vol. 5, p. 57-61).[22]

When *Activist* #147 is officially interrogated by the People's Police for her support of Robert Havemann (a well-known activist) during his prosecution in 1979, she and #148 intensify their disloyalty rather than being intimidated. The security official observes that both *Persistent Activists* prepared for preventive measures by the MfS following the interrogation: they concealed connections and evidence, watched for signs of surveillance, reached out to Western organizations for help, and planned to engage with the MfS in case they 'make contact' to learn what they know, before gradually distancing themselves from the organization (again, in the case of #147). As part of these efforts, they decide to move the protest letters to West Berlin after all, so that they can (threaten to) publish them in case of arrest (p. 9-10). This is in line with the security agent's traitor profile, and justified the official prosecution of *Persistent Activists* for public agitation (p. 11-13).

Settlement. MfS interrogators settled the loyalties of *Academic Reformers* in individual interrogations. #65 and #68 were instructed to conceal their informing about group activities vis-a-vis third parties (vol. 5, p. 165-166), and reflected on their "political recovery" in public statements to their colleagues and SED party members (vol. 6, p. 180-181, 184-185). This loyalty was rewarded with the 'deletion' of their official SED party reprimands (vol. 27, p. 339). Seeing as #66 did not admit private disloyalty,

[19] Quotidian disloyalty included "posession of anti-Marxist, hostile" literature, contacts to opposition groups in Poland (who rejected their offer to send monetary aid anonymously) and other "operatively interesting" individuals abroad (who mainly helped obtain illegal literature), preparing 'open letters' for publication (publication record: one not submitted, one rejection, but two published in the FRG), all the while devising plans to avoid detection and identify potential MfS agents among recruits (vol. 4, p. 54; vol. 5, p. 58).

[20] In the absence of information about #67's connections, they posited the existence of an 'overarching group' worthy of a separate operation, though from the present files it appears that his ambitions to mobilize durable groups in the GDR remained mostly in a conceptual stage (vol. 1, p. 96; vol. 2, p. 243; vol. 4, p. 164).

[21] Security officials did of course not believe in such justifications (e.g. noting that #66 was at least aware of her husband's political deviance). Presumably they also did not believe #67 when he previously rejected collaboration with the MfS on the grounds that "his nerves could not handle it" (vol. 2, p. 243). In turn, political deviants merely demonstrated but did not privately change their loyalty. For instance, when #66 was threatened by the interrogator—"her and her husband's apparent refusal to contribute to the resolution of the matter only delayed it in a for them unfavorable fashion"— she is paraphrased with stating that "she and her husband had already discussed what would happen if they could no longer follow their current careers".

[22] *Reformer* #67 reportedly "had the impression that statements 'meant as jokes' were known to MfS and treated as serious lines of questioning", but otherwise did not reveal much about the interrogation to the informant.

the MfS decided to archive her case without a reprimand (vol. 5, p. 119-121; vol. 27, p. 348). By contrast, since #68 and #64 admitted their private disloyalty without repenting, they were officially reprimanded by the SED and lost their academic positions, yet maintained their political disloyalty (vol. 4, p. 166; vol. 6, p. 171; vol. 27, p. 339).[23]

Despite very similar behavior, *Persistent Activists* were right not to expect the same leniency from the MfS. They were both arrested shortly after the above-mentioned preparations, and sentenced to between nine and twelve months of prison for 'riotous assembly' on 31.9.1979. In January 1980, the case worker noted that #147 had been released (possibly as part of a general 'amnesty') and, finding "thus far no evidence of hostile activities directed against the GDR", closed the file (p. 38). As it turns out though, the case worker misjudged her allegiance:[24] a year later the Stasi had once more started to check her 'politically-negative contacts', though they could not observe any exceptional deviance. This intelligence failure is accounted for by her "mistrust of new acquaintances as she constantly fears unofficial MfS informants" (p. 43), rather than any decline in private disloyalty "vis-à-vis the societal conditions in the GDR", which justifies the use of "disruptive measures" to discredit her vis-à-vis her fellow activists, such as by identifying criminal in place of political deviance (p. 42-44). In turn, *Persistent Activist* writes letters in protest of her professional discrimination and travel bans. After five years of surveillance and several more preventive interrogations, in which *Persistent Activist* made no statements or concessions as to her contacts and refused collaboration with the MfS, she "promises not to engage in [exceptional] political activities" in exchange for an end to the "alleged" hindering of her career by the MfS.[25] Yet, rather than reporting that she ceased any of the political contacts that prompted her renewed observation, the security official conceded that she made new contacts instead, but nevertheless decided that "the main reasons for her operative control are no longer given", justifying the (final) closure of the file in 1986 (p. 103).

Overall, the MfS officially labeled *Academic Reformers* and *Persistent Activist* alike, yet only those who refused to demonstrate loyalty were repeatedly punished and simultaneously least likely to shift allegiance. While one could argue along with Hirschman (1970) that these activists had a 'special [private] attachment' to Socialism or the GDR, their continued deviance was heavily influenced by "oppositional contacts", whom they refused to denounce and who for the most part did not denounce them in return. Under these conditions, based on the (rarely detailed) mentions of unofficial labeling *prior* to exceptional deviance in MfS reporting, it appears that such labels reinforced disloyalty. By comparison, MfS reporting on the *Academic Reformers* who shifted towards loyalty makes no mention of unofficial labeling.

Abusive Escapist and Brother's Keeper

Initial suspicion. By 1989 emigration to the West was perhaps the most popular form of political deviance in the GDR, including among members of the military. In July that year, one soldier was facing a divorce from his wife, and in the presence of a People's Police officer and another fellow solider stated to her that "When I come home the next time, I will finish you and then I make a run over [to the FRG]". And during the drive to his post he states to a fellow soldier: "you better arrest me,

[23] #67 in particular did not reveal his obscure contacts to the MfS, and is perceived to engage in 'voice' in subsequent surveillance (vol. 27, p. 341).

[24] The time period between incarceration, release, and closure of the file is so brief that one might suspect the case worker of trying to rid themselves of it before *Persistent Activist* became active again. Notably, #148 is not discussed again in the file, but #147 appears to have made contact with (similarly persistent) *Academic Reformer* #67—possibly in relation to their shared support of well-known activist Havemann in his prosecution.

[25] Presumably not a difficult promise since at least from her file, *Persistent Activist* did not commit exceptional political deviance since her release, and much like *Academic Reformers* was barely exceptionally deviant in the first place.

otherwise I would run again".[26] As a first official measure, *Abusive Escapist* was temporarily placed under arrest until a 'preventive talk' (p. 6-7).

And in November 1988, another soldier's grievances are reported by an informant: *Brother's Keeper* does not want to be positioned at a border unit where the army wants to place him, would refuse to use his weapon for humanitarian reasons, had enjoyed to "finally" watch West German television in a hotel room his parents had rented, and shared the story of how his brother had attempted to leave the GDR with a rubber boat. When prompted by the informant "why he didn't partake" in the exit attempt and whether he really "wants to stay 1.5 years in the artillery unit [that he was in at the time]", *Brother's Keeper* replied that "it [exit] could have gone wrong" and that he "doesn't give a fuck" about staying in the artillery (p. 114). Initial investigations find that his brother is a 'persistent emigration applicant', had been arrested for his emigration attempt, and that *Brother's Keeper* is familiar with border security measures, making his political deviance particularly problematic. The MfS quite agrees with *Brother's Keeper* that he should not have been called to the border unit, given the unofficial information about his "political instability" (p. 116-117).

Abusive Escapist and *Brother's Keeper* react differently to similar accusations of quotidian political deviance. The former appears embedded in the military and responds to the 'incidental' accusation of disloyalty with compliance, such that the MfS facilitates the resolution of his personal problems to control him. By comparison, *Brother's Keeper* denies private disloyalty (his knowledge of an exit attempt), and defies attempts to fit his brother's exit into a traitor profile. Both suspects have no problems settling on loyalty and disloyalty outcomes with the MfS, respectively, seeing as their private loyalty is already aligned with the perceptions of security officials.

Label confrontation. The "reprimanded" *Abusive Escapist* is compliant in his preventive talk with the MfS, confirming the statements he made in the presence of his colleague. He explains that he has no 'practical thoughts' to emigrate, not having thought of life in the FRG and having no knowledge about border security. *Abusive Escapist* admits that he can't control himself under the influence of alcohol and that he is worried to be separated from his child in case of divorce. After the talk, an informants reports that he made similar remarks to him: the 'exit' statement was not genuine, and "he did not care about anything but that his wife gives up on the thought of divorce".

By contrast, the MfS investigates whether *Brother's Keeper* knew of his brother's political deviance without reporting it, which would warrant legal measures, but can merely ascertain that he could have learned through family contacts. They organize a preventive talk with the soldier, where he confirms that his brother had been arrested for an exit attempt, but denies any knowledge of it. Moreover, *Brother's Keeper* makes explicit that he condones the "criminal offenses" of his brother, which confirms his "political instability" to the MfS.

Settlement. The MfS made arrangements such that the brother of *Abusive Escapist* was used to 'positively influence' him, and found that "after the wife agreed to continue the marriage" they observed no more disloyalty. They noted that he should remain under unofficial control due to his impulsiveness. By comparison, the MfS ensured that *Brother's Keeper* is moved to the builder's company rather than the border. Noting that he has a "strong connection to his parents" they viewed the matter as closed, and transferred the file to another department.

[26] Seeing as the *Stasi* has no prior record on his political deviance, the "again" here presumably indicates leaving his post to go home without authorization.

5. Discussion

Does labeling disloyalty stop betrayal? The discussed cases from the *Stasi*'s track record would suggest that labels 'work' when politicized behavior is exceptional to the group to which the labeled seek to belong, as it prompts them to demonstrate their loyalty. This is particularly the case when the political deviance of the labeled is in fact incidental, and they do not receive support for disloyalty from fellow group members or rival authorities. But when state security labeled individuals for behavior that was perceived as loyal by group members, it tends to have no effect or increases disloyalty. Ironically, such labels may even be invited by suspects in order to boost their status as a defector or receive support for disloyalty. As a particularity of the GDR case during the 1970s and 80s, defectors were not as vilified as they often are in conflict settings, leading to at times significant support both within the GDR and abroad. Notably in those cases, labeling does not elicit behavior change, but still reinforces the imagined traitor profile of authorities amongst security officials and potentially among 'ordinary' group members.

Some organizational dynamics that relate to the study of state repression are worth highlighting as well. First, in line with existing theories on defector identification, reliable informants appear to decrease the need for political arrests (Steinert 2022) and other forms of 'indiscriminate' labeling (see Kalyvas 2006), but not because of the high-quality information they provide to facilitate the correct identification of targets. At least in the GDR, informants could give security officials the impression that defectors can be controlled through unofficial means, which do not appear in official statistics, and provided them with reassurances to justifiably close cases without further investigation. Through unofficial labeling, loyal informants are more effective at stopping disloyalty than state agents. But since 'true defectors', perceived and privately disloyal, hardly operate in social groups where they are vilified, labeling them at best leads them to hide their private disloyalty from the eyes of security officials in the future.

Second, MfS officials devoted far more reporting (pages) to exceptional than to quotidian disloyalty, seeing as the latter were settled more quickly with fewer justifications. This means that in terms of workload for the official, controlling quotidian political deviance earlier may have been preferable to waiting for later exceptional deviance that requires more effort to control. To successfully close a case, the official who observed quotidian political deviance had to either conclude that the suspect had no intention or opportunity to defect (risking blame if they are wrong), or had to justify measures that allowed for a short-term disambiguation of private loyalty. This may partially explain why traitor profiles practically maintain themselves: once created, they allow security officials to open cases (for imagined disloyalty) and close them quickly (for lack of perceived or private disloyalty).

Third, permanent surveillance was only considered for those exceptional political deviants who repeatedly demonstrated their disloyalty and were the least likely to shift allegiances. Security agents may have made the lives of defectors particularly difficult when they felt that the incorrigibly disloyal are making it difficult for them and 'their' community. Finally, when officials were confident in perceptions of disloyalty but could not officially label it, they resorted to labeling social deviance that fit their traitor profile. This meant that marginalized social groups had both a good chance of being perceived as defectors by informants (who disliked social deviants), and of being officially labeled by authorities.

There are a number of alternative explanations for changes in political deviance than labeling it. These range from basic demographic characteristics (e.g. age, gender, socio-economic status) and mechanisms to explain compliance (e.g. social influence, power relations) to changes in the international environment (e.g. détente under new leadership from 1971 onwards) and incentive structures (e.g. the increasing economic deterioration of the GDR relative to the FRG, decreasing ability to punish given the number of defectors). Many of them can be integrated into the framework proposed here: a young NVA soldier may be more likely to comply with authority than his middle-aged superior; women may be perceived as intentionally disloyal less frequently than men seeing as their motivations are sexualized more often

in the process; labeling suspects while providing monetary and other loyalty incentives becomes more difficult with a stagnating economy and official tolerance for East-West contact; and a perceptibly disloyal upbringing may make it more likely for suspects to begin a 'defector career'. A statistical comparison to assess the relative importance of these factors for political deviance is an avenue for future research. The more modest goal in this paper was to show how labeling defectors affects politicized behavior, and to demonstrate how one might interpret the requisite data to that end.

PAPER 3
Defection from Covenants in Conflict: Experimental Evidence on the Effects of Punishment*

Mirko Reul[1]

[1]Graduate Institute of International and Development Studies, Department of Political Science/International Relations, Geneva, Switzerland.

Abstract

Does punishment increase cooperation with covenants in conflict? From regime critics to enemy-informants, groups punish their members for disloyal behavior, at times ostracizing them if they refuse to cooperate with loyalty expectations. That selective punishment elicits conditional cooperation is largely supported by experimental evidence, yet observational studies suggest that punishment is frequently indiscriminate, not perceived as justified, and that defection persists among those ostracized. Moreover in conflict settings, the culpability of defectors moderates the efficacy of punishment, given imperfect information about disloyal behavior: (1) sanctions may be ineffective where disloyalty is not intended, and (2) who is punished for defection is shaped by popular perceptions. This paper presents results from a lab experiment with N=240 subjects to evaluate the effect of punishment on defection. In the experiment, two groups engage in a nested Tullock contest with uncertain opportunities to demonstrate loyalty. The control condition establishes a baseline for contributions to the contest and defection as side-switching. Two between-subject treatment conditions introduce (1) unequal opportunities to demonstrate loyalty, and (2) communication between group members. Individuals are expected to shift their allegiance in response to misidentification: inequality increases defection as participants unjustifiably accuse their peers of disloyalty, while communication increases cooperation as groups coordinate on appropriate loyalty expectations and punishments.

*I especially thank Noah Bacine, who contributed to this paper with numerous discussions of the design, co-drafting of instructions and survey questions, and implementing the experiment in the lab under difficult conditions. I also thank Jake Bowers, Janine Bressmer, Juliette Ganne, Paroma Ghose, Sungmin Rho, Alessandra Romani, David Sylvan, Eliza Urwin, and discussants at the Gothenburg University GLD and the EUI-Graduate Institute workshops for their helpful comments and suggestions on several versions of the design. This study is fully conducted through the Nuffield College Centre for Experimental Social Sciences (CESS), and received approval from the center's ethics review board. The project was supported by a SNF research grant 188287.

1. Motivation

The labeling of individuals as defectors can lead to allegiance shifts. During the 1936 Arab Uprising in Palestine, a family with friendly relations to Jews was attacked by insurgents for their refusal to pay them in exchange for refraining from attacks on nearby Jewish vineyards, and the surviving family members subsequently aided the British occupation (Cohen 2008, 161). By contrast, another Palestinian who sold land to Jews "expressed his regret" at a gathering of village leaders, promising them to "abstain from such acts in the future" (Cohen 2008, 110-112). In 1980s East Germany, the Ministry for State Security (MfS, *Stasi*) denied a doctor vacation in Hungary on suspicion that she may defect to the West, prompting her to abstain from work and ultimately pressure authorities to tolerate her emigration (BArch, MfS, BV Potsdam, Abt. XX, Nr. 790, 32-36). By contrast, a company manager whom the MfS accused of espionage in the 1970s reluctantly collaborated with state security, admitting culpability and providing information on his past intelligence activities (BArch, MfS, AP 14791/72).

In each of these cases, individuals find themselves in *loyalty conflicts* where they are expected to cooperate with one of two rival social groups. Political actors in each group punish deviant behavior that not only "departs from the normative" but poses a threat to political order (Raybeck 1991, 23). And those who are punished for perceived disloyalty either prove their allegiance by increasing cooperation, or defect to the rival. This paper discusses preliminary results from a lab experiment on the determinants of defection in conflict. A challenge to analyzing loyalty conflicts in the lab is that participants do not share a social identity (see Tajfel and Turner 1986), and consequently lack loyalty expectations for each other, have no incentive to cooperate or defect beyond strategic considerations, or punish each other for perceived disloyalty. This paper focuses on generating the conditions of loyalty conflicts without harming participants: under what circumstances are laboratory participants willing to punish each other for perceived disloyalty? And given the existence of loyalty conflicts, under what conditions does punishment lead to shifts in cooperation or defection?

Existing empirical studies focus on *collective* defection, as groups of people "voice" or "exit" against the will of authorities in response to grievances with political order (Hirschman 1970), be they motivated by interactions between the two (Hirschman 1993; Pfaff and Kim 2003), organizational characteristics (Albrecht and Ohl 2016; Dworschak 2020; Staniland 2012), or social justice appeals (Ophir 2020) that lead to the fragmentation of state institutions and popular dissent; or by ensuing group conflict over relative deprivation (Gurr 1968), the exclusion of ethnic groups from state power (Cederman et al. 2013), violent repression (Schutte 2017), and changes in territorial control (Kalyvas 2008). In particular, conflict studies focus on the interaction between repression by political regimes, on the one hand, and contentious mobilization against them, on the other (e.g. Della Porta 1995; Kocher et al. 2011; Lichbach 1987; McAdam et al. 2001; Schutte 2017; Sullivan 2016b; Wickham-Crowley 1987). In the context of group conflict (Coser 1956; Simmel 1955), shared perceptions of unjustified punishment are associated with collective resistance against authorities (Wood 2003), while uncertainty about loyalties in asymmetric civil conflict leads authorities to punish indiscriminately when they perceive defection (Kalyvas 2006). Theoretical models have highlighted the importance of punishment for individual behavior (Axelrod 1986; Bhavnani 2006; Heckathorn 1990), yet empirical research falls short of tracing the *individual reactions* to punishment that groups perceive as justified, seeing as related outcomes—from refusing to contribute to group goals to actively supporting a rival—are challenging to observe systematically.

To induce loyalty conflicts, a modified Tullock contest primes participants on making personal sacrifices to win with their ingroup against a rival outgroup (Bornstein 1992; Rapoport and Bornstein 1987; Tullock 1980), yet equally allows them to defect for personal gain. The design is based on existing experiments on loyalty and cooperation with groups in conflict. In social psychology, loyalty is the willingness to forgo individual rewards in the service of group goals for the purpose of enhancing ingroup

welfare at the expense of outgroups. The propensity of individuals to demonstrate loyalty increases with their social identification, and perceptions of disloyalty may lead to negative reactions by ingroup members (Levine and Moreland 2002). However, this literature stops short of investigating the behavior of deviant group members after their social exclusion from groups (see Abrams et al. 2018; Castano et al. 2002; Ditrich and Sassenberg 2016; Lindström and Tobler 2018; Marques and Paez 1994; Travaglino et al. 2014; Zdaniuk and Levine 2001). Related economic games investigate social dilemma, where the individual incentive to free-ride on the cooperation of others threatens efficient group outcomes (Olson 1965). Punishing free-riders consistently increases loyalty as cooperation while 'anti-social' punishment does not, but competing groups are not considered, and individual reactions to punishment are not theorized beyond the maximization of personal benefits (Andreoni and Gee 2012; Casari and Luini 2009; Cinyabuguma et al. 2006; Fehr and Gächter 2000; Grechenig et al. 2010; Herrmann et al. 2008; Nikiforakis and Normann 2008; Ostrom et al. 1992). Moreover, experiments on group contests neither allow for uncertainty regarding individual loyalties nor for defection as side-switching (Abbink et al. 2010; De Dreu et al. 2016; Dechenaux et al. 2015; Mago et al. 2016; Sheremeta 2018), even though both are key to explaining violence in conflict settings (e.g. Bhavnani et al. 2011; Kalyvas 2008; Liu 2022; McLauchlin 2010; Staniland 2012). For example, Abbink et al. (2010) find that group competition and the possibility of punishment increase cooperation, compared to competitions between individuals without punishment, but neither study the effects of punishment on individual cooperation nor allow participants to defect to the other group.

Assuming loyalty conflicts are induced among participants, the study focuses on the effects of punishment on both cooperation with an ingroup, and on defection as side-switching to an outgroup. Two experimental conditions test how these effects are moderated by common features of real-world loyalty conflicts. First, defectors may be *unequally misidentified*, due in particular to the biased use of perceived rather than actual disloyalty given imperfect information (see Kalyvas 2006, 89; Davenport 2005; Levine and Moreland 2002; Lyall et al. 2015; McLauchlin and Parra-Pérez 2018). The effects of such misidentification, based on a distance between observable behavior and disloyal intentions, are near impossible to observe in real-world settings. Unequal opportunities to demonstrate loyalty are expected to decrease cooperation with the group and increase defection to the rival. Second, loyalty expectations and defection are *socially constructed*, based on shared norms about acceptable and unacceptable behavior, and perceptions where individuals fall within that range. Even when political authorities clearly delineate expectations for loyalty, defectors may be sanctioned by their peers outside of official 'loyalty trials'.[1] And in spite of procedurally fair official trials, defectors may be vilified or protected based on popular perceptions.[2] The design incorporates peer-based deliberation of expectations and defector culpability by allowing for ingroup communication.

The following three sections respectively discuss the design of the group contest at the core of the study, experimental treatments, and hypothesized effects. Section 5 presents preliminary results based on pilot studies, and Section 6 discusses the ability of the design to induce loyalty conflicts and identify causal effects.

[1] See for example the swift popular condemnation and sentencing of a Palestinian defector, who as it turned out was coerced into collaboration with Israel on threat of imprisonment, and who may not have known that Israeli intelligence was using him to assassinate his cousin (Williams 2001, 30-32; Human Rights Watch 2001, 45-46; al-Bitawi 2016, 35; see Abdel-Jawad 2001).

[2] In contrast to the Palestinian case, consider the experience of Browe Bergdahl—a U.S. Army soldier who left his outpost in Afghanistan without authorization in 2009, was captured by the Taliban, released in a controversial exchange deal in 2014, and sentenced guilty of desertion in 2017 (BBC 2017). Given the loss of US soldiers in search for Bergdahl and the release of Taliban in the prisoner exchange, politicians publicly called for Bergdahl's punishment as a "traitor" (Morgan 2020; Seck 2021), while affiliates of Bergdahl were threatened by those who villified him (Rosenberg 2016).

2. Game Design and Procedures

The experiment is implemented with participants from the subject pool at the Nuffield Center for Experimental Social Sciences (CESS) Oxford, in August (first pilot), September (second pilot), and October 2022 (main study). For the pilot of the design presented here, a total of $N = 80$ subjects was recruited to participate in a decision-making experiment that lasted approximately 2 hours (including 30 minutes for arrival and payment procedures).[3] Eight participants jointly participate in each session, with up to two sessions implemented simultaneously. On arrival, participants are asked to sign a consent form (see Appendix C.II), seated individually in front of computers, and randomly assigned to two groups with $n = 4$ players in each group. Participants interact only with members of the same group throughout the experiment (fixed matching protocol), and are identifiable only by a constant number. Initial instructions inform participants that they make decisions as a member of one of two 'teams', that each player will make choices individually, but that their final bonus payment depends on the aggregate choices in each team.

Figure 3.1 presents the five *stages* of the game that participants play repeatedly over 20 rounds, as well as the timing of experimental treatments that modify key aspects of the game between sessions. During Rounds 1-5, participants engage in a Tullock contest, where contributions from a personal endowment determine the probability for winning a prize as a group. In Round 6, participants are introduced to Stage 3, which allows them to punish each other after choosing their contributions to the group contest. And from Round 11, they are introduced to Stage 4, where two randomly selected participants have the opportunity to switch their contributions to the opponent. Participants then play all stages for ten rounds. They have complete information about possible moves, payoffs and rounds in each part of the game, but not about the distribution of randomly assigned treatments (see Appendix C.III), and are informed that experimenters are not allowed to deceive participants. The details of each stage are described in the following sections, and the details of each treatment in Section 3.

After Round 20, a short survey prompts participants about their experience with similar games and the motivations behind key decisions in the game (see Appendix C.IV).[4] After the survey, they privately receive a guaranteed show-up fee and a completion bonus of respectively £5, as well as an additional payment based on their decisions and the decisions of other participants in the game. Assuming all participants adhere to Nash equilibrium play they would earn £34 on average (see Section 2.6), which is slightly above the realized average earnings of £33.

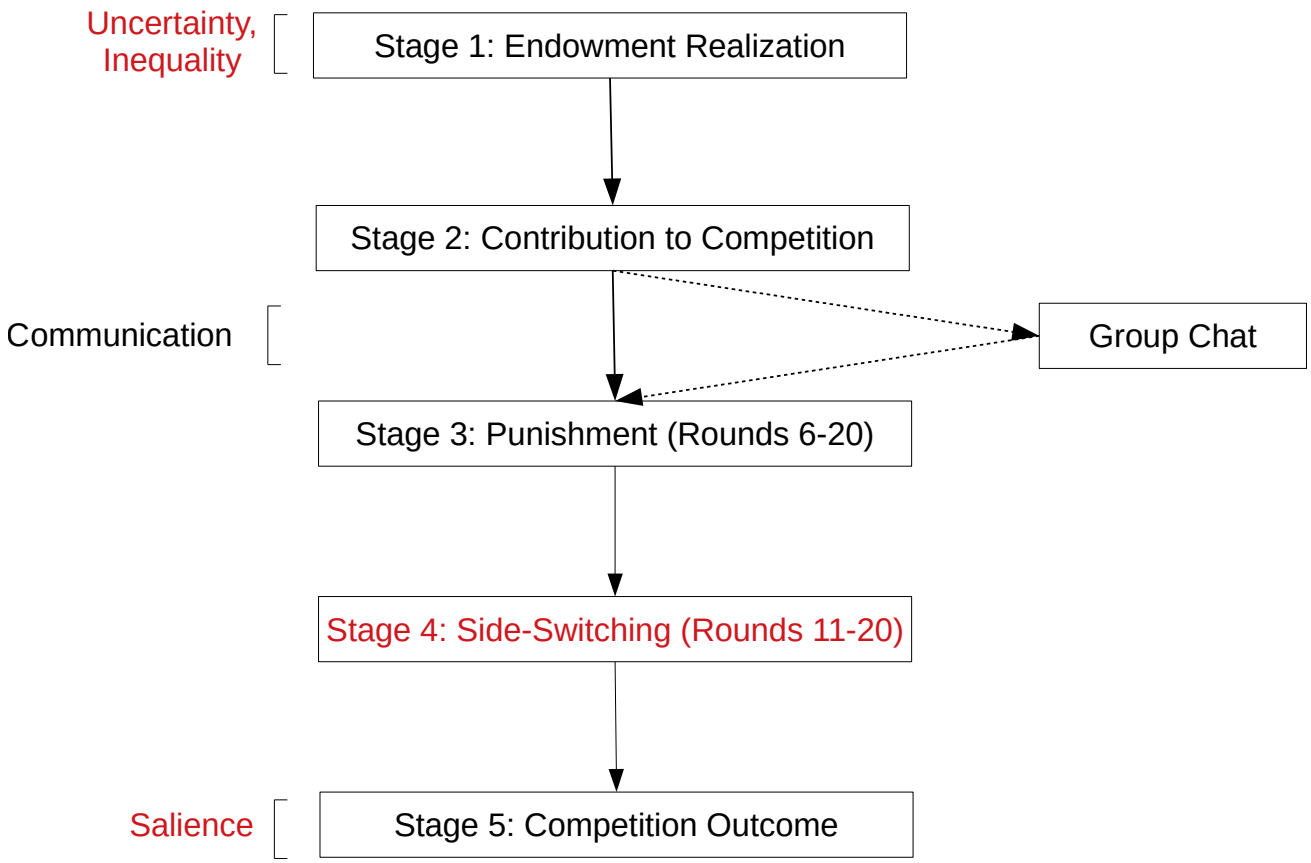
2.1. Stage 1: Endowment Realization

In every round, each participant receives an endowment of experimental currency units (ECU) or 'points', denoted by e_i. The endowment is drawn from a distribution that ranges from 20 to 100, in increments of 20. In the control condition, the probability of receiving any individual endowment is equal, conditional on the sum of endowments across groups being equal. In other words, players with

[3] $N = 48$ participants were recruited for the pilot study of the present design in September, and an additional $N = 32$ participated in the pilot of a previous design in August. The following discussion focuses on the design and results from the latest pilot study in September.

[4] Screens have no time limit, but the experimenter ensures that participants do not idle significantly longer than other participants on a given screen. Instructions are displayed on screen and additionally distributed to participants in paper form. The experiment is implemented using *oTree* (Chen et al. 2016).

Figure 3.1. Loyalty Game Flow Diagram



Note: Boxed text constitutes a game design stage consisting of decisions made by participants, text outside boxes indicates experimental modifications of the core game. Design features highlighted in red deviate significantly from existing experimental studies on cooperation and defection in general.

the same identifier in each group receive the same endowment:

$$\begin{aligned} & e_i = 60\epsilon_i \text{ ECU} \\ & \epsilon_i \in [\frac{1}{3}, \frac{2}{3}, 1, 1\frac{1}{3}, 1\frac{2}{3}] \\ & i \in [1, 4] \end{aligned} \tag{3.1}$$

Participants are informed that they may receive unequal endowments, and that the sum of endowments is identical in every team (symmetric conflict), but they are not informed about the endowment distribution. This ensures that participants are not deceived during the experiment, and have an equal albeit random chance of receiving an equitable payout, but have imperfect information about other group member's opportunities to cooperate.

At the end of every round, unspent points are deposited in each participant's personal fund and converted to cash ($125\ ECU = 1£$).

2.2. Stage 2: Contribution

After learning their endowment, participants independently choose to purchase $c_i \in [0, e_i]$ tokens as a contribution to their team's 'competition fund'. The competition is modeled as a contest between two groups, where players make irreversible contributions to increase the probability of winning a prize that is shared among members of the winning group (for recent reviews see Dechenaux et al. 2015; Sheremeta 2018). Formally, the sum of contributions in each group is denoted by C_I, and follows a *perfect-substitutes* function:

$$C_I = sum(c_i, ..., c_n) \tag{3.2}$$

Participants are informed that contributions are visible to group members (public information).

2.3. Stage 3: Punishment

This stage only takes place after Round 5, which is when the possibility of punishment is introduced. After each participant chooses their contribution to the competition, each group member's choice is revealed to other group members, including the total contributions across rounds. Then, each player i has the opportunity to assign up to ten punishment tokens to each other member j of their group:

$$l_{ij} \in [0, 10] \tag{3.3}$$

Each token that a participant assigns reduces their earnings by 1 *ECU* (see Abbink et al. 2010; Fehr and Gächter 2000; Ostrom et al. 1992).

Only the player j who receives the most punishment tokens is punished, with parity resolved by random choice:

$$L_{jI} = \sum_{i \neq j}^{n} l_{ij} \geq L_{kI}, \text{for all } k \neq j \wedge I = I \tag{3.4}$$

Players then see a screen with the points that each participant received, but not from whom. It follows that the group can identify and punish at most one participant as a defector per round. The cost of purchased tokens is deducted from the purchasing participant regardless of who is punished.

The impact of punishment is limited to a penalty:

$$\begin{aligned} w &= 3l_{ij} \\ s_i &= L_{jI} \cdot w \end{aligned} \tag{3.5}$$

Where w denotes the strength of the sanction per token assigned, and s_i the penalty that is applied to the punished player's final earnings after the outcome of the competition is determined. It follows that sanction strength increases with assigned punishment tokens, with a cost-to-punishment ratio of 1 : 3

(see Nikiforakis and Normann 2008).[5]

2.4. Stage 4: Side-Switching Choice

This stage only takes place after Round 10. With the outcome of Stage 3 revealed, two participants are assigned the opportunity to switch their initial contribution over to the other group. Punished participants are almost certain to receive the opportunity to switch, whereas all other participants have an equal probability. If no participant was punished, all participants have the same probability to receive the opportunity:

$$p_i(D_i|L_{jI}) = \begin{cases} 0.99, \text{if } L_jI > 0 \\ 0.01, \text{if } L_jI = 0 \\ \frac{1}{n}, \text{if } L_jI > 0 \ni L_1^k \text{for all } k \neq j \wedge I = I \end{cases} \tag{3.6}$$

Participants are informed that switching opportunities are assigned at random and independently each round, but are not informed about the probability distribution.

Switched contributions are subtracted from the fund of the participant's group, and added to the fund of the rival group, such that overall contest performance is the sum of ingroup and outgroup contributions:

$$C_I = sum(c_i, ..., c_n) + sum(c_j d_j, ..., c_n d_n) \tag{3.7}$$

Participants are informed that their side-switching choice is always private information.

2.5. Stage 5: Competition Outcome

During the first five rounds, this stage occurs after Stage 2 contributions. The outcome of the competition is determined probabilistically with a *Tullock lottery* function (Tullock 1980). The probability for a group to win linearly increases with competition funds as per Equation (3.2):

$$p_I(C_I, C_J) = \begin{cases} \frac{C_I}{C_I + C_J}, \text{if } C_I + C_J > 0 \\ 0, \text{if } C_I + C_J = 0 \end{cases} \tag{3.8}$$

Following Abbink et al. (2010), with zero contributions in both groups, neither group wins the competition and there are no additional payoffs for any player in that round.

In Rounds 1-10, each member of the winning group receives the same, additional payoff, based on an egalitarian distribution of a $V_I = 960$ ECU contest prize:[6]

$$v_{iI} = 240 \ ECU \tag{3.9}$$

[5] Following Abbink et al. (2010) in the unlikely event where this would result in total negative earnings at the end of the experiment, earnings are set to zero.

[6] See Gunnthorsdottir and Rapoport (2006) on egalitarian versus proportional prize distributions in a nested group competition.

After Round 10, this stage follows Stage 4 defection choices. The probability for each group to win increases with contributions from group members, and contributions by defectors from the rival group, as per Equation (3.7). The outcome of the competition is determined by Equation (3.8), as during the first ten rounds. As before, each loyal member of the winning group receive an additional payoff equal to Equation (3.9), while loyal members of the losing group receive no additional payoff.

Each participant who switched to the winning group receives a payoff equal to:

$$v_{iJ}(d_i = 1) = 192\ ECU \tag{3.10}$$

The prize would make participants indifferent between staying in their team and switching, assuming that all participants adhere to the Nash equilibrium (see Equation (3.13)).

Participants receive feedback on the outcome of the competition, the public contributions of their team members, and their private payoffs before moving to the next round. This includes their remaining private funds after contributions, potential competition winnings, side-switching choices, as well as costs from purchasing penalty tokens and being penalized:

$$\mu_i = e_i - c_i - l_i - s_i + v_i \tag{3.11}$$

Outgroup performance is not shown to minimize intergroup coordination.

2.6. Game Discussion

The average endowment received by participants amounts to $\hat{e_i} = 60\ ECU$. Following conventional group contests with perfect substitutes performance functions (Konrad 2009, 133-136; Katz et al. 1990; Nitzan 1991; Sheremeta 2018), given an equal valuation and distribution of the prize and equal group sizes ($n_I = n_J = n$), the symmetric Nash equilibrium for the competition contribution of each player is given by:

$$c_i = \begin{cases} \frac{V_i}{4n} = 60, \text{if } c_i \geq e_i, \\ e_i, \text{if } c_i < e_i \end{cases} \tag{3.12}$$

In contrast to existing studies, the equilibrium contribution is constrained by realized endowments, and is otherwise equal to the expected value of the endowment rather than the expected winnings from the contest. From a rational choice perspective, participants should therefore contribute the expected endowment to the contest, and keep any remaining endowment to themselves, conditional on the expected contributions of group members being equal.

As discussed in previous studies (see Abbink et al. 2010; Fehr and Gächter 2000), utility-maximizing participants are not expected to punish given that efforts toward punishment are wasted. Assuming that all participants contribute the Nash equilibrium, the marginal benefit for a single participant to defect is set such that they are indifferent between cooperation and defection. In other words, the conditional winnings after shifting the Nash contribution to the outgroup are equal to the conditional winnings for

cooperating with one's own group:

$$\begin{aligned} & v_i(p_J|d_i = 1) = 240(p_I|d_i = 0) \\ & \Rightarrow v_i(\frac{300}{480}|d_i = 1) = 240(\frac{240}{480}|d_i = 0) \\ & \Rightarrow v_i(d_i = 1) = 192\ ECU \end{aligned} \tag{3.13}$$

Limited opportunities to defect preclude coordination on defection, which is not usually possible in settings where disloyalty is viewed negatively by most group members and political authorities. The high probability for punished participants to have the side-switching opportunity ensures that it is possible to compare their choice directly to participants who were not punished, without deceiving participants.

Overall, participant payoffs given Nash equilibrium play amount to:

$$\mu_i = 20(60\ ECU - 60\ ECU + 0.5 \cdot 240\ ECU) = 2400\ ECU = £24 \tag{3.14}$$

Therefore, participants are expected to earn an average of £12 per hour, plus the show-up and completion payments. Note that in one of the treatment conditions, some participants will by chance receive a lower per-round payoff, but all participants received at minimum the expected additional £10/hour, in line with CESS Oxford guidelines.

The design is comparable to experimental group contests with homogenous endowments (e.g. Abbink et al. 2010; Cason et al. 2012; Sutter and Strassmair 2009). In the group contest with punishment that is most comparable to the present design, Abbink et al. (2010) finds usage of 10% of the maximum possible punishment points, and significant over-expenditure on the competition above the Nash equilibrium.

However where Abbink et al. apply all assigned punishment tokens, in the present study only a single defector can be punished in a given round, such that as in loyalty conflicts, punished group members constitute a minority that is being excluded from the group. Punishment in the present study therefore differs from the most commonly used peer-based mechanism in public goods experiments. The constraint provides incentives to identify defectors rather than punish randomly, and coordinate with other group members to do so (see Andreoni and Gee 2012; DeAngelo and Gee 2018). Participants have an incentive to assign points in order to avoid being punished themselves, but this strategy is inefficient given that all group members have the same opportunities to punish.

The design further bears some similarities to contests where the valuation of the prize differs between players (see Gunnthorsdottir and Rapoport 2006; Sheremeta 2009). Prize heterogeneity may be interpreted as heterogeneity of abilities or costs of contributions, and similarly drive some players to exert higher effort than others. But prize valuation manipulates participant's intrinsic motivations to exert effort without determining it: those who highly value the prize can still defect, and those with lower prize valuation can cooperate. By contrast, heterogeneity in endowments manipulates participant's perceptions of other's commitment: players who highly value the prize but have a low endowment may be falsely perceived as defectors, and players who do not value the prize but have a high endowment may secretly defect.

3. Experimental Design

Table 3.1 depicts the $2x2x3$ factorial design of the experiment. Columns indicate the three within-subject conditions over 20 rounds in each session, and rows indicate the between-subject conditions that are randomly assigned to different sessions.

In the control condition described above, participants first play the group contest (*C*, Round 1-5), then the group contest with punishment (*CP*, Round 6-10) and finally the group contest with punishment and defection (*CPD*, Round 11-20), using the same group and identifier within their group. Treatments are applied to all rounds of a session: in one treatment condition participants in each group receive unequal endowments (Section 3.1), and in another they have the opportunity to chat (Section 3.2). The full 'loyalty game' therefore combines the baseline above with unequal endowments and opportunities to communicate via an ingroup chat.

Table 3.1. Experimental Design

Treatment (Between-Subject)	**Rounds** (Within-Subject)			**Survey**	**Participants**
	1-5	6-10	11-20		
Control	C	CP	CPD	X	60
Endowment Inequality	C-Ineq	CP-Ineq	CPD-Ineq	X	60
Communication	C-Comm	CP-Comm	CPD-Comm	X	60
Endowment Inequality + Communication	C-Ineq-Comm	CP-Ineq-Comm	CPD-Ineq-Comm	X	60
Sum: 30 Sessions					240

Note: C = Group Contest (Rounds 1-5), CP = Group Contest with Punishment (Rounds 6-10), CPD = Group Contest with Punishment and Side-Switching (Rounds 11-20). *Ineq*uality and *Comm*unication treatments are assigned at the session level.

3.1. Treatment 1: 'Inequality'

This condition adds a change to the distribution of endowments. In every session, exactly two participants in each group are randomly assigned the 'poor' type, denoted by r_i. The probability of being poor is equal for all participants:

$$\begin{aligned} &p(r_i = \text{Poor}) = \frac{1}{2}n \\ &r_i \in_R [\text{Poor}, \text{Rich}] \end{aligned} \tag{3.15}$$

Player types determine the probability of drawing a particular endowment. To increase the contrast

between types while minimizing the risk of revealing them to participants,[7] the probability of receiving higher endowments decreases for poor and increases for rich players. *Poor* players are most likely to receive an endowment of 20 or 40 points, whereas *rich* players are most likely to receive an endowment of 80 or 100 points. Formally, Equation (3.1) is modified by:

$$P(\epsilon_i|r_i) = \begin{cases} p_{\frac{1}{3}e_i} = 0.4, p_{\frac{2}{3}e_i} = 0.4, p_1 = 0.1, p_{1\frac{1}{3}e_i} = 0.05, p_{1\frac{2}{3}e_i} = 0.05, \text{if } r_i = \text{Poor} \\ p_{\frac{1}{3}e_i} = 0.05, p_{\frac{2}{3}e_i} = 0.05, p_1 = 0.1, p_{1\frac{1}{3}e_i} = 0.4, p_{1\frac{2}{3}e_i} = 0.4, \text{if } r_i = \text{Rich} \end{cases} \tag{3.16}$$

In this condition, poorer cooperators may be perceived as disloyal even though they contributed all available endowments to their group, and rich defectors may signal to their group that they are cooperating.

3.2. Treatment 2: 'Communication'

Following Cason et al. (2012), members of the same group are allowed to communicate via a public chat. In sessions where chat is allowed, it occurs after Stage 2 and before Stage 3 (see Figure 3.1). The opportunity to chat lasts two minutes in rounds 6 and 11, and otherwise one minute.[8]

The purpose of intragroup communication is threefold: to induce a stronger ingroup-outgroup boundary, to increase expectations for individual commitment to group efforts (see Cason et al. 2012; Sutter and Strassmair 2009), and to encourage the deliberation of loyalty among players of the same group. In contrast to conditions without communication, participants are able to justify their contributions to the competition fund, and a systematic analysis of chat messages will allow for a validation that participants 'take' the induced stimuli, in particular:

- Deliberate to identify defectors
- Accuse low contributors of defection (even if they choose not to punish them)
- Respond to suspicions of defection with defiance or counter-accusations

Since conflict resolution between groups is not relevant to this experiment, there are no opportunities for intergroup communication that might otherwise lead to an agreement on minimal effort in both groups (see Cason et al. 2012). Note that side-switching has no impact on communication (only contributions are switched and only for a given round), while ingroups are able to deliberate expectations for appropriate contributions, communication does not guarantee agreement or commitment to an agreed upon strategy.

4. Measurement and Hypotheses

The following discussion distinguishes between theoretical propositions, i.e. participant behavior that should be induced through game design features, and hypotheses, i.e. behavior change that should be caused by variation in treatments. Table 3.2 provides an overview of key measures. *Punishment intensity* is measured by actualized penalties for at maximum one player per group in a given round, and

[7] This is necessary since revealing participant types would reduce uncertainty induced by heterogeneous endowments.

[8] This is to shorten the duration of the experiment, while still leaving sufficient time for participants to discuss how changes to the game in rounds where they are introduced affect their decision-making.

Table 3.2. Key Measures

Variable	**Notation**	**Range**
Participant		
Type	r_i	['Random', 'Poor', 'Rich']
Cooperation, Cohesion	$c_i^e, C_{\sigma I}^e$	$\{0, 1\}$
Defection	$d_i \vert D_i$	$\{0, 1\}$
Punished	L_i	$\{0, 1\}$
Punishment intensity	s_i	$[0, 90]$
Group won competition	V_{iI}	$\{0, 1\}$
Round	t_i	$[1, 20]$
Group identification	x_{i1}	$[1, 10]$
Session		
Endowment Inequality	Z_1	$\{0, 1\}$
Communication (Chat)	Z_2	$\{0, 1\}$

defection by the decision to switch contributions to the opposing team, given the opportunity to switch. *Cooperation* is measured by the size of a participant's contribution as a proportion of their endowment:

$$c_i^e = \frac{c_i}{e_i} \tag{3.17}$$

Group cohesion is simply the standard deviation of cooperation within a group, i.e. cohesion is maximized when all group members contribute the same proportion of their endowment:

$$C_{\sigma I}^e = \sqrt{\sum (c_i^e - \bar{c_i^e})^2 \frac{1}{n-1}} \tag{3.18}$$

Finally, whether a participant's assigned group won irrespective of individual defection is given by:

$$V_{iI} = \sum_i^n v_{iI} > 0 \tag{3.19}$$

4.1. Inducing Loyalty Conflicts

First, the game should induce loyalty conflicts among participants. Moreover to the extent that there are overlaps in design features, the relationships between punishment, cooperation and defection should be consistent with existing experimental studies. Related experiments with Tullock contests between groups find that cooperation, measured in monetary contributions to a group, is increased by (a) higher prize valuation (Baik 1993; Sheremeta 2009), (b) communication and identification with groups (Cason et al. 2012; Charness et al. 2014; Mago et al. 2016; Sheremeta and Zhang 2010; Sutter and Strassmair 2009; Zdaniuk and Levine 2001), and (c) the possibility of punishment (Abbink et al. 2010). To the

extent that endowment heterogeneity is comparable to heterogeneity in valuation, it would be expected that higher endowments increase cooperation, as do communication (including group identification through communication), and punishment:

Proposition 1. *Rich participants cooperate more than poor participants.*

Proposition 2. *Group identification increases in sessions with communication.*

Proposition 3. *Cooperation increases in sessions with communication.*

Proposition 4. *Cooperation in Part 2 is higher than in Part 1.*

In typical contest experiments, cooperation has a negative connotation: the efficiency of a Tullock contest decreases in contributions (above zero or in some cases, above one token), and it is commonly found that expenditure exceeds the Nash equilibrium, pointing to behavioral factors that are not captured by the vast game theoretical literature on the subject (Dechenaux et al. 2015). By contrast in other experimental games with voluntary contribution mechanisms (VCM), such as 'public goods' and 'common pool resources' (e.g. Fehr and Gächter 2000; Henrich et al. 2006; Ostrom et al. 1992), cooperation is a positive outcome since 'defection' refers to relatively low contributions (or withdrawals in the case of a Common Pool Resource game), up the point where participants are free-riding on other people's effort entirely.

In real-world conflict settings, cooperation has a positive connotation to group members insofar as it demonstrates *loyalty*, and those whose contributions are insufficient may be punished for their *disloyalty*. But rather than being a purely intrinsic individual property (see Hirschman 1970; James and Cropanzano 1994), loyalty is socially constructed as personal sacrifice that enhances group welfare at the expense of a rival group (Levine and Moreland 2002). In the present study, contributions represent the personal sacrifice of group members in cooperation against a rival. Groups are expected to develop a norm around appropriate cooperation, rather than disregarding the past cooperation of others when choosing contributions (see Mago et al. 2016):

Proposition 5. *Cohesion of cooperation increases over time.*

The motivation and ability to demonstrate loyalty are private information in conflict settings. Intrinsic motivation, based on social identification (James and Cropanzano 1994; Tajfel and Turner 1986), cannot be discerned directly, but only approximated by observation of behavior. Loyalty in the context of the experiment is therefore based on "epistemic actions", which includes presuppositions about endowments, observations of decisions and statements, and the possible appearances of these actions to other participants (see Baltag 2002). Participants are expected to state uncertainty around the meaning of their contributions:

Proposition 6. *Participants state uncertainty about other's true cooperation in relation to their decisions.*

A participant's *intended* commitment may differ from *perceived* commitment, leading to misperceptions of loyalty. Existing conflict studies treat misperception as an information problem, whereby authorities falsely punish when they fail to identify disloyalty (e.g. Kalyvas 2006; Steinert 2022). Relatedly, Grechenig et al. (2010) suggest that punishment should be constrained where information is imperfect, as participants punish even when they know that they might unjustifiably sanction high contributors (see Herrmann et al. 2008). Overall, it is expected that punishment is related to observed contributions:

Proposition 7. *The likelihood of being punished decreases in contributions.*

However, even if intentions are revealed through communication or actions, participants can misperceive loyalty when (a) they do not trust the signal of the sender (see Farrell 1987), or (b) disagree on

appropriate contributions. It follows that punishments which appear justified to the sanctioning agent may or may not appear justified to the punished (see Sherman 1993; Tyler and Huo 2002). In the present study, any punishment may be deemed subjectively unfair, as there is no a priori commitment to a fixed strategy. It follows that 'poor' participants should be punished more frequently, given that they appear consistently disloyal for lower contributions:

Proposition 8. *Participants with the 'poor' type are more likely to be punished than those with the 'rich' type.*

Experiments in economics consider preferences for competing groups (McCarter et al. 2013) or group formation (Charness et al. 2014) without considering mechanisms for social exclusion as drivers of behavior. By contrast in social psychology, ostracism from the group is studied as a function of norm violation (Abrams et al. 2018; Castano et al. 2002; Ditrich and Sassenberg 2016; Lindström and Tobler 2018; Marques and Paez 1994; Travaglino et al. 2014; Zdaniuk and Levine 2001), however the behavior of former ingroup members is rarely considered after their ostracism. In conflict settings, disassociation from the group may lead to association with a rival group. For the present study to mirror this understanding, it is expected that participants consider side-switching in terms of group loyalty:

Proposition 9. *Participants frame the decision to switch their contributions in terms of (dis)loyalty.*

Since Proposition 9 and Proposition 6 concern participant accounts of decisions, they are evaluated through exploratory coding of chat messages and open survey responses rather than quantitative estimation. The conditional average treatment effect of communication on identification with the group (Proposition 2) is estimated with naive OLS:

$$\hat{x}_{i1} = \beta_0 + \beta_1 Z2 + \epsilon_i \tag{3.20}$$

Propositions related to *cooperation* outcomes (1, 3, 4, 5) are evaluated using panel regression models with random effects for participants, controlling for communication and team identification (Proposition 2):

$$\begin{aligned} \hat{c}^e_{it} &= \beta_0 + \beta_1 r_i + \beta_2 Z_2 + \beta_3 (t_i > 6) + \beta_4 x_i + \epsilon_i + \mu_{it} \\ \hat{C}^e_{\sigma I} &= \beta_0 + \beta_1 t_i + \beta_2 Z_2 + \beta_3 x_i + \epsilon_i + \mu_{it} \end{aligned} \tag{3.21}$$

In turn, the prevalence of punishment (Propositions 7 and 8) is evaluated with a linear estimator proposed by Lin (2013), which centers covariates (here: participant contributions) and interacts them with randomized treatments (here: player type) to reduce biases in OLS estimations of experimental treatments as described in Freedman (2008):

$$\hat{L}_i = \beta_0 + \beta_1 r_i + \beta_2 c^e_i + \beta_3 c^e_i r_i + \epsilon_i \tag{3.22}$$

For all estimators, standard errors are clustered at the session level.

4.2. Effects of Punishment on Defection

Following the labeling approach to social deviance (Becker 1963; Lemert 1951; Matza 2010), defiance and procedural justice theories (Sherman 1993; 2014; Tyler and Huo 2002), it is expected that punishment has an exclusionary effect on the punished. A participant who is punished may come to accept their role

as a disloyal member, up until the point where they feel excluded from their group. Starting in Part 2, the opportunity to disassociate from ones group is to lower contributions (loyalty as group support), and starting in Part 3, participants who are punished are expected to defect to the opponent (loyalty as group membership). On average, participants who were punished are expected to decrease cooperation and increase defection:

H1. *Punished participants decrease their cooperation in the subsequent round compared to participants who were not punished.*

H2. *When given the opportunity to switch sides, punished participants are more likely to defect than participants who were not punished.*

The estimands for these hypotheses are the (conditional) average treatment effects of punishment (given the opportunity to defect) on cooperation and defection, respectively:

$$\begin{aligned} ATE_{c^e_{it}} &= \sum_{i=1,t=7}^{n,20} \frac{c^e_i(L_{it-1}=1)}{n} - \sum_{i=1,t=7}^{n,20} \frac{c^e_i(L_{it-1}=0)}{n} \\ CATE_{d_{it}} &= \sum_{i=1,t=11}^{n,20} \frac{d_{it}(L_{it}=1|D_{it}=1)}{n} - \sum_{i=1,t=11}^{n,20} \frac{d_{it}(L_{it}=0|D_{it}=1)}{n} \end{aligned} \tag{3.23}$$

Since punishment is endogenous to the game in all experimental conditions, the estimate of this effect must control for within-subject covariates. First, the intensity of the penalty may increase the exclusionary effect of punishment: on average, smaller penalties that have a minimal impact on payoffs are expected to be tolerated, whereas larger penalties are expected to decrease cooperation and increase defection. And second, winnings in prior rounds and higher competition funds are expected to increase participant confidence that loyalty is rewarding, decreasing defection. As above, the Lin (2013) adjustment is used to reduce bias in the OLS estimator for the treatment effect:

$$\begin{aligned} \hat{c}^e_{it} &= \beta_0 + \beta_1 L_{it-1} + \beta_2 s_{it-1} + \beta_3 V_{iIt-1} + \beta_4 s_{it-1} L_{it-1} + \beta_5 V_{iIt-1} L_{it-1} + \epsilon_i \\ \hat{d}_{it} &= \beta_0 + \beta_1 L_{it} + \beta_2 s_i + \beta_3 V_{iIt} + \beta_4 s_i L_{it} + \beta_5 V_{iIt} L_{it-1} + \epsilon_i \end{aligned} \tag{3.24}$$

As an alternative estimation strategy to the Lin estimator, a panel regression model as above is considered:

$$\begin{aligned} \hat{c}^e_{it} &= \beta_0 + \beta_1 L_{it-1} + \beta_2 s_{it-1} + \beta_3 V_{iIt-1} + \epsilon_i + \mu_{it} \\ \hat{d}_{it} &= \beta_0 + \beta_1 L_{it-1} + \beta_2 s_{it} + \beta_3 V_{iIt} + \epsilon_i + \mu_{it} \end{aligned} \tag{3.25}$$

The study is designed to test whether inequality and lack of communication about disloyalty cause defection (Z_1, Z_2). First, in sessions where endowments are unequally distributed, 'poor' participants are unable to demonstrate loyalty as much as other group members, more likely to be punished, and hence more likely to defect. While this experience only applies to two participants per group, they may be repeatedly excluded and defect as a result. Moreover, the defection of 'poor' participants might lead to fragmentation in the group, particularly when they lose to the opponent, increasing defection by those who are not punished.

H3. *Participants in groups with unequally distributed endowments are more likely to defect, compared to participants in groups with equally distributed endowments.*

Second, in sessions with communication within groups, participants are able to coordinate their strategies with respect to cooperation, punishment and defection. They are more likely to develop a shared expectation for appropriate contributions, less likely to misperceive each others cooperation as disloyal, and less likely to be punished unjustifiably. Moreover, given that identification with the group increases in the chat condition, members are less likely to want to abandon their group without being punished:

H4. *Participants who can communicate via chat are less likely to defect than those who cannot.*

Finally, sessions that combine communication and endowment inequality allow participants to signal their unequal endowment distributions, reducing punishments due to misperceptions, and attempt to dissuade group members from choosing to defect. Some groups may fail to dissuade its members from disloyalty, or coordinate to refrain from punishment in order to prevent defection. Conversely, they might not trust 'poor' participant's consistent signals that their endowment is low, or claim to have a low endowment themselves. But compared to the condition with only endowment inequality, defection is reduced by communication and the associated group identification.

H5. *Participants who can communicate about unequal endowments are less likely to defect, compared to participants in the Inequality condition without communication.*

The estimands for these hypotheses correspond to a series of conditional average treatment effects and their difference:

$$\begin{aligned} CATE_{H3} &= \sum_{i=1}^{n} \frac{d_i(Z_1 = 1|Z_2 = 0)}{n} - \sum_{i=1}^{n} \frac{d_i(Z_1 = 0|Z_2 = 0)}{n} \\ CATE_{H4} &= \sum_{i=1}^{n} \frac{d_i(Z_2 = 1|Z_1 = 0)}{n} - \sum_{i=1}^{n} \frac{d_i(Z_2 = 0|Z_1 = 0)}{n} \\ CATE_{H5} &= CATE_{H4} - CATE_{H3} \end{aligned} \tag{3.26}$$

A series of OLS regressions suffices to estimate these effects without participant covariates:

$$\hat{d}_i = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2 \tag{3.27}$$

5. Results

This section analyzes results from six comparable pilot sessions with $N = 48$ participants, though only 3/6 sessions correspond exactly to the design described above, and the current number of conditions and participant types is not balanced (see Appendix C.I for details). Moreover, two sessions had the highly unlikely outcome of one team winning 95% (*Comm*) and 100% (*Ineq-Comm*) of the time, despite initially similar cooperation, which led to slightly aberrant behavior.[9] The focus is therefore on analyzing to

[9] In the *Comm* condition, the highly unlikely outcome of one team winning 95% of the time, despite initially similar contributions, lead the other team to believe that the game is "rigged" and coordinate on low cooperation and defection. By the same token, the winning team eventually decided that low contributions would allow them to win the game more efficiently, given low opponent contributions (see Cason et al. 2012; Sutter and Strassmair 2009). Even more unlikely in one

what extent the design induces loyalty conflicts, by evaluating the propositions from Section 4.1. Given the low number of observations I do not attempt to interpret the (non-)significance of the results, but they are shown along with effect sizes in Appendix C.I.3.

5.1. Cooperation

Figure 3.2 shows how cooperation changes over the course of a session in different treatment conditions.[10] In line with Proposition 4, there is a slight increase in cooperation with the onset of punishment ($\hat{\beta} = 4\%$). However in conditions with communication, participants appear less worried that they might be punished, because they can signal endowments, justify contributions, and agree not to punish each other. This sense of unity is reflected in the effects of communication on team attachment. Figure 3.3 (B)-(D) illustrate how communication increases team attachment ($\hat{\beta} = 3$) in line with Proposition 2, yet how counter to Propositions 3 and 5, this identification translates to slightly lower levels of cooperation ($\hat{\beta} = -1\%$) and cohesion ($\hat{\beta} = -2\%$). This may be an artifact of the few *Comm* sessions run thus far, where there was a tendency to coordinate on low cooperation particularly in two almost permanently losing groups. Yet, it may also suggest that communication merely increases the willingness of group members to conform with a perceived group consensus, which is not necessarily that cooperation is beneficial.

The expectations for Proposition 1 are similarly reversed: Figure 3.3 (A) shows how rich participants contribute on average *less* of their endowment than poor participants ($\hat{\beta} = -11\%$). This suggests that there is a fundamental difference between the private *opportunity* to demonstrate loyalty and the *motivation* to do so. When the private motivation to contribute to group efforts is either known or not questioned, group members may conform to implicit norms around appropriate effort. When loyalty is perceived but not known, those who can demonstrate it are more easily perceived as cooperators. Rich participants keep a larger proportion of their endowment to themselves, compared to poor participants who must match the effort of other group members in order to be seen as cooperative.

5.2. Punishment and Defection

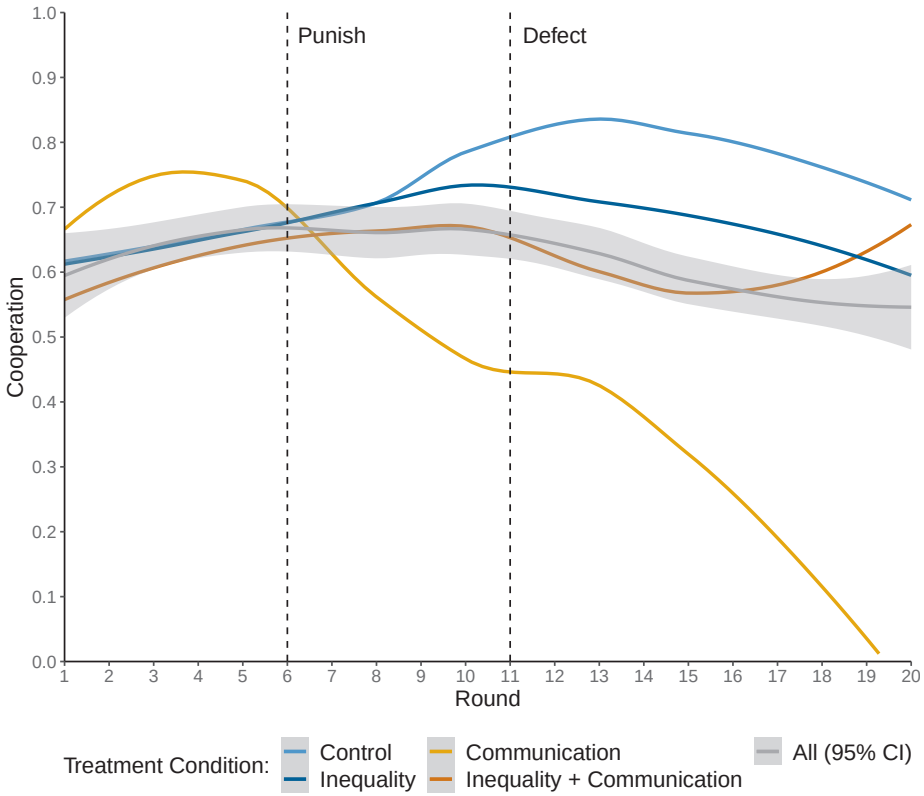
On average, participants punish a group member in 10% of rounds (similar to Abbink et al. 2010, 435), with $s_i = 21$ median punishment intensity. This is plausible seeing as almost all participants understand that punishment is inefficient and costs their peers real-world currency. Though participant responses in the survey suggest that they punish low contributors in line with propositions 7 and 8,[11] these could not be validated with the proposed estimator: while punishment decreases in contributions and rich player types, the effect size is extremely negligible. This is possibly due to the limited number of punishment observations and imbalance in participant types in the sample, but also due to the dampening effect of communication on punishment in general: poor participants were punished $\Delta = 27\%$ more frequently than rich participants in the *Ineq* condition without communication (19 vs. 11 times), compared to $\Delta = 16\%$ in the *Ineq-Comm* condition (18 vs. 13 times).

of the three *Ineq-Comm* conditions, one group consistently lost despite, at some point, contributing slightly more than the other, leading at least some participants to treat the size of their contributions as irrelevant and decreasing communication in the chat.

[10] In general based on survey feedback, participants adjusted their cooperation in response to (1) strategic considerations and learning (e.g. previous contest winnings, other team member's contributions), (2) team agreement (in *Comm*), (3) being punished, (4) endowments received, (5) experimentation ("curious about the result"), or (6) tried not to adjust and stick to a given proportion.

[11] Based on survey feedback, participants punished when (1) contributions were below expected or agreed upon levels, (2) after being punished themselves, (3) to avoid being punished by others (in *NoComm* conditions), or (4) never because (a) it was costly/inefficient, (b) their team agreed not to (in *Comm* conditions), or (c) it was otherwise deemed 'inappropriate'.

Figure 3.2. Cooperation by Experimental Condition



Note: LOWESS for cooperation over time by experimental condition, administered at the session level.

Figure 3.4 shows the defector count for each experimental condition, with the coloring indicating the proportion of defectors who were punished. In the *NoComm* conditions, the number of punished exceeds that of 'strategic' defectors who switched their contributions without being punished ($\hat{\beta} = 69\%$). Moreover, participants who won the competition in the previous round defect less than those who lost, though the effect size is smaller ($\hat{\beta} = -33\%$). This corroborates the notion that participants rely on strategic considerations to decide whether or not to switch sides in the absence of punishment,[12] and that punishment increases defection particularly in the absence of communication (H 2).

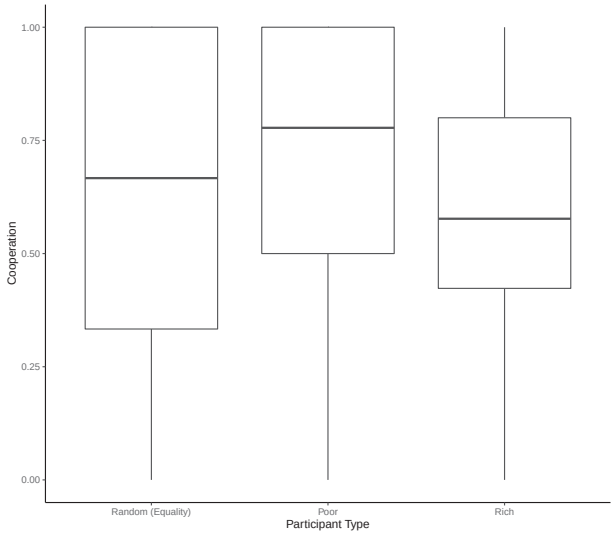
In line with H 4 and H 5, communication decreases defection compared to control ($\Delta = -17.5\%$) and inequality ($\Delta = -2.5\%$). Yet counter to H 3, inequality decreases rather than increases defection in the absence of communication ($\Delta = -15\%$). These effect sizes are highly suggestive at best, seeing as defection was relatively high in the single pilot run of the control condition (62.5%).

5.3. Social Construction of Loyalty

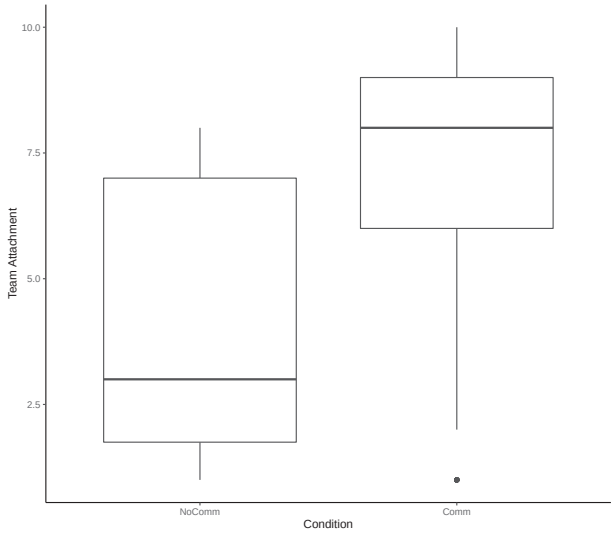
Chat messages and survey responses provide a less general yet more accurate picture about the extent participants 'took' the treatments designed to induce loyalty conflicts. First in line with Proposition 6,

[12] Survey feedback suggests that participants defected when (1) team loss was anticipated (low competition fund, previously losing), (2) after being punished, (3) the team agreed to (this overlapped with the team losing unusually often), (4) experimentation/curiosity, and (5) never due to self-identifying as a "loyal member" and/or the team agreeing not to (in *Comm* conditions). Some participants note that they were "tempted" but still remained loyal to conform with the team consensus.

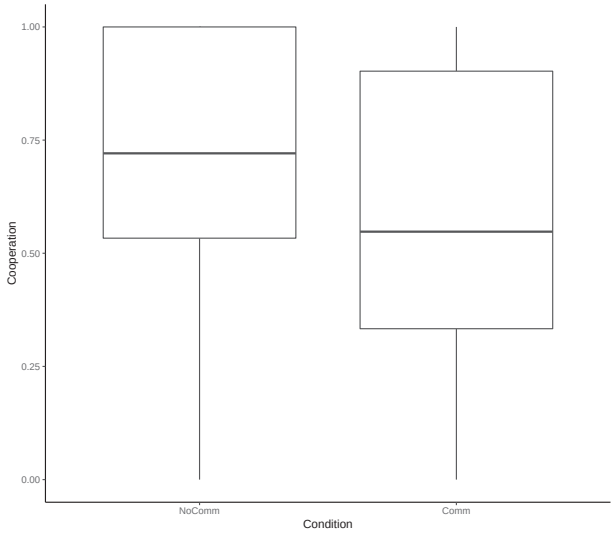
Figure 3.3. Cooperation Propositions (1, 2, 3, 5)



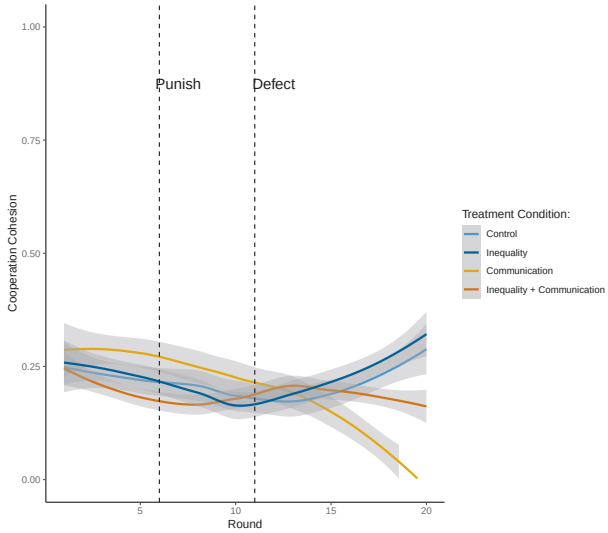
(A) Type and Cooperation



(B) Communication and Team Attachment



(C) Communication and Cooperation



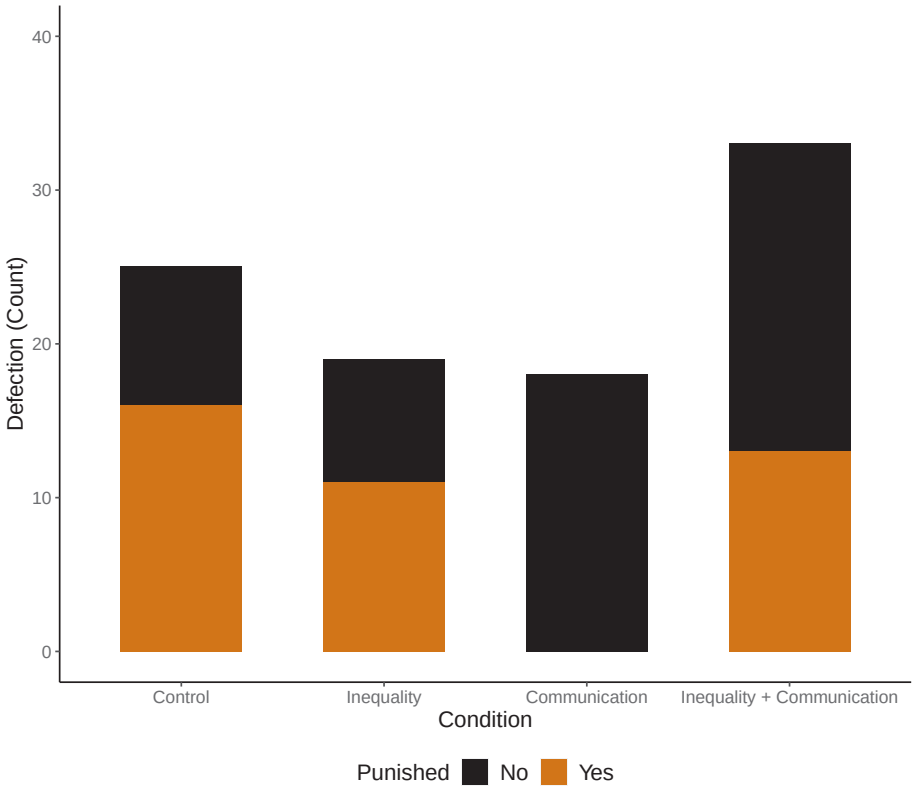
(D) Cohesion by Condition

Note: Data visualizations for propositions related to cooperation that can be measured with quantifiable data.

participants were uncertain about each others endowments, and used the chat to reduce that uncertainty in *Comm* conditions. Most chat messages were geared towards coordinating optimal contributions, and participants tended to be honest about declaring their endowments. But they still voiced suspicions of disloyalty as signals about endowments were not necessarily believed, which could lead to punishment. For example, in a group that decided to punish the lowest contributor in each round:

Player 2: "ok guess it is 3 now"
Player 2: "even though I dont trust 1 to always go all in"
Player 4: "I don't think we should penalise 3 because they've gone the most in overall"

Figure 3.4. Defection by Experimental Condition



Note: Absolute defector counts among participants given the opportunity to defect, by treatment condition and punishment.

> Player 3: "you must beleave =_)))"
> Player 2: "if the assignment is random, u should not have so much lower"
> Player 1: ":)"
> (*Comm* condition, 28.09.2022, Round 9).

Relatedly in the survey, participants describe how they attempted to deceive their peers about their true cooperation:

> "I tried to take points in a multiple of 20 so it didn't look like I was taking any for myself, once penalty points became a thing, as I felt like if I looked to be giving all my points as competition tokens then I SHOULD be less likely to be penalised."
>
> "Although we agreed as a team at first to go all in I always kept a spare amount in order to maximize my earnings while (trying) to give my team the impression that I was going all in so that they would so our chances of winning would increase."

Deception to avoid punishment was expected, but not deception to maintain the motivation of others to cooperate. Communication could therefore increase cooperation irrespective of identification with the group, as some members trick their teammates into levels of cooperation that they themselves do not observe.

Second in line with Proposition 9, participants understand side-switching in terms of 'loyalty', e.g. in the following chat messages:

> Player 2: "yeah loyalty for the team I would propose"
> ...
> Player 4: "the other team must be choosing to switch"
> Player 2: "they dont seem to have any sense of loyalty :)"
> (*Comm* condition, 28.09.2022, Round 11,13)

> Player 2: "I plan on being loyal if chosen"
> Player 1: "same"
> (*Ineq-Comm* condition, 09.09.2022, Round 11).

Similarly in response to the survey question whether and when to side-switch:

> "never – loyal member"
> "I didn't switch, I felt it was better to be loyal since we were supposed to be working together."
> "never switched. we seemed to always win anyways and felt loyal to my tem"

Jointly, inequality and punishment can lead to defection, particularly when participants feel unreasonably punished. This is evident from survey responses as to why participants defected, for instance:

> "I did so increasingly as a matter of principle, being annoyed with my own team as I was persistently being penalised regardless of how much money I invested. In the last instance, I realised that my decision to switch was not even rationally motivated, but due to annoyance at my team."

Unexpectedly, uncertainty did not always evolve around opportunities to demonstrate loyalty, but can also evolve around conflicting expectations for appropriate contributions. This happened explicitly in one instance in an (earlier pilot) *Comm* condition, where a participant was 'singled out' for refusing to cooperate with the group, leading to a sequence of punishment, decreasing cooperation, and group fragmentation:

> Player 3: "Played it mean this time."
> Player 4: "Feeling luckier this round [...] doesn't mean I was being mean though"
> ...
> Player 2: "if you receive more than 100, you should put 100, if less then at least half of that"
> Player 4: "there are no moral imperatives because you got put max and everyone could put max and still not win"
> ...
> Player 2: "0??"
> Player 4: "yes it is fair given I got penalised"
> ...
> Player 4: "if I get penalised I will put in 0 you can bully all you like"
> ...
> Player 4: "I wanted to make a contribution but was very clear that if I go penalised I would put in zero"

..
Player 4: "I got penalised every time last time so went on what I said I would do if penalised again"
Player 2: "it seems all of us will send it to the other team :)"
(*Comm* condition, 12.08.).

It may be fruitful to code such sequences systematically after the main study, in order to assess the extent to which punishment is understood as a label that leads to exclusion from the group.

6. Discussion

In terms of the experimental design, a key issue pertains to statistical power of the treatment effects. Only half the participant pool has an opportunity to defect for ten rounds, and each of the *CATE* are estimated for only half the total sample. Drawing on work by Blair et al. (2019), a simulation-based power analysis of the design suggests that with current effect sizes, resampled to $N = 240$, the probability of obtaining a statistically significant result ranges from 25% for H 3, over 37% for H 4, to 75% for H 5. In other words, none of the estimated causal effects would reach the conventional target of 80% of statistical power if effect sizes were to remain at current levels during the main study.[13] This is slightly worrying, but could be due to the limitations in the pilot data and the particularly experienced participant pool at the CESS over the summer (see Appendix C.I).

The game design fares reasonably well when it comes to inducing loyalty conflicts, on top of the competitive incentives that are usually induced with group contests. First, even when participants can declare their endowments, they are uncertain about how much other team members are truly 'sacrificing'. By itself this uncertainty does not necessarily lead to punishment: participants do not consistently attribute losses to other group members, accepting that there is a random chance that they might lose despite acceptable contributions; and they tolerate lower contributions than their own, accepting that there is disagreement about appropriate cooperation levels. But they punish poor participants more than rich participants, which suggests that the *Inequality* treatment is taken.

Second, the *Communication* treatment is taken as well, given the significantly higher levels of team attachment and coordination compared to conditions without communication. Participants associate loyalty with staying in their team, and consider it less appropriate to defect in the *Comm* condition. They do not explicitly identify and punish defectors who switch sides, which was expected given that there are no straightforward signals to learn who had the opportunity to defect, and whether or not they did.

Overall, the design of the game goes a long way towards inducing the necessary conditions for loyalty trials. The laboratory context lacks important contextual factors, such as the fear and risks to defection that the Palestinians and East Germans in the motivating examples presumably experienced. But like those individuals, participants in the game faced trade-offs between personal and group goals, were induced to cooperate with these goals, and punish those who failed to cooperate. We reduce demonstrations of loyalty to monetary contributions in the lab, but note that personal sacrifices in the real world may entail a wide range of behaviors that have nothing to do with financial resources. By the same token, some constraints on behavior may be visible in the real world (e.g. whether somebody

[13] This is based on simulated data with $N = 240$, using the main effects of treatment conditions as the 'true' difference in the probability to defect for each participant. After assigning a hypothetical side-switching decision for each participant-round across conditions, and randomly assigning a realized treatment condition, the estimate for all possible *CATE* and *ATE* is compared against the true outcome. The process is simulated $S = 500$ times. The mean number of times that the null is rejected when it should be (p-value $\leq .05$) is the power for each effect.

is too poor to pay authorities), but the distance between loyal *intentions* and *perceptions* is similarly difficult to establish in the real world as it is in our study. As a caveat, though side-switching was seen as disloyalty by participants, our design did not link ingroup contributions and defection, such that participants did not attempt to identify side-switchers among them. Introducing additional design features that frame loyalty conflicts more explicitly in terms of side-switching is an avenue for future research.

FINAL REMARKS

The papers in this manuscript make complementary arguments about the relationship between the labeling of political deviance as disloyalty, on the one hand, and the consequences for defection, on the other. The first paper shows how labeling affects allegiance based on the same mechanisms in vastly different context. We find that labeling deviance as disloyalty increases defection particularly when quotidian behavior is perceived as disloyal by group members, and that tolerance for political deviance increases conformity. And we show how loyalty trials always polarize conformers against defectors, though the latter are minorities in most conflict settings. Of particular interest to the study of conflict is perhaps the relationship between loyalty expectations and allegiance shifts. Governments who feel existentially threatened are unlikely to change loyalty expectations or encourage tolerance for political deviance even when they are internationally shamed for human rights violations. Where possible, the practices that governments use to protect defectors from their rivals should be as transparent and stable as the norms that ensure the sovereignty of their own rule, and decision-makers should take the consequences for defection in their own communities into account. To reference a salient example at the time of writing: countries who let the risk of infiltration by Russian agents and popular resentment of the Ukrainian invasion guide restrictive visa policies not only deny opportunities for 'exit' to aggrieved Russians in the country, but also promote the image that 'being Russian' is tantamount to disloyalty, with the associated consequences for discrimination and defection.

The second paper shows how allegiance is co-produced between authorities and their subjects, drawing on the notion that political deviance is labeled based on three 'moments' of loyalty. Where labeling introduces conflicts between 'images' of what constitutes defection and perceptions of disloyalty, or conflicts between perceptions and the 'true' allegiance of the labeled, allegiances are prone to shift. In that regard, I corroborate a key mechanism from the first paper, namely that false defectors are more likely to conform than defect, compared to true defectors whose labeling has little impact on their political deviance. Moreover, I highlight the importance of third-moment loyalty and unofficial labeling, which are rarely considered in mainstream conflict studies, but quite well understood in interpretive scholarship. Authorities might 'correctly' label individuals as defectors, but in contrast to criminal deviance, what constitutes defection could still be legitimately 'false' to the labeled (as in the GDR). And by the same token, authorities might refrain from labeling defection that is very much considered unacceptable to group members, as is the case among Palestinians for 'security coordination' with Israel today (Albzour et al. 2019). Particularly in civil war settings, therefore, it might be preferable to enforce official norms against defection that are in line with unofficial perceptions and increase tolerance for political deviance, as opposed to refraining from official labeling that encourages group members to label already marginalized social deviants.

From these two papers, labeling appears to *reinforce* prior disloyalty rather than causing 'primary' defection: given that an individual is labeled, allegiance shifts towards conformity are more common than shifts from conformity towards defection. This is plausible both when defectors are vilified—there are strong social incentives to conform—and when they are glorified—initial disloyalty is intrinsically motivated. However, in line with similar concerns in criminology (see Sherman 2014), there is the possibility that the effects of labeling are under-estimated in observational studies, seeing as unofficial

labels that might lead to the defiance of third-moment images are rarely observed systematically.

The third paper attempts to address this concern, by studying a novel game that induces key conditions of loyalty conflicts in the lab. Results from the pilot study give some support for the notion that punishment leads to defection. The design induces most of the key aspects of loyalty dynamics in conflict, though when communication is allowed, some participants identify with their group sufficiently to resist the temptation of punishing disloyalty and defecting entirely, and at times tolerate the defection of others in the face of endogenous conflict asymmetries. This is in line with the micro-dynamics of labeling in real-world conflict settings: defectors are not uniformly vilified, as can be seen from Paper 2 where punishment and defection occurred mostly in the presence of a strong social norm against disloyalty, but did not occur or affect allegiance otherwise. However, the motivation behind the paper is to study the effects of punishment given that defection is threatening the group, and future designs might want to test different framing conditions to analyze how they affect behavior.

Overall, this project contributes to our understanding of conflict by viewing the stability of political order through the quotidian behaviors of 'ordinary' individuals beyond activists and rebels, by considering the relationship between international rivalries and domestic repression, and advocating for a wider understanding of political behavior as a function of loyalty expectations and labeling. It crosses disciplinary boundaries by integrating research from conflict studies, sociology, history, and social psychology. And it introduces a novel perspective on the defiance of authoritarian rule in the GDR, as well as on 'collaboration' with Israel in Palestine today. Aside from corroborating some of the propositions in this manuscript with additional evidence, future research might examine the applicability of these findings to other fields of study. How does the ostracism of minorities affect their political behavior in the community they end up with? When political parties exclude members for violating core ideological principles, do the labeled correct their views or join a different party? Under what conditions does the labeling of governments for violation of international norms lead to an exclusion of its diplomats from international forums, and does such exclusion strengthen or weaken the norm itself? As was hopefully shown here, particularly when the goal is to integrate different methodological approaches, the lens of labeling moments of loyalty could be helpful to answering such questions.

PAPER 1 APPENDICES

Appendix

A.I. Agent Fitness

Figure A.1 illustrates how agent fitness (y-axis) varies with deviance (x-axis), loyalty expectations and incentives (line colors represent different structural conditions), and the perceived deception of labeled defectors (line style represents the agent's condition). First, across all lines, disloyalty is beneficial as personal sacrifice is reduced: agent fitness increases with deviance from loyalty expectations (δ_i^2).

Second, loyalty expectations interact with loyalty incentives. When incentives are low but loyalty expectation are high (k = -1, $\lambda = 1$, sienna curve), the marginal benefits for disloyal behavior increase. And when both incentives and loyalty expectation are high ($k = 1$, $\lambda = 1$, black curve), the marginal increase in benefits for disloyal behavior remain low. In contrast, when expectations are low and loyalty incentives negative, (k = -1, $\lambda = 0$, grey curve), the marginal increase in benefits for loyal behavior remain low. Where loyalty incentives are balanced given moderate loyalty expectations ($k = 0$, $\lambda = 0.5$, red curve), defection and conformity are equally beneficial.

Third, where the disparity between private and labeled behavior is large ($k = 0$, $\lambda = 0.5$, $|i - p|l_A = 0.5$, dotted line), fitness is significantly lower than where the disparity is small ($k = 0$, $\lambda = 0.5$, $|i - p|l_A = 0.1$, dashed line).

A.II. General Model Robustness

The model was implemented in Python (Van Rossum and Drake Jr 1995), drawing on several packages for processing and data visualization (Harris et al. 2020; Hunter 2007; Tange 2011; Virtanen et al. 2020; Waskom 2021; Wes McKinney 2010). The code is available on the website of the corresponding author.

We conduct two types of robustness checks. First, we sweep the same parameters as for the general model in Section 5.1, and analyze additional model outcomes. Second, we sweep additional parameters and discuss how they affect the results, compared to a baseline that makes agents equally likely to increase conformity or defection. In the baseline, initial allegiance, loyalty expectations and incentives are set to $\bar{i} = \bar{p} = \lambda = k = 0.5$, while other auxiliary parameters are kept at the values used in the general model (see Table A.4). Table A.1 gives an overview of outcomes for all parameter sweeps compared to the baseline.

A.II.1. Type I and Type II Errors

We show how the model captures the micro-dynamics of labeling defectors. We define misidentification as the ratio of labeled and true defectors, formally:

$$\rho = \frac{A^{III} + A^{IV} + 1}{A^{II} + A^{IV} + 1} \tag{A.1}$$

ρ represents the trade-off between types of errors political authorities make when conducting loyalty trials: 'Type I' errors when defectors are over-identified—more defectors are falsely labeled than hidden ($\rho > 1$); and 'Type II' errors when defectors are under-identified—fewer defectors are falsely labeled than hidden ($\rho < 1$). When $\rho = 1$, defectors are perfectly identified, such that only true conformers and true defectors are observed.

Figure A.2 (A) shows mean identification for combinations of loyalty expectations and initial behavior. 'Type I'-errors occur when loyal behavior initially exceeds expectations, and 'Type II'-errors when it does not meet expectations: defection is on average under-identified where $\lambda > \bar{i}, \bar{p}$, and over-identified where $\lambda < \bar{i}, \bar{p}$. In settings where initial loyalty far exceeds expectations, there are no loyalty trials and therefore no misidentification occurs ($\rho = 1$). Where behavior meets or slightly exceeds expectations, agents shift allegiance only indirectly, in observing loyalty trials that correctly identify true defectors and conformers. But overall, allegiance shifts rarely occur without misidentification.

ρ differs from misperceptions, which may be based on the quality of information used to label defectors, or on biased usage of available information. Misperceptions are given by the distance between perceived and private behavior, formally:

$$\Delta_{pi} = \bar{p} - \bar{i} \tag{A.2}$$

Misperceptions may lead to misidentification, contingent on sufficiently high expectations and low group tolerance. Figure A.2 (B) shows the most prevalent defector type at the end of the simulation for each possible value of Δ_{pi}, with all other parameters as in the baseline ($\lambda = 0.5$). Assuming loyalty expectations are initially not met, when $\Delta_{pi} > 0$ secret defectors are above suspicion (black cells, bottom right quadrant), and when $\Delta_{pi} < 0$ true defectors are labeled as if they engaged in severe disloyalty (sienna cells, bottom left quadrant). False defectors almost always constitute a minority, such that true conformers are on average the most prevalent defector type even when loyal individuals are initially perceived as disloyal (grey cells, upper quadrants). Irrespective of misperceptions, allegiances eventually shift toward defection when extreme loyalty exceeds moderate expectations.

A.II.2. Auxiliary Parameters

Agent Characteristics

Figure A.3 shows defector type probability by (A) mean group tolerance, (B) spread of tolerance and (C) spread of private and public allegiance. (A) As tolerance increases, true defection decreases and true conformity increases. Once tolerance for defection exceeds disloyal behavior, secret defection approaches true conformity, while false defection decreases. Compared to the baseline, the probability of labeled defection is higher when agents are either perfectly intolerant or tolerant of defection. (B) As heterogeneity of tolerance increases, true conformity slightly increases as true defection decreases. Maximum heterogeneity essentially makes all agents either perfectly tolerant or intolerant, which is sufficient to trigger cascades of defection. Misidentification is not affected by the spread of tolerance compared to the baseline. (C) Behavior heterogeneity generally increases defection, given borderline conformity with moderate loyalty expectations. Defection slightly exceeds conformity where the spread of behavior is equal to or smaller than the spread of tolerance (here: $\sigma_i, \sigma_p \leq \sigma_q = 0.1$).

Overall, changes to tolerance and allegiance distributions lead to minor changes in model outcomes, but only for extremely low or high values when they change who is subjected to loyalty trials. Compared to the baseline, extreme increases in the (heterogeneity of) tolerance and initial loyalty lead to increases in conformity, and vice-versa for defection.

Simulation Parameters
Experimentation with the number of agents and generations shows that setting $N, G = 1000$ is sufficient to converge on a given distribution of defector types. Other simulation parameters delineate spatial and temporal aspects of model runs. Figure A.4 shows the probability of defector types given changes in (A) agent pairings, (B) agents subjected to and affected by loyalty trials per generation, and (C) probability of mutation.

P represents the frequency of labeling opportunities, and therewith the ability of individuals to alter perceptions of each others loyalty. The probability of true defection tends to increase with such opportunities, but all else equal the marginal change is minimal beyond three interactions per period. Notably, increasing opportunities to label in each generation leads to a quicker convergence on defection, and therefore more true defectors on average. T represents the size of the population who learn of and are affected by loyalty trials. Given that T affects which agents are subjected to loyalty trials in every generation, the probability of true conformity and defection fluctuates, but which type dominates does not change beyond extreme values. True defectors are more prevalent when all agents are affected by loyalty trials in every generation, and secret defection dominates true conformity when agents are barely affected by loyalty trials. Finally following Riolo et al. (2001), M controls the randomness of the evolutionary adaptation mechanism, where higher values increase the probability that agents adopt characteristics from the initial distribution of parameters. As for other simulation parameters, mutation changes outcomes when set to extreme values that preclude mutation or regularly re-seed the agent distribution.

A.III. Extension Robustness

A.III.1. Source Material on Defection

We discuss the reliability and validity of the empirical materials that were used for this paper. First, we briefly discuss the primary data from archives and interviews that was used to inform the model specification. Second, we discuss existing empirical data on defection and justify why we do not deem it suitable for model contextualization or validation. Qualitative evidence for model contextualization stems from secondary sources, as discussed in Section 5.2.

Model Specification
For the GDR, primary data was collected from the Stasi Archives. Over 111 kilometers of records were secured after the collapse of the regime, including internal instructions, reports by informers, denunciations, interrogation protocols, and descriptions of individual activities prior and after they were labeled. Given that these documents were not intended for publication, they are more reliable than other accounts of defection, such as from public news reports (see Balcells and Sullivan 2018; and Horz and Marbach 2022; Maddrell 2013; Piotrowska 2020; Steinert 2022 for recent large-N studies that draw on the archives). As part of this study, a purposeful sample of 453 files was reviewed between February 2020 and November 2021. Time to review files in the reading rooms was limited, the use of analytical software on original documents prohibited, and random sampling strategies would not yield a large share of complete files on defection. Therefore, files were purposefully sampled to maximize internal validity, based on three criteria: (1) completed surveillance cases, (2) targeting individuals, (3) who were suspected of defection. We use some of this evidence to illustrate the effects of labeling on defection, and more generally refined the model specification based on it.

For the OPT, primary data was collected in semi-structured interviews with 20 respondents during two months of fieldwork in Israel and the West Bank in November 2019 and February 2022. Among the

respondents were Palestinians targeted by Israeli security (or their parents in cases where the targets were children), officials familiar with recruitment strategies, journalists, as well as researchers at human rights organizations and universities. Each interview lasted between one and three hours. Interview questions were adapted to the context of the interview and respondent backgrounds, but generally covered the following topics:

1. Targeting of repression: who is labeled as a defector based on which criteria?
2. Source of threat perceptions: how is the criminalization of behavior legitimized by political actors?
3. Changing loyalty expectations: how and when did delineations of acceptable and unacceptable behavior change?
4. Tolerance: under what conditions is disloyal behavior acceptable?
5. Reactions by the labeled: how do targeted individuals respond to experiences of repression in the short- and long-term?
6. Reactions by peers: how do third parties (family members, neighbors, co-workers, superiors, rival authorities) respond to labeling?

As with the archival data, interviews were used to inform the model specification and interpret evidence on loyalty trials during the Second Intifada (see Table A.8 for a list of interviews).

Statistical Evidence on Defection

Table A.2 summarizes loyalty expectations and estimates for the proportion of defecting subjects in each conflict setting. Available data is at best a poor approximation of true defection, with no indication of other defector types. Note however that in line with our theoretical framework, the overall share of the population that reportedly defected increases with the personal sacrifice required to demonstrate loyalty, though it constitutes a small minority in both settings. Our model merely represents relative increases in conformity and defection, but does not predict this 'true' proportion of defectors among conformers.

Table A.3 provides an overview of the sources we used to generate Table A.2. While we consider these authors reliable secondary sources, the primary sources lack reliability and internal validity due to biases in reporting by primary sources, coverage and coding of cases.

First, given that our conceptualization of defection includes any form of deviant behavior perceived to threaten political order, much defection is not documented in official statistics. It is only through a detailed examination of the materials that we can observe the criminalization of quotidian behaviors, and at times distinguish between observed and fabricated accusations of more exceptional disloyalty (e.g. BArch, MfS, HA IX, Nr. 25283, Bl. 34-36; MfS, HA IX, Nr. 25609, Bl. 9-127; Kelly 2010, 179).

Second, the evidence that is observable suffers from biases in government sources (see Becker 2017). Such biases stem from attempts to charge individuals with criminal offenses when there was insufficient evidence for political motivations or outgroup affiliation (e.g. Joester 1999, 317-318), and more generally from misidentification of defectors (e.g. Jalal 2015; Human Rights Watch 2001, 47-48; Williams 2001, 32-36).

Third, the coverage of statistical data is insufficient to precisely contextualize our model for the chosen periods. As shown in Table A.3, where systematically collected statistics that approximate the described behaviors are available, they cover only selected time periods (e.g. Human Rights Watch 2001), and some

estimates are not based on a systematic data construction methodology (e.g. Eisenfeld and Eisenfeld 1999).

Overall, our model is not suitable for top-down coding approaches of defection, as perceptions shape whether defection is reported in the first place.

A.III.2. Counterfactuals: Reward and Punishment

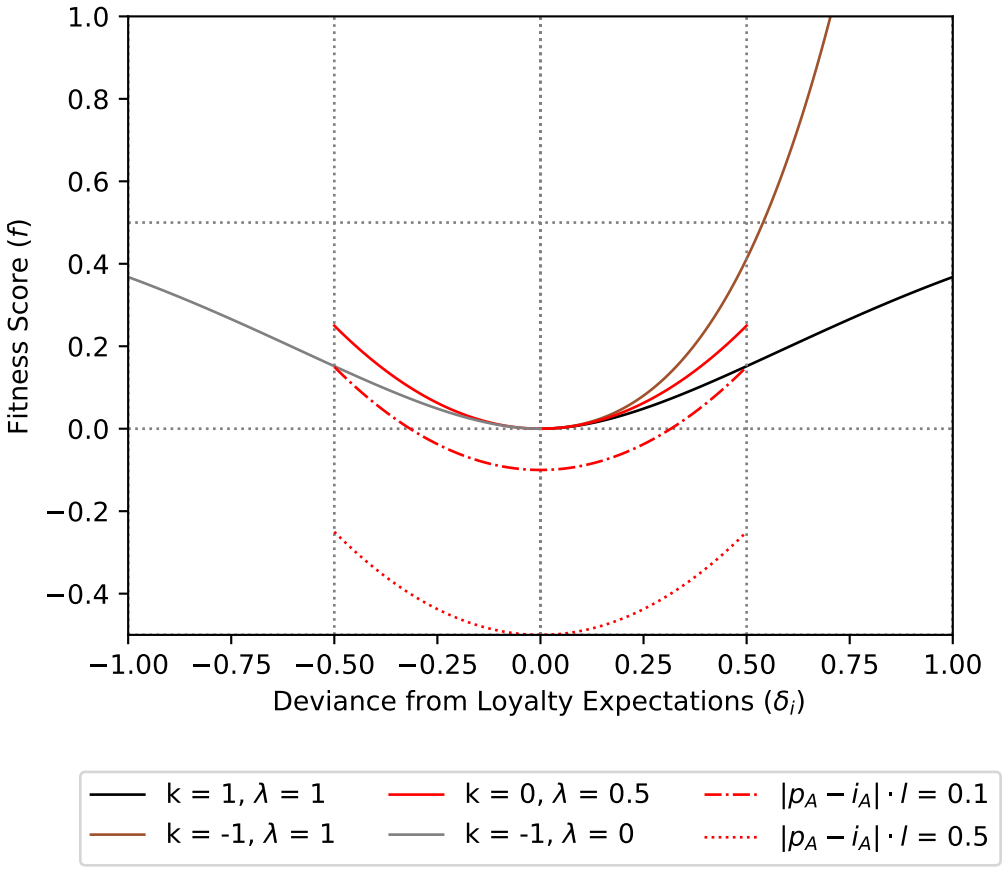
In this section, we consider how changes in reward and punishment (k) affect the direction of the observed allegiance shift in the GDR and the OPT. k represents exogenous social and material incentives for (dis)loyalty at the group level, whether by other members or authorities.

In the GDR amongst East Germans, it was generally accepted and rewarding to maintain Western contacts, and Western state and non-state organizations provided protection to political dissidents to counter GDR sanctioning for affiliation with Western organizations and the church. By comparison in the OPT during the Second Intifada, it was highly rewarding for Palestinians to distance themselves from Israeli institutions and individuals. Israeli authorities could not protect collaborators from the severe social punishments should their activities become known, and political actors in the West Bank were generally willing to grant amnesty if informing had not led to the death of Palestinians. Even where Israeli authorities commit to protecting Palestinian collaborators (e.g. via residence permits for Israel), the bureaucratic hurdles to such protection are significant, and extremely few would consider it rewarding (see Jalal 2015; Pulwer and Cohen 2016; Sherwood 2011).

Figure A.5 (A) and (B) show how defector type prevalence changes as a function of k in the GDR and the OPT, respectively. All else equal in the GDR, loyalty incentives do not alter the outcome shown in Section 5.2. That is, even where Western protection of dissidents faltered, or GDR state security punished outgroup contact severely, cascades towards defection eventually ensue. For the OPT however, defection increases when incentives for disloyalty outweigh those for disloyalty. Loyalty trials only generated conformity in the OPT where it was socially rewarding, such that Palestinians who refused collaboration were protected from reprisals.

A.IV. Supplementary Figures and Tables

Figure A.1. Agent Fitness and Loyalty



Note: Plot of Equation (1.11): $f_A = \frac{\delta_i^2}{e^{k\delta_i}} - |p_A - i_A| \cdot l_A$. Inspired by Dino Dini's Normalized Tunable Sigmoid Function.

Table A.1. Comparative Statistics for Model Parameters

Model	Parameter Change	A^I	A^{II}	A^{III}	A^{IV}	Allegiance Δ_λ
Baseline	—	479.87	184.84	113.94	221.35	8.65%
Min k	$k = -1$	88.18	228.11	55.42	628.29	-33.43%
Max k	$k = 1$	622.63	159.35	108.2	109.82	20.36%
Min $\bar{q}$	$\bar{q} = 0.025$	245.62	209.52	93.58	451.28	-15.31%
Max $\bar{q}$	$\bar{q} = 0.9$	332.24	232.04	93.66	342.06	-6.52%
Min σ_q	$\sigma_q = 0.025$	232.41	223.67	90.88	-16.83%	-16.83%
Max σ_q	$\sigma_q = 0.9$	328.98	224.61	108.19	338.22	-5.11%
Min σ_i, σ_p	$\sigma_i, \sigma_p = 0.025$	367.68	312.51	77.7	242.12	-6.21%
Max σ_i, σ_p	$\sigma_i, \sigma_p = 0.9$	332.24	232.04	93.66	342.06	9.51%
Min P	$P = 1$	434.83	242.85	111.21	211.11	5.55%
Max P	$P = 10$	230.87	220.55	85.32	463.26	-17.61%
Min T	$T = 1\%$	481.92	488.28	13.9	15.91	-0.001%
Max T	$T = 100\%$	250.85	212.11	93.01	444.02	-14.75%
Min M	$M = 0$	198.68	199.3	52.41	549.61	-23.61%
Max M	$M = 1$	275.04	271.93	113.74	339.29	-9.92%

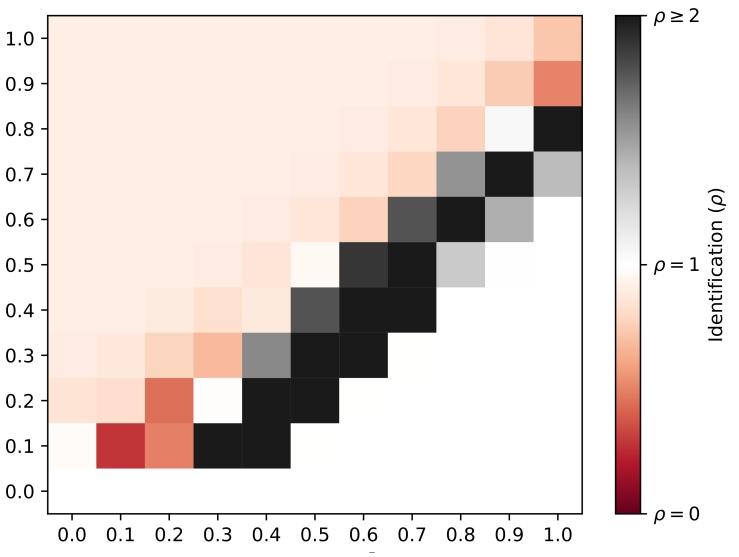
Note: The listed parameters are varied holding all other parameters constant at a baseline that makes agents indifferent between loyalty and disloyalty. Baseline parameters are shown in Table A.4, and results correspond to Experiment 61 in Table A.5.

Table A.2. Observable Data for GDR & OPT Settings

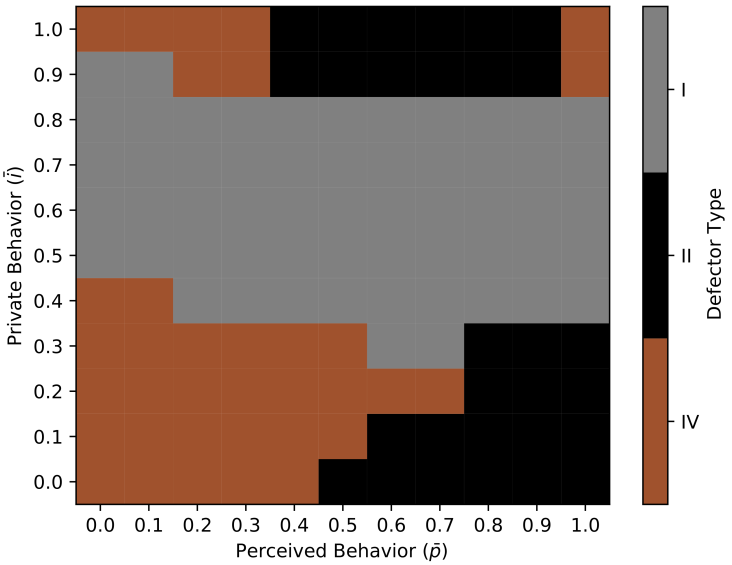
Loyalty Expectation	Observable Defection	GDR (1971-1989)	OPT (2000-2004)
Low $(0.1 - 0.3)$	Plan Revolution	0.0003%*†	N.A.
	Land Selling	—	0.00009%†
	Enemy-Informing	0.004%*†	0.02%†
Moderate $(0.4 - 0.6)$	Political Prisoners	1.55%†	0.01%†
	Protest Turnout	1.58%*	N.A.
	Refuse Authority Support	N.A.	N.A.
High $(0.7 - 0.9)$	(Attempt) Illegal Emigration	2.3%*†	—
	Work in Rival Area	—	3.7%*
	Personal Outgroup Contact	N.A.	N.A.
Sum		$\sim 5.41\%$	$\sim 3.84\%$

Note: Defector estimates include *the population at risk of being labeled and †officially labeled suspects, relative to total population size. To arrive at these estimates, count data from the sources listed in Table A.2 was summed and divided by the mean population across the time period. '—' denotes that behavior was not observed, 'N.A.' that no statistic was available. Estimates are unreliable and do not always cover the exact time period indicated. Sources are provided in Table A.3.

Figure A.2. Misidentification and Misperception



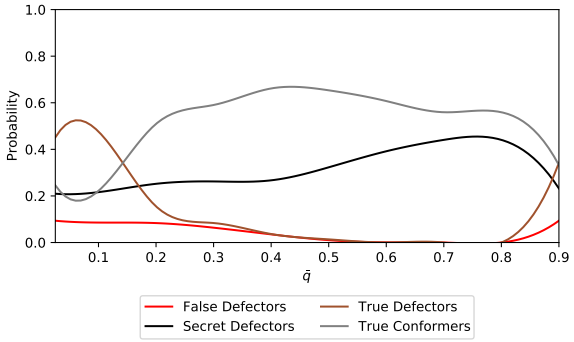
(A) Misidentification



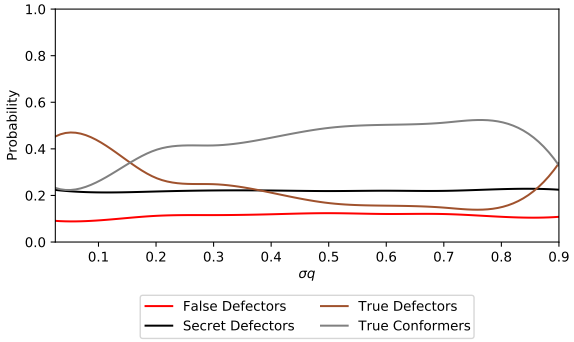
(B) Misperception

Note: All parameters are at their default values given in Table A.4. **(A)** Heatmap cells represent average misidentification of defectors (ρ) across generations and simulations, for the same experimental conditions as the general model. **(B)** Heatmap cells represent the most prevalent defector type in the last generation, summed across simulations, for possible combinations of private and perceived behavior.

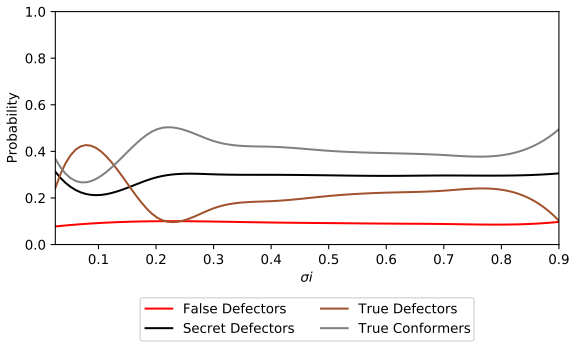
Figure A.3. Agent Parameters and Defector Type Prevalence



(A) Tolerance Mean



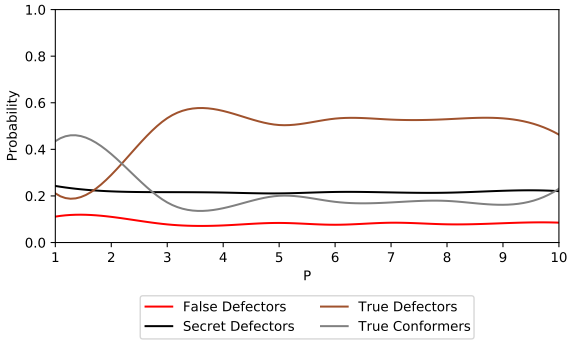
(B) Tolerance Spread



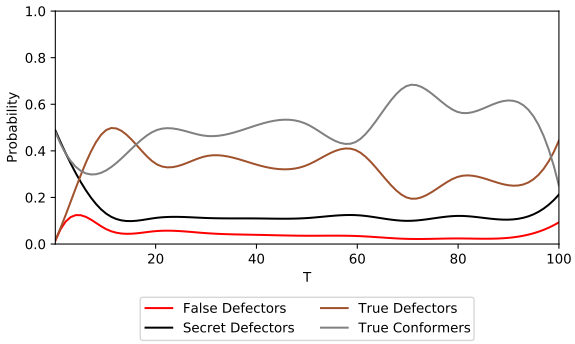
(C) Behavior Spread $\sigma_i = \sigma_p$

Note: Showing the probability of defector types by initial (A) mean tolerance ($\bar{q} \in [0.025, 0.9]$), (B) tolerance spread ($\sigma_q \in [0.025, 0.9]$), (C) spread of behavior and perceptions ($\sigma_i = \sigma_p \in [0.025, 0.9]$). Initially, $\lambda = \bar{i} = \bar{p} = 0.5$, all other parameters correspond to defaults in Table A.4. To arrive at a smooth probability distribution, the prevalence of each defector type was averaged across simulations, and values between experimental conditions were interpolated using cubic splines.

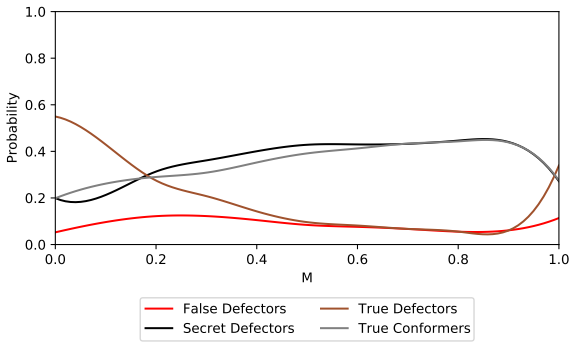
Figure A.4. Defector Type Prevalence by Simulation Parameter



(A) Agent Pairings



(B) % Agents Affected



(C) Mutation Probability

Note: Showing the probability of defector types by (A) agent pairings ($P \in [1, 10]$), (B) agents updated per generation ($T \in \{0.1\%, 1\%, 10\%, ...100\%\}$, with steps of 10%), (C) mutation probability ($M \in [0, 1]$). Initially, $\lambda = \bar{i} = \bar{p} = 0.5$, all other parameters correspond to defaults in Table A.4. To arrive at a smooth probability distribution, the prevalence of each defector type was averaged across simulations, and values between experimental conditions were interpolated using cubic splines.

Table A.3. Sources for Table A.2

Case	Statistic	Years	Sources
GDR	Population	1971-1989	Franzmann 2009
GDR	Illegal Emigration	1971-1988	Eisenfeld 1999
GDR	Spies Accused/Arrested	1971-1989	Maddrell 2013
GDR	Political Prisoners	1979-1982, 1984-1988	Horz and Marbach 2022
GDR	Selected Events Turnout*	1973-1983	Eisenfeld and Eisenfeld 1999
GDR	Protest Turnout 1989	1989	Lohmann 1994
GDR	Revolutionary (KPD/ML) Affiliation	1980	Eisenfeld and Eisenfeld 1999
OPT	Population	2000-2004	Palestinian Central Bureau of Statistics 1999, PCBS
OPT	Fatalities/Death Penalties	2000-2004	B'Tselem 2021a; 2021b
OPT	Work in Israel/Settlements	2000-2004	Palestinian Central Bureau of Statistics 2005
OPT	Land Sales*	2001	Human Rights Watch 2001
OPT	Security prisoners*	2000-2001	Human Rights Watch 2001

Note: Statistics marked with * were not systematically compiled, but merely mentioned figures in the cited texts.

Table A.5. General Model Results

	λ	$\bar{p} = \bar{i}$	A^I	A^{II}	A^{III}	A^{IV}	**Identification** ρ	**Allegiance** Δ_λ
Exp. 1	0.0	0.0	1000.0	0.0	0.0	0.0	1.0	250.83
Exp. 2	0.1	0.0	360.09	208.37	150.58	280.96	0.97	127.32
Exp. 3	0.2	0.0	1.71	153.93	2.75	841.61	0.85	-192.63
Exp. 4	0.3	0.0	0.09	108.08	0.14	891.7	0.89	-293.28
Exp. 5	0.4	0.0	0.0	97.89	0.0	902.11	0.9	-393.79
Exp. 6	0.5	0.0	0.0	95.89	0.0	904.11	0.9	-494.2
Exp. 7	0.6	0.0	0.0	95.32	0.0	904.68	0.9	-594.55
Exp. 8	0.7	0.0	0.0	95.17	0.0	904.83	0.9	-694.85
Exp. 9	0.8	0.0	0.0	95.13	0.0	904.87	0.9	-795.12
Exp. 10	0.9	0.0	0.0	95.13	0.0	904.87	0.9	-895.35
Exp. 11	1.0	0.0	0.0	95.13	0.0	904.87	0.9	-995.57
Exp. 12	0.0	0.1	1000.0	0.0	0.0	0.0	1.0	344.66
Exp. 13	0.1	0.1	849.24	112.93	28.49	9.33	0.28	447.99
Exp. 14	0.2	0.1	82.58	238.25	48.33	630.84	0.81	-120.04
Exp. 15	0.3	0.1	1.61	128.67	2.88	866.84	0.87	-279.37
Exp. 16	0.4	0.1	0.08	101.45	0.15	898.32	0.9	-381.39
Exp. 17	0.5	0.1	0.0	96.45	0.0	903.55	0.9	-482.88
Exp. 18	0.6	0.1	0.0	95.47	0.0	904.53	0.9	-584.08
Exp. 19	0.7	0.1	0.0	95.28	0.0	904.72	0.9	-685.07
Exp. 20	0.8	0.1	0.0	95.15	0.0	904.85	0.9	-785.92
Exp. 21	0.9	0.1	0.0	95.13	0.0	904.87	0.9	-886.68
Exp. 22	1.0	0.1	0.0	95.13	0.0	904.87	0.9	-987.38
Exp. 23	0.0	0.2	1000.0	0.0	0.0	0.0	1.0	443.21
Exp. 24	0.1	0.2	955.97	31.24	12.16	0.62	0.5	540.94
Exp. 25	0.2	0.2	794.42	136.55	48.85	20.19	0.45	434.58
Exp. 26	0.3	0.2	15.99	231.04	19.58	733.39	0.77	-251.3
Exp. 27	0.4	0.2	1.55	117.32	2.95	878.18	0.88	-359.3
Exp. 28	0.5	0.2	0.08	98.46	0.16	901.3	0.9	-463.84
Exp. 29	0.6	0.2	0.0	95.58	0.0	904.42	0.9	-567.07
Exp. 30	0.7	0.2	0.0	95.21	0.0	904.79	0.9	-669.58
Exp. 31	0.8	0.2	0.0	95.14	0.0	904.86	0.9	-771.67
Exp. 32	0.9	0.2	0.0	95.13	0.0	904.87	0.9	-873.48
Exp. 33	1.0	0.2	0.0	95.13	0.0	904.87	0.9	-975.06

Table A.5. General Model Results

	λ	$\bar{p} = \bar{i}$	A^I	A^{II}	A^{III}	A^{IV}	Identification ρ	Allegiance Δ_λ
Exp. 34	0.0	0.3	1000.0	0.0	0.0	0.0	1.0	543.05
Exp. 35	0.1	0.3	986.25	4.61	9.1	0.04	3.07	542.68
Exp. 36	0.2	0.3	944.77	31.71	22.58	0.95	0.99	524.04
Exp. 37	0.3	0.3	690.14	176.72	80.1	53.04	0.68	336.01
Exp. 38	0.4	0.3	13.88	180.06	20.19	785.87	0.83	-324.3
Exp. 39	0.5	0.3	1.5	109.62	3.0	885.87	0.89	-436.54
Exp. 40	0.6	0.3	0.08	97.22	0.16	902.54	0.9	-544.29
Exp. 41	0.7	0.3	0.0	95.36	0.0	904.64	0.9	-649.71
Exp. 42	0.8	0.3	0.0	95.17	0.0	904.83	0.9	-753.88
Exp. 43	0.9	0.3	0.0	95.13	0.0	904.87	0.9	-857.27
Exp. 44	1.0	0.3	0.0	95.13	0.0	904.87	0.9	-960.12
Exp. 45	0.0	0.4	1000.0	0.0	0.0	0.0	1.0	643.04
Exp. 46	0.1	0.4	997.83	0.24	1.92	0.0	2.79	556.25
Exp. 47	0.2	0.4	978.77	4.6	16.58	0.04	5.51	527.84
Exp. 48	0.3	0.4	931.75	31.59	35.24	1.43	1.59	477.27
Exp. 49	0.4	0.4	553.14	186.85	107.41	152.61	0.88	197.12
Exp. 50	0.5	0.4	13.33	153.98	20.68	812.0	0.85	-396.54
Exp. 51	0.6	0.4	1.5	106.84	3.04	888.63	0.89	-514.02
Exp. 52	0.7	0.4	0.08	96.87	0.16	902.89	0.9	-625.14
Exp. 53	0.8	0.4	0.0	95.33	0.0	904.67	0.9	-732.65
Exp. 54	0.9	0.4	0.0	95.17	0.0	904.83	0.9	-838.22
Exp. 55	1.0	0.4	0.0	95.13	0.0	904.87	0.9	-942.77
Exp. 56	0.0	0.5	1000.0	0.0	0.0	0.0	1.0	743.04
Exp. 57	0.1	0.5	1000.0	0.0	0.0	0.0	1.0	643.04
Exp. 58	0.2	0.5	997.15	0.24	2.61	0.0	3.44	551.83
Exp. 59	0.3	0.5	973.46	4.56	21.91	0.07	7.5	503.94
Exp. 60	0.4	0.5	927.94	31.43	39.14	1.49	1.77	425.83
Exp. 61	0.5	0.5	479.87	184.84	113.94	221.35	0.96	86.51
Exp. 62	0.6	0.5	13.29	149.57	20.95	816.19	0.86	-469.7
Exp. 63	0.7	0.5	1.51	105.92	3.05	889.52	0.9	-592.09
Exp. 64	0.8	0.5	0.08	96.68	0.16	903.09	0.9	-706.35
Exp. 65	0.9	0.5	0.0	95.34	0.0	904.66	0.9	-815.62

Table A.5. General Model Results

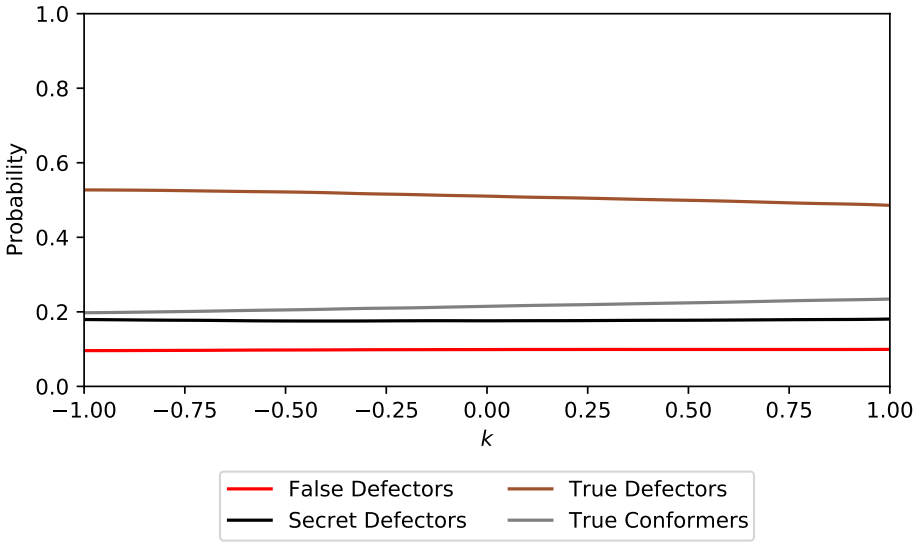
	λ	$\bar{p} = \bar{i}$	A^I	A^{II}	A^{III}	A^{IV}	**Identification** ρ	**Allegiance** Δ_λ
Exp. 66	1.0	0.5	0.0	95.17	0.0	904.83	0.9	-922.64
Exp. 67	0.0	0.6	1000.0	0.0	0.0	0.0	1.0	842.89
Exp. 68	0.1	0.6	1000.0	0.0	0.0	0.0	1.0	742.89
Exp. 69	0.2	0.6	1000.0	0.0	0.0	0.0	1.0	642.89
Exp. 70	0.3	0.6	994.24	0.24	5.52	0.0	5.95	548.47
Exp. 71	0.4	0.6	976.6	4.58	18.76	0.05	6.43	465.15
Exp. 72	0.5	0.6	925.32	31.06	41.85	1.77	1.89	378.53
Exp. 73	0.6	0.6	165.89	213.89	71.82	548.41	0.77	-333.15
Exp. 74	0.7	0.6	13.32	147.18	21.21	818.28	0.86	-543.82
Exp. 75	0.8	0.6	1.51	105.66	3.06	889.77	0.9	-671.05
Exp. 76	0.9	0.6	0.08	96.61	0.16	903.15	0.9	-787.83
Exp. 77	1.0	0.6	0.0	95.32	0.0	904.67	0.9	-898.77
Exp. 78	0.0	0.7	1000.0	0.0	0.0	0.0	1.0	923.17
Exp. 79	0.1	0.7	1000.0	0.0	0.0	0.0	1.0	823.17
Exp. 80	0.2	0.7	1000.0	0.0	0.0	0.0	1.0	723.17
Exp. 81	0.3	0.7	1000.0	0.0	0.0	0.0	1.0	623.17
Exp. 82	0.4	0.7	996.7	0.24	3.06	0.0	3.75	523.59
Exp. 83	0.5	0.7	985.91	4.6	9.46	0.03	3.46	424.08
Exp. 84	0.6	0.7	917.15	40.24	40.45	2.17	1.78	314.15
Exp. 85	0.7	0.7	156.07	206.38	71.65	565.89	0.78	-408.28
Exp. 86	0.8	0.7	13.13	145.8	21.15	819.91	0.86	-619.09
Exp. 87	0.9	0.7	1.5	105.49	3.07	889.93	0.9	-750.18
Exp. 88	1.0	0.7	0.08	96.55	0.16	903.22	0.9	-869.56
Exp. 89	0.0	0.8	1000.0	0.0	0.0	0.0	1.0	956.05
Exp. 90	0.1	0.8	1000.0	0.0	0.0	0.0	1.0	856.05
Exp. 91	0.2	0.8	1000.0	0.0	0.0	0.0	1.0	756.05
Exp. 92	0.3	0.8	1000.0	0.0	0.0	0.0	1.0	656.05
Exp. 93	0.4	0.8	999.99	0.0	0.0	0.0	1.0	556.05
Exp. 94	0.5	0.8	999.29	0.24	0.47	0.0	1.31	456.05
Exp. 95	0.6	0.8	989.16	4.61	6.21	0.02	2.17	356.08
Exp. 96	0.7	0.8	872.69	56.88	38.68	31.76	1.55	199.44
Exp. 97	0.8	0.8	124.59	217.69	65.58	592.14	0.77	-504.84

Table A.5. General Model Results

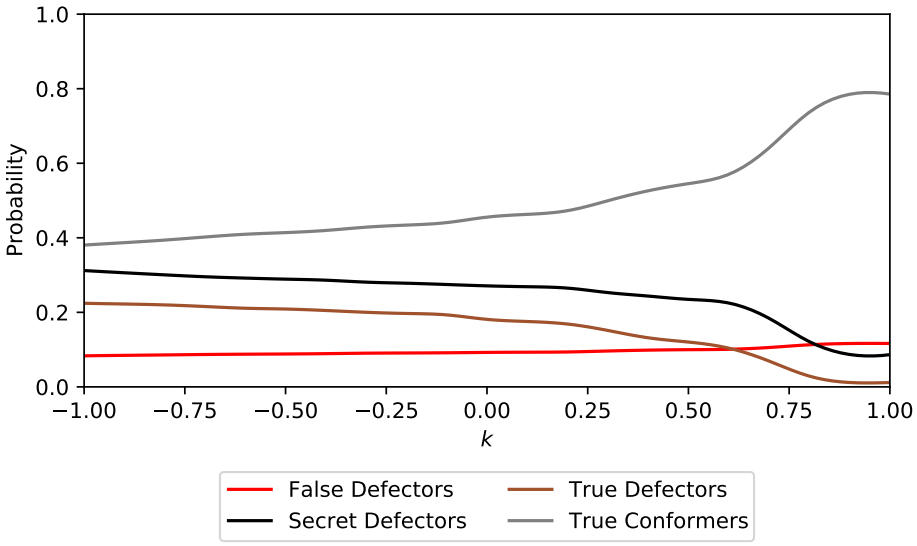
	λ	$\bar{p} = \bar{i}$	A^I	A^{II}	A^{III}	A^{IV}	Identification ρ	Allegiance Δ_λ
Exp. 98	0.9	0.8	12.58	147.09	20.88	819.46	0.86	-694.35
Exp. 99	1.0	0.8	1.49	106.06	3.05	889.4	0.9	-829.64
Exp. 100	0.0	0.9	1000.0	0.0	0.0	0.0	1.0	978.07
Exp. 101	0.1	0.9	1000.0	0.0	0.0	0.0	1.0	878.07
Exp. 102	0.2	0.9	1000.0	0.0	0.0	0.0	1.0	778.07
Exp. 103	0.3	0.9	1000.0	0.0	0.0	0.0	1.0	678.07
Exp. 104	0.4	0.9	1000.0	0.0	0.0	0.0	1.0	578.07
Exp. 105	0.5	0.9	999.99	0.0	0.01	0.0	1.01	478.07
Exp. 106	0.6	0.9	999.14	0.24	0.61	0.0	1.44	378.07
Exp. 107	0.7	0.9	988.87	4.61	6.5	0.02	2.21	278.11
Exp. 108	0.8	0.9	577.19	206.91	45.12	170.77	1.05	-114.6
Exp. 109	0.9	0.9	76.11	244.53	54.16	625.19	0.75	-607.21
Exp. 110	1.0	0.9	12.1	153.66	20.59	813.65	0.86	-770.05
Exp. 111	0.0	1.0	1000.0	0.0	0.0	0.0	1.0	992.21
Exp. 112	0.1	1.0	1000.0	0.0	0.0	0.0	1.0	892.21
Exp. 113	0.2	1.0	1000.0	0.0	0.0	0.0	1.0	792.21
Exp. 114	0.3	1.0	1000.0	0.0	0.0	0.0	1.0	692.21
Exp. 115	0.4	1.0	1000.0	0.0	0.0	0.0	1.0	592.21
Exp. 116	0.5	1.0	1000.0	0.0	0.0	0.0	1.0	492.21
Exp. 117	0.6	1.0	999.99	0.0	0.01	0.0	1.01	392.21
Exp. 118	0.7	1.0	999.21	0.24	0.55	0.0	1.38	292.22
Exp. 119	0.8	1.0	987.82	5.83	6.31	0.04	2.08	191.6
Exp. 120	0.9	1.0	202.0	372.67	47.95	377.37	0.51	-470.26
Exp. 121	1.0	1.0	60.77	273.89	48.23	617.11	0.72	-686.35

Note: General model results with default parameters as defined in Table A.6. Each experimental condition is simulated $S = 30$ times, and each simulation lasts $G = 100$ generations. Simulations are seeded with $N = 1000$ agents, $P = 3$ agent pairings per generation, $M = 0.1$ probability of agent mutation, $T = 10\%$ of agents updating per generation, $k = \lambda$ loyalty incentives, $\sigma_i, \sigma_q = 0.1$ dispersion of agent parameters, and $\bar{q} = 0.1$ initial agent tolerance. Agent parameter values are drawn from the normal distribution.

Figure A.5. Loyalty Incentives and Defector Type Prevalence



(A) GDR



(B) OPT

Note: Parameters other than k as per Table A.6. To arrive at a smooth probability distribution, the prevalence of each defector type was averaged across simulations, and values between experimental conditions were interpolated using cubic splines.

Table A.4. Model Parameters and Outcomes

Parameter	**Description**	**Values**	**Default**
Agent-Level			
i_A	A's private behavior	Variable $\in [0,1]$	$\mathcal{N}(\bar{i}, \sigma_i)$
p_A	A's perceived allegiance	Variable $\in [0,1]$	$\mathcal{N}(\bar{p}, \sigma_p)$
q_A	A's tolerance for disloyalty	Variable $\in [0,1]$	$\mathcal{N}(\bar{q}, \sigma_q)$
l_A, d_A	A was labeled, is defecting	Variable $\in \{0,1\}$	$(0,0)$
Group-Level			
$\bar{i}, \bar{p}$	Mean allegiance, perceptions	Variable $\in [0,1]$	Sweep $[0,1]$
$\bar{q}$	Mean tolerance	Variable $\in [0,1]$	0.1
$\sigma_{i,p}, \sigma_q$	Spread in allegiance, tolerance	Variable $\in [0,1]$	0.1
λ	Loyalty expectations	Fix $\in [0,1]$	Sweep $[0,1]$
k	Reward & Punishment	Fix $\in [-1,1]$	λ
Simulation			
N	Number of agents	Fix $\mathbb{Z}^+$	1000
S	Simulations per experiment	Fix $\mathbb{Z}^+$	30
G	Generations per simulation	Fix $\mathbb{Z}^+$	100
P	Agent pairings per generation	Fix $\in [1,10]$	3
T	Agents updated per generation	Fix $\in [0.1\%, 100\%]$	10%
M	Probability to mutate	Fix $\in [0,1]$	0.1
Outcomes			
A^{I-IV}	Defector type	Variable $\in \{I, II, III, IV\}$	—
Δ_λ	Group conformity	Variable $\in [0,N]$	—
$\bar{\sigma}_i$	Cohesion	Variable $\in [0,1]$	—
ρ	Identification ratio	Variable $\in [0,N]$	—

Note: 'Default' shows the initial parameter values used in the baseline (see Experiment 61 in Table A.5). 'Variable' parameter values change over the course of a model run (only the initial value is fixed).

Table A.6. Parameters for Empirical Contextualization

	Group Members	$\boldsymbol{\lambda}$	$\bar{\boldsymbol{i}}$	$\bar{\boldsymbol{p}}$	$\boldsymbol{\sigma_i}$	$\boldsymbol{\sigma_p}$	$\bar{\boldsymbol{q}}$	$\boldsymbol{\sigma_q}$	$\boldsymbol{P}$	$\boldsymbol{T}$	$\boldsymbol{G}$
GDR (1971)											
	0-100%	0.7			0.05	0.1	0	0.01	1	0.5%	1800
	0-95%		0.8	$i_A - 0.1$							
	95-99%		0.5	$i_A - 0.1$							
	99%-100%		0.2	$i_A - 0.1$							
OPT (2000)											
	0-100%	0.6			0.1	0.2	0.1	0.05	3	1%	400
	0-59%		0.7	$i_A + 0.1$							
	59-79%		0.5	$i_A + 0.5$							
	79-99%		1.0	$i_A - 0.5$							
	99-100%		0.05	$i_A = p_A$							
Both Settings											
	0-100%	Sweep $k \in [-1, 1]$									
	0-100%	$M = 0.1$									
	—	$N = 1000$									
	—	$S = 30$									

Note: Agent-level parameters are drawn from the normal distribution. Most parameters are applied across agents in each empirical setting. For behavior and perception parameters i and p, we draw different parameters for a given share of the group. To capture the notion that representative agents are individually perceived as more or less loyal than they behave, perceptions are drawn at the agent-level: $\mathcal{N}(\mu = \bar{p} = i_A \pm \epsilon, \sigma = \sigma_p)$. Mutation is based on the parameter values that the majority of the group is initially seeded with.

Table A.7. Extension Results

	k	A^I	A^{II}	A^{III}	A^{IV}	**Identification ρ**	**Allegiance Δ_λ**
GDR							
Exp. 1	-1.0	197.53	179.36	95.94	527.17	0.87	-425.81
Exp. 2	-0.9	198.69	178.52	96.11	526.68	0.88	-424.74
Exp. 3	-0.8	200.12	177.66	96.47	525.75	0.88	-423.39
Exp. 4	-0.7	201.72	177.17	96.78	524.33	0.88	-421.94
Exp. 5	-0.6	203.67	176.11	97.35	522.87	0.88	-419.92
Exp. 6	-0.5	205.17	175.53	97.64	521.66	0.88	-418.48
Exp. 7	-0.4	206.98	175.36	97.91	519.75	0.89	-416.93
Exp. 8	-0.3	209.26	175.53	98.33	516.88	0.89	-415.04
Exp. 9	-0.2	210.5	176.03	98.5	514.96	0.89	-413.9
Exp. 10	-0.1	212.68	176.19	98.68	512.44	0.89	-412.13
Exp. 11	0.0	214.8	175.98	98.78	510.44	0.89	-410.6
Exp. 12	0.1	216.99	176.25	99.03	507.73	0.89	-408.9
Exp. 13	0.2	218.59	176.36	99.02	506.02	0.89	-407.89
Exp. 14	0.3	220.34	176.81	99.14	503.7	0.9	-406.52
Exp. 15	0.4	222.35	177.31	99.06	501.28	0.9	-405.25
Exp. 16	0.5	224.14	177.5	99.09	499.26	0.9	-404.07
Exp. 17	0.6	226.05	177.98	99.03	496.93	0.9	-402.81
Exp. 18	0.7	228.42	178.62	99.01	493.95	0.9	-401.18
Exp. 19	0.8	230.65	179.19	99.03	491.14	0.9	-399.39
Exp. 20	0.9	232.28	179.56	99.06	489.09	0.9	-398.18
Exp. 21	1.0	234.34	180.65	99.32	485.69	0.9	-396.47
OPT							
Exp. 1	-1.0	380.41	312.07	83.25	224.26	0.53	-133.45
Exp. 2	-0.9	386.96	306.14	84.59	222.32	0.54	-127.68
Exp. 3	-0.8	393.84	300.54	85.71	219.91	0.55	-121.88
Exp. 4	-0.7	402.07	295.45	86.87	215.61	0.56	-114.59
Exp. 5	-0.6	409.48	291.87	87.67	210.98	0.57	-108.06
Exp. 6	-0.5	413.89	289.04	88.23	208.84	0.57	-104.21
Exp. 7	-0.4	419.68	286.25	89.14	204.93	0.58	-98.56
Exp. 8	-0.3	428.51	280.99	90.38	200.13	0.59	-90.41

Table A.7. Extension Results

	k	A^I	A^{II}	A^{III}	A^{IV}	**Identification ρ**	**Allegiance Δ_λ**
Exp. 9	-0.2	434.03	278.12	90.93	196.92	0.59	-85.25
Exp. 10	-0.1	440.72	274.65	91.49	193.14	0.6	-79.07
Exp. 11	0.0	455.44	271.01	92.46	181.09	0.61	-65.17
Exp. 12	0.1	462.84	268.73	93.02	175.41	0.62	-57.88
Exp. 13	0.2	472.75	264.99	93.67	168.59	0.62	-48.33
Exp. 14	0.3	499.42	252.96	96.25	151.37	0.67	-20.29
Exp. 15	0.4	526.09	243.66	98.57	131.68	0.7	7.16
Exp. 16	0.5	545.26	234.74	99.77	120.24	0.74	26.39
Exp. 17	0.6	570.15	225.04	100.88	103.93	0.78	51.36
Exp. 18	0.7	641.09	184.01	105.59	69.31	1.03	125.8
Exp. 19	0.8	736.87	121.37	112.91	28.85	1.66	228.37
Exp. 20	0.9	785.03	86.89	116.36	11.73	2.14	280.04
Exp. 21	1.0	785.54	86.42	116.44	11.6	2.16	280.43

Note: Parameter sweep of k for Appendix A.III.2, holding other parameters constant as shown in Table A.6.

Table A.8. List of Interviews

Date	**Location**
03.11.2019	Ramallah
04.11.2019	Jerusalem
06.11.2019	Tel Aviv
12.11.2019	Ramallah
12.11.2019	Ramallah
13.11.2019	Birzeit
20.11.2019	Givat Haviva
26.11.2019	Jerusalem
26.11.2019	Ramallah
28.11.2019	Jerusalem
04.12.2019	U.S. (VoIP)
17.02.2022	Birzeit
20.02.2022	Ramallah
22.02.2022	Birzeit
22.02.2022	Ramallah
23.02.2022	Ramallah
25.02.2022	Ramallah
27.02.2022	Jerusalem
21.03.2022	Jerusalem (VoIP)
05.04.2022	Ramallah (VoIP)

Note: Due to the sensitivity of the subject, all research partners are kept anonymous. Not listed are informal conversations with Palestinians and Israelis about the subject.

PAPER 2 APPENDICES

B.I. Data Construction

There are two major challenges to data construction from the 'Stasi-Archives': the corpus of information is too vast for comprehensive examination, and access to it is limited due to the sensitivity of its contents. A major part of the files that document MfS activities was secured after the collapse of the regime, including 15,500 shredded paper bags, out of which around 500 have been reconstructed thus far. Until recently, a dedicated institution, the *Office of the Federal Commissioner for the Records of the State Security Service of the former German Democratic Republic* (BStU), was solely responsible for documenting, maintaining, and providing access to these files. Over 111 kilometers of records are distributed across 13 locations in Eastern Germany, about 44km of which stem from ministry headquarters in Berlin (20km of those had been archived by the *Stasi*).[1]

Due to the sensitive nature of the information the files contain, and the human rights violations committed to obtain it, researchers must request special permission to view uncensored files, and can only do so in reading rooms at BStU headquarters in Berlin after signing a confidentiality agreement. After reviewing the files, censored copies of the originals may be requested. Time to review files in the reading rooms was limited, and the use of analytical software on original documents prohibited. Therefore, data construction took place in four steps:

1. Sampling Stage 1: query archivist file descriptions from the archive's database
2. Sampling Stage 2: purposeful sample of files to maximize completeness of surveillance and labels
3. Sampling Stage 3: purposeful sample of file sections to request censoring of relevant pages
4. Coding: systematic coding of censored copies and construction of the dataset.

I provide a description of each step below.

B.I.1. Sampling Stage 1: Database Query

The Stasi conducted its operations across wide parts of society, from opposition activists over state-owned enterprise employees to its own case workers. The challenge is to select a reasonable sample of relevant files, without omitting particular social groups, indicators for loyalty, forms of labeling, or behavioral responses to labeling. I considered two general approaches: subjective (purposeful) sampling and statistical sampling (Kepley 1984).

Statistical sampling may be based on a random sampling strategy (e.g., selecting every n-th file), however, random sampling is not technically support by the BStU at the time of writing. Even if it were this would yield very little evidence on labeling or loyalty, seeing as a single file may contain a host of documents that give no indication of either, such as lists of persons of interest, surveillance of organizational operations rather than individuals, and MfS-internal directives. Stratifying the random

[1] See https://www.bstu.de/ueber-uns/bstu-in-zahlen/.

sampling process would have required a way to group files into categories that are meaningful for the research topic. Beyond a specific signature, the fields that one can use to filter queries include:

1. Person research (using name and date of birth)
2. MfS department
3. Location of file in BStU system
4. Duration of procedure ['Laufzeit']
5. Full text (description of file by archivist)

For some research questions, filtering the corpus on these criteria may reduce the number of files sufficiently, while retaining an unbiased sample of documents. A study on the surveillance of protestant church members in the GDR, for instance, might weight by MfS department to oversample files from the *Hauptabteilung (HA) XX*, which dealt with political surveillance and repression of state institutions, cultural organizations, media and opposition movements. But this strategy would under-represent other sections that presumably hold relatively few files on church members, such as the *HA II* for domestic espionage, possibly violating not only the first but also the second criterion for an optimal sampling strategy. Overall, a statistical sampling strategy is not currently feasible at the Stasi archives, and not appropriate for this project. Instead, I use a purposeful sample that systematically identifies as many documents material for this project as possible.

I limited the scope of the study to *written files concerning the surveillance of individual GDR citizens for suspicion of disloyalty*, as different from surveillance of organizations, or individuals who were surveilled for other reasons, such as ensuring that they are sufficiently loyal to travel abroad. The relevance of a file is further determined by its completeness: a file should at least contain the indicators for disloyalty that arose the suspicion of the Stasi, and ideally describe the entire surveillance process until its conclusion. Where a file indicates that an individual was labeled by a Stasi official, the file would ideally describe the surveillance of the individual after the labeling occurred as well.

Based on these criteria, I designed a 'full text' search query to compile a list of documents from the entire corpus of the archive that is accessible. I therefore construct a search query with two main components to compile an initial list of files. The first component selects all files marked as 'archived'. This means that the list contains all cases that the *Stasi* regarded as closed, thus increasing the chances that files detail the behavior of individuals until their loyalty or exit from the GDR is observed. The second component searches for *keywords in the file description* which indicate that an individual has been labeled. Because the process by which archivists described the contents of a file is not documented and unlikely uniform, the pilot study in 2018 served to identify keywords that relevant files were likely described with. The overall search query in pseudocode:

```
Archivierter Vorgang ODER
        Operativer Vorgang UND
        Abgeschlossener Vorgang UND
        Etikettierung:
                Aufklaerung ODER
                Inhaftierung ODER
                Zufuehrung ODER
                Vernehmung
```

```
[English Translation]
Archived Procedure OR
        Operative Procedure AND
        Concluded Procedure AND
        Labelling:
                Clarification OR
                Incarceration OR
                Police Escort OR
                Interrogation
```

Note that the search only included files stored at MfS headquarters in Berlin. One could in principle repeat the process for each regional office as well, thus reducing the bias towards cases in Berlin, though the workload for all parties involved would be quite high given the large number of files that the process ultimately yields.

B.I.2. Sampling Stage 2: Files for Review

Figure B.1 shows an overview of the Stage 2 file selection criteria. I classified all files based on its description, and assigned a code for whether a file was deemed relevant (within the scope of the project), i.e. mentioned either disloyal behavior ('indicators') or contact with state security agents ('labeling'). Borderline cases that referenced criminal deviance, such as "breach of exchange control regulations (selling of gold- and silver ingots to VEB Münze)", were highlighted for a second pass of review, but ultimately dropped due to limitations in the number of files that can feasibly be requested by a single researcher (see Figure B.2 for an overview of reasons for exclusion). This decision-tree is applied twice: once based on the description of the file in the database to decide whether it warrants an access request, and once upon reviewing the uncensored file at BStU headquarters. This process led to a reduction of the 1248 files in the query-based sample to 453 files that were requested for review.

As shown in Figure B.3, one third of files selected for the preliminary sample originate from MfS department *HA XVIII*, which was responsible for the surveillance of firms, with a focus on foreign trade, science and technology, as well as the defense industry. Other frequently sampled departments include *HA II*, *HA IX*, and *HA XX* (each around 15%). These were respectively tasked with counter-intelligence (including within the MfS), criminal prosecutions, and the surveillance of opposition in culture, media and churches. Another 5% of files stems from the 'Confidential Repository', which concern procedures that the MfS considered particularly sensitive, such as disloyalty of Stasi officials. Most files from sections *HA XIX* (transport and post surveillance) and *ZKG* (coordination group against emigration attempts) were included, with many descriptions pointing to operative procedures concerning border-crossings.

B.I.3. Sampling Stage 3: Sections of Files

With access to the uncensored files at the reading room but insufficient resources to analyze them, the goal is to identify sections that may be relevant for the purposes of this project, to request censored copies. The density of material on a single individual ranges from a one-page report to multiple volumes with multiple reports on stakeouts, 'subversive' meetings, denunciations, interrogations, and lists of contacts.

I ensured consistency with the above Stage-2 sampling criteria, and created a database of individual 'surveillance windows'. A window is a sequence of 'episodes' during which the behavior of a subject or subjects is observed for the purposes of identifying their political loyalty. Each episode is a document where an MfS case worker gives an account of their activities. A single document can at maximum

be coded as a single episode, even when it includes accounts by multiple individuals (for example a report may include denunciations by informants as attachments), or the described activities span several months (for example, quarterly reports may repeatedly synthesize intelligence from operations that date back several years).

Surveillance windows may mention multiple individuals, but only windows for those individuals for whom the surveillance operation was opened and is concluded are considered. I neither coded windows that target organizations or discuss aggregate information on networks of individuals, nor 'open' windows that were not concluded (see above). However, I coded separate windows for multiple individuals who were surveilled on suspicion of disloyalty and whose cases were closed. I also included windows that only consisted of closure reports, *if* the closure report described the initial suspicion for disloyalty. Otherwise, the file was coded as incomplete, seeing as it was possible for case workers to write 'operative person controls' without actually conducting any surveillance beyond the material that was already available on the suspect.

For each set of files on a single surveillance window, the following types of sections or 'episodes' are identified for detailed coding:

1. **Initial opening report or analysis**: Must at least contain reasons for opening of investigation. May include denunciations, and metadata about the subject.
2. **Initial contact with label**: First episode where the subject is categorized negatively for an activity, including official labeling through preventive interrogations, as well as unofficial labeling.
3. **First official contact with label**: first episode where the subject is officially labeled by authorities (only applies if initial contact was not official).
4. **Latest background report or interrogation**: accounts of subject activities *prior* to the surveillance (latest background interrogation takes precedence over report about interrogation).
5. **Most comprehensive analytical report**: accounts of subject activities *during* surveillance (longest by page number takes precedence over latest report).
6. **Closure report**: must contain conclusions about the subjects' allegiance or activities *after* an investigation and possibly labeling.

These selection rules are based on experiences from the pilot study in 2018, as well as experiences with the material in the first week of the primary study in 2020, which prompted a re-evaluation of the sampling process to further reduce the material. The review of the original files was completed after around 2.5 months at the archives in the fall of 2021. The censoring of files that were used for systematic coding was completed by a case worker at the archives in May 2022.

Finally over the course of a year, archivists were asked to sequentially search for additional files on the individuals that were selected for systematic coding. As requesting the files themselves would have taken several more years and put too much workload on case workers for potentially no gain (naively requested files would likely have duplicated information from already reviewed files), the search results were systematically coded for information on additional surveillance windows instead. This included information on the first and last date of surveillance mentioned, criminal prosecutions (if any), and brief MfS notes regarding surveillance activities. A preliminary analysis of this data suggests that 5% of sampled individuals were disloyal prior to the observed window; 11% were (still) disloyal after the observed window (4% of those seemingly left the GDR); and 2% may have been recruited as informants

for the *Stasi* prior or after the surveillance. For case studies, this data complements the information in the sampled files. And once the systematic coding procedure is complete, it can be used to estimate how much information on sampled individuals is missing from the data.

B.I.4. Coding

The 6308 censored pages are coded manually using MaxQDA (VERBI Software 2020), and involves three passes of each "surveillance window". At the time of writing, coding is completed for 10% of cases. The first pass serves to identify metadata (episode ID, dates, MfS author etc.), as well as the authorities, subjects and group members associated with each paragraph in each account. This includes surveilled subjects, as well as their relationship to third parties where such relationships are made explicit, but not relationships between third parties that are not the primary subject of surveillance.

The second pass codes (1) the accounts subjects give about their private loyalty, (2) accounts by group members about the subject's private and perceived loyalty (including labeling), and (3) accounts that authorities give of their own perceptions and traitor profile, as well as their activities, including labeling. An account is coded as a subject account if the subject can be presumed to be the original narrator of the account, and it is not narrated through any other individual (subject or object), except for an MfS employee. These accounts are typically found in interrogation protocols, reports that paraphrase such interrogations, and letters or other documents written by subjects. Subcodes can be interpreted as indicators for the labeling and private loyalty of the subject. By the same token, third party accounts are presumed when the MfS is attaching copies of letters, informant reports or quoting them directly. Of note, authorities give accounts of labeling by third parties, and at times of subject accounts of labeling. When in doubt, accounts are coded as authority accounts.

For subject accounts, main codes relate to how they present their private behavior:

- present actions as compliant or innocent: "I attempted, through active professional and societal work [...] to contribute to the solution of concrete tasks and contradictions in the Sociallist development of the GDR. Because of my work, which was indeed in the spirit of the [SED] party program, In was awarded a [ ] medal"
- state uncertainty or no recollection: "I do not recall that we organized any circle activities in Leipzip"
- admit non-compliance: "In the last five years I belonged to a conspirational circle with members of the SED, whose activities was aimed at changing the power relations of the GDR [...] For my membership and activity in the circle I am solely responsible and am willing to accept the consequences"

Analogous codes are used for group members reporting on the loyalty or disloyalty of subjects. In addition, third parties have the following codes related to the labeling of subjects:

- supporting: "The father of the accused [ ] reassured them, that he will take the necessary steps in the FRG [to arrange for their exit]
- rejecting or denying support: "They refused to receive the statement of solidarity unofficially, but demanded an official statement to maintain their 'legal position' "
- bedeviling [if the third party is making negative remarks about the subject for an activity or about the activity itself, but not denouncing the activity to authorities in the coded segment]:

"His parents are against his behavior but he is not letting them meddle"; "The IM was appalled and distanced himself officially from him, telling him that he is a hypocrite who for years deceived his comrades about his positive party-attitude. Due to his hostile attitude against our state he also took a hostile attitude against him"

- denouncing [if the third party is reporting the subject's activity to authorities]: "The IM [ ] reported via phone to the undersigned and reported, that [ ] revealed his hostile activities to him in a conversation. At the request of the IM a short-term meeting was arranged."

I code the accounts of the MfS for their relation to perceived and imagined disloyalty, along the following dimensions:

- Threat appearance (including loyalty conflicts): "In August of of the year 1973 the father of the accused went ill. She received permission for a visit of multiple weeks into the FRG. This encouraged the accused in her decision, to leave the GDR together with her family illegally."
- Motive (political, obscure, personal): "Due to her inimical-hateful attitude against the GDR, she subordinated herself under the systematic espionage activities of her husband."; "the possible influences and motives for such behaviors and statements of [ ] could not be determined"; "As reasons [for leaving the oppositional group] Mrs. [ ] stated that it was difficult to get all the people involved together, particularly since the wife of [ ] was working as a teacher in the countryside and could not participate
- Activity or predicate (negative, obscure, positive): "Further enemy literature passed to IM"; "Shortly before the implementation of the visa restrictions with the VR Polen the subject stayed for a short time in Poland at the office of the new unions. Details about this are unknown."; "The [illegal] literature distributed by the group was given to the MfS voluntarily as part of the interrogation"
- Background (negative, non-political, positive): "His political development proceeded coherently over his active participation in the pioneer organization, the FDJ until his decision to become a member of the SED"; "it was emphasized [in his school] that decisive political statements during debates were lacking. [ ] showed reservations in taking responsibilities within the FDJ, because he spent his free time with personal interests."; "her husband is critically influenced by his parents, seeing as his father left the SED"

Finally, I code how authorities justify their own activities, with meta-codes including the judgement, labeling, and surveillance of subjects, as well as organizational communication strategies. For example:

- judge subject—highlight magnitude of threat potential: "it stands to reason that [ ] and [ ] are two members of an already existing, larger group"
- judge subject—defend subject account: "it does not seem useful to implement criminal measures. Of note, the mentioned individuals were open and objective during the interrogations, and there are starting points for a promising re-education."
- labeling—restrict mobility—deny travel: "due to the activities of the circle around [ ] regarding the events in the VR Poland, a travel ban for [ ] and [ ] was implemented"
- labeling—control without reprimand—unofficial refusal to publish: "According to her own statements, a crucial role played that her 'critical questions' were previously rejected"

- labeling—criticize publicly: "unofficially it became known that [ ] received a disciplinary and party reprimand due to revisionist tendencies"
- labeling—talk things out: "On 27.6.89 a preventive talk was conducted with [ ]. The measure was taken on the basis of a verbal decision with the head of the UA in conclusion of the confirmation by the OEI from 26.6.89. The head of division, Mj. Büchner, participated in the preventive talk"
- organizational communication—defer to higher authority: "Based on the conception confirmed by the assistant minister—Major General Mittig—for the prevention of further enemy activities [...] interrogations were implemented in cooperation with the HA IX"
- surveillance—investigate area of residence: "Investigations in the neighborhood in May 1976 about the couple showed that they have a good repute"

Other codes that were assigned 'inductively' relate to 'oppositional norms' (where loyalty to the GDR is labeled as negative), subject or third-party awareness of surveillance, opportunities to demonstrate loyalty, .

Finally, the third pass re-classifies the existing sets of codes into categories that will be used for statistical analysis. First, for all "negative' codes in subject, third party and authority accounts, behavior is classified in terms of the deviance itself (e.g. "maintain West contacts", "watch Western television", "obtain illegal literature"), and constructions of disloyalty (sabotage, enemy-informing, voice, exit, outgroup contact, other quotidian non-compliance). Second, key sets of codes are arranged visually, by 'actor' (subject, group member, authority) on the y-axis and time on the x-axis. Relevant sets include all codes related to labeling, subject allegiance (loyalty, disloyalty), and object allegiance. See Figure B.5 for the 'Deviant Defector' example. This final step significantly reduces the density of information coded for each case, but has the advantage that it will eventually allow for a statistical estimation of the relationship between labeling and political deviance.

1. Inductively code MfS case worker accounts of their activities and the contents of their reports, on the one hand, and the impact that labeling has on the subject, on the other hand (Garfinkel 1967). This data is used to explain how individuals are constructed as loyal or disloyal, and how these constructions may be contested by the surveilled subjects.
2. Deductively code descriptions of individual activities into types of allegiance, as well as bedeviling by their peers and official labeling by authorities.

B.II. Supplementary Figures and Tables

Table B.1. List of Archival Sources

Signature	Accessed
MfS AP 14791/72	10.02.2020
MfS GH 107/80	10.02.2020
MfS HA I 14995	11.02.2020
MfS HA I 20026	12.02.2020
MfS HA IX 25845	12.02.2020
MfS HA IX/11 AK 1242/83	12.02.2020
MfS HA XVIII 16238	12.02.2020

Table B.1. List of Archival Sources

Signature	Accessed
MfS HA XVIII 38416	12.02.2020
MfS HA XVIII 39520	13.02.2020
MfS HA XVIII 39521	13.02.2020
MfS HA XVIII 24974	14.02.2020
MfS HA XVIII 25116	14.02.2020
MfS HA XX 1471	14.02.2020
MfS HA XXII 5312	14.02.2020
MfS HA XVIII 28852	17.02.2020
MfS HA XVIII 30737	17.02.2020
MfS HA XVIII 30745	17.02.2020
MfS AOP 16183/81	18.02.2020
MfS AU 5812/77	20.02.2020
MfS GH 14/65	20.02.2020
MfS GH 16/77	20.02.2020
MfS GH 2/88	20.02.2020
MfS GH 21/86	20.02.2020
MfS GH 48/78	20.02.2020
MfS GH 17/79	24.02.2020
MfS GH 19/62	24.02.2020
MfS GH 59/82	24.02.2020
MfS HA I 18698	25.02.2020
MfS HA II 24250	25.02.2020
MfS HA II 24455	25.02.2020
MfS HA XXII 18390	25.02.2020
MfS HA II 32945	26.02.2020
MfS HA II 34916	26.02.2020
MfS HA II 34917	27.02.2020
MfS HA II 35301	27.02.2020
MfS HA II 35322	27.02.2020
MfS HA II 38991	03.03.2020
MfS HA II 38994	03.03.2020
MfS HA II 38995	03.03.2020
MfS HA II 38996	03.03.2020
MfS HA I 18623	04.03.2020
MfS HA II 40418	04.03.2020
MfS HA IX 24368	04.03.2020
MfS HA IX 24419	05.03.2020
MfS HA IX 24823	05.03.2020
MfS HA IX 25267	05.03.2020
MfS HA IX 25268	05.03.2020
MfS HA IX 25283	05.03.2020
MfS HA IX 25304	05.03.2020
MfS HA IX 25353	05.03.2020
MfS HA IX 25368	05.03.2020
MfS HA IX 25387	05.03.2020
MfS HA IX 25458	05.03.2020

Table B.1. List of Archival Sources

Signature	Accessed
MfS HA IX 25468	06.03.2020
MfS HA IX 25505	06.03.2020
MfS HA IX 25582	06.03.2020
MfS HA IX 25625	06.03.2020
MfS HA VII 305	06.03.2020
MfS HA II 39419	09.03.2020
MfS HA II 39421	09.03.2020
MfS HA VII 1779	09.03.2020
MfS HA XX/9 1934	09.03.2020
MfS HA XX/9 2125	09.03.2020
MfS HA XX 350	10.03.2020
MfS HA XX 351	10.03.2020
MfS HA XX 360	10.03.2020
MfS HA XX 4891	10.03.2020
MfS HA XX 4908	10.03.2020
MfS HA XX 15456	11.03.2020
MfS HA XX 16338	11.03.2020
MfS HA XX 16405	11.03.2020
MfS HA XX 17809	11.03.2020
MfS HA XX 20617	11.03.2020
MfS HA XX 20622	11.03.2020
MfS HA XX 20638	12.03.2020
MfS HA XX 21217	12.03.2020
MfS HA XX 21260	12.03.2020
MfS HA XX 21262	12.03.2020
MfS HA XX 21266	12.03.2020
MfS HA XX 23001	12.03.2020
MfS HA XX 23027	12.03.2020
MfS HA XIX 5159	16.07.2020
MfS HA XIX 5499	16.07.2020
MfS HA XIX 6186	16.07.2020
MfS HA XIX 6191	16.07.2020
MfS HA XIX 6192	16.07.2020
MfS HA XIX 6193	16.07.2020
MfS HA XIX 6195	16.07.2020
MfS HA XIX 6197	17.07.2020
MfS HA XIX 6198	17.07.2020
MfS HA XIX 6199	17.07.2020
MfS HA XIX 6200	17.07.2020
MfS HA XIX 6201	17.07.2020
MfS HA XX/AKG 3891	17.07.2020
MfS HA XX/AKG 5767	17.07.2020
MfS HA XX/AKG 5857	17.07.2020
MfS HA XX/AKG 6215	17.07.2020
MfS HA XX 23050	05.10.2020
MfS HA XX 23053	05.10.2020

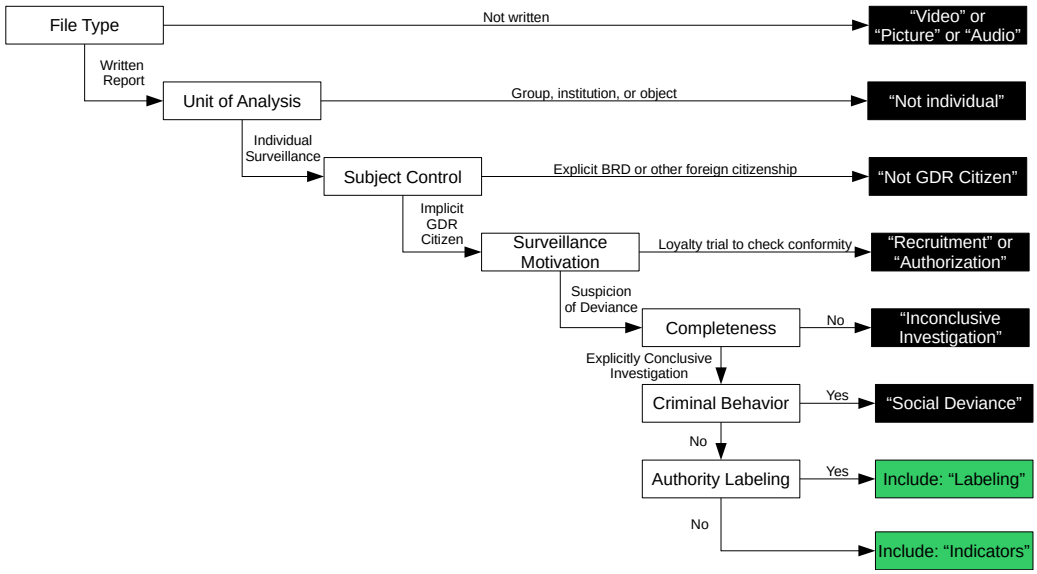
Table B.1. List of Archival Sources

Signature	Accessed
MfS HA XX 23069	05.10.2020
MfS HA XX 23071	05.10.2020
MfS HA XX 23085	05.10.2020
MfS HA XX 23090	05.10.2020
MfS HA XX 23101	05.10.2020
MfS HA XX 23132	05.10.2020
MfS HA XX 23144	05.10.2020
MfS ZKG 16371	06.10.2020
MfS ZKG 23168	06.10.2020
MfS ZKG 23603	06.10.2020
MfS ZKG 2391	06.10.2020
MfS ZKG 24125	06.10.2020
MfS ZKG 24740	06.10.2020
MfS ZKG 27462	06.10.2020
MfS HA XVIII 10584	07.10.2020
MfS HA XVIII 12534	07.10.2020
MfS HA XVIII 3607	07.10.2020
MfS HA XVIII 5783	07.10.2020
MfS HA XVIII 6121	07.10.2020
MfS HA XVIII 6129	07.10.2020
MfS HA XVIII 6268	07.10.2020
MfS HA XVIII 6320	07.10.2020
MfS HA XVIII 12537	08.10.2020
MfS HA XVIII 12585	08.10.2020
MfS HA XVIII 17369	08.10.2020
MfS HA XVIII 19974	08.10.2020
MfS HA XVIII 20378	09.10.2020
MfS HA XVIII 25478	09.10.2020
MfS HA XVIII 25480	09.10.2020
MfS HA XVIII 22626	29.03.2021
MfS HA XVIII 26290	29.03.2021
MfS HA XVIII 27236	31.03.2021
MfS HA XVIII 27817	31.03.2021
MfS HA XVIII 27823	31.03.2021
MfS HA XVIII 27825	31.03.2021
MfS HA XVIII 28592	31.03.2021
MfS HA XVIII 28071	01.04.2021
MfS HA XVIII 28150	01.04.2021
MfS HA XVIII 28454	07.06.2021
MfS HA XVIII 29283	07.06.2021
MfS HA XVIII 29656	07.06.2021
MfS HA XVIII 29968	08.06.2021
MfS HA XVIII 30095	08.06.2021
MfS HA XVIII 30226	08.06.2021
MfS HA XVIII 30773	09.06.2021
MfS HA XVIII 31158	09.06.2021

Table B.1. List of Archival Sources

Signature	Accessed
MfS HA XVIII 32032	09.06.2021
MfS HA XVIII 32980	09.06.2021
MfS HA XVIII 33040	10.06.2021
MfS HA XVIII 35058	19.07.2021
MfS HA XVIII 35168	19.07.2021
MfS HA XVIII 36905	20.07.2021
MfS HA XVIII 37797	20.07.2021
MfS HA II 32130	21.07.2021
MfS HA XVIII 28434	21.07.2021
MfS HA XVIII 38403	21.07.2021
MfS HA XVIII 38587	21.07.2021
MfS HA XVIII 38593	21.07.2021
MfS HA XXII 598	22.07.2021
MfS HA XXII 6037	22.07.2021
MfS AIM 1151/73	23.07.2021
MfS AOP 7920/91	01.11.2021
MfS AOP 8750/65	01.11.2021
MfS AOPK 635/86	02.11.2021
MfS AOPK 13012/82	03.11.2021
MfS GH 1/75	03.11.2021
MfS GH 10/89	03.11.2021
MfS GH 11/83	03.11.2021
MfS GH 22/67	04.11.2021
MfS GH 27/69	05.11.2021
MfS GH 28/79	05.11.2021
MfS GH 44/61	08.11.2021
MfS GH 55/78	08.11.2021
MfS GH 57/80	08.11.2021
MfS GH 61/59	08.11.2021
MfS GH 337/79	09.11.2021
MfS GH 339/79	09.11.2021
MfS GH 78/81	09.11.2021
MfS GH 361/79	10.11.2021
MfS HA XIX 7355	10.11.2021
MfS HA XIX 9362	10.11.2021
MfS HA XX/AKG 6249	10.11.2021
MfS HA XX/AKG 6275	10.11.2021
MfS HA XX/AKG 6769	10.11.2021

Figure B.1. Stage 2 and Stage 3 Sample: File Selection Decision-Tree



Note: Green boxes are included, black boxes excluded from the sample.

Figure B.2. Stage 2 Sample: Justifications for Exclusion of Files

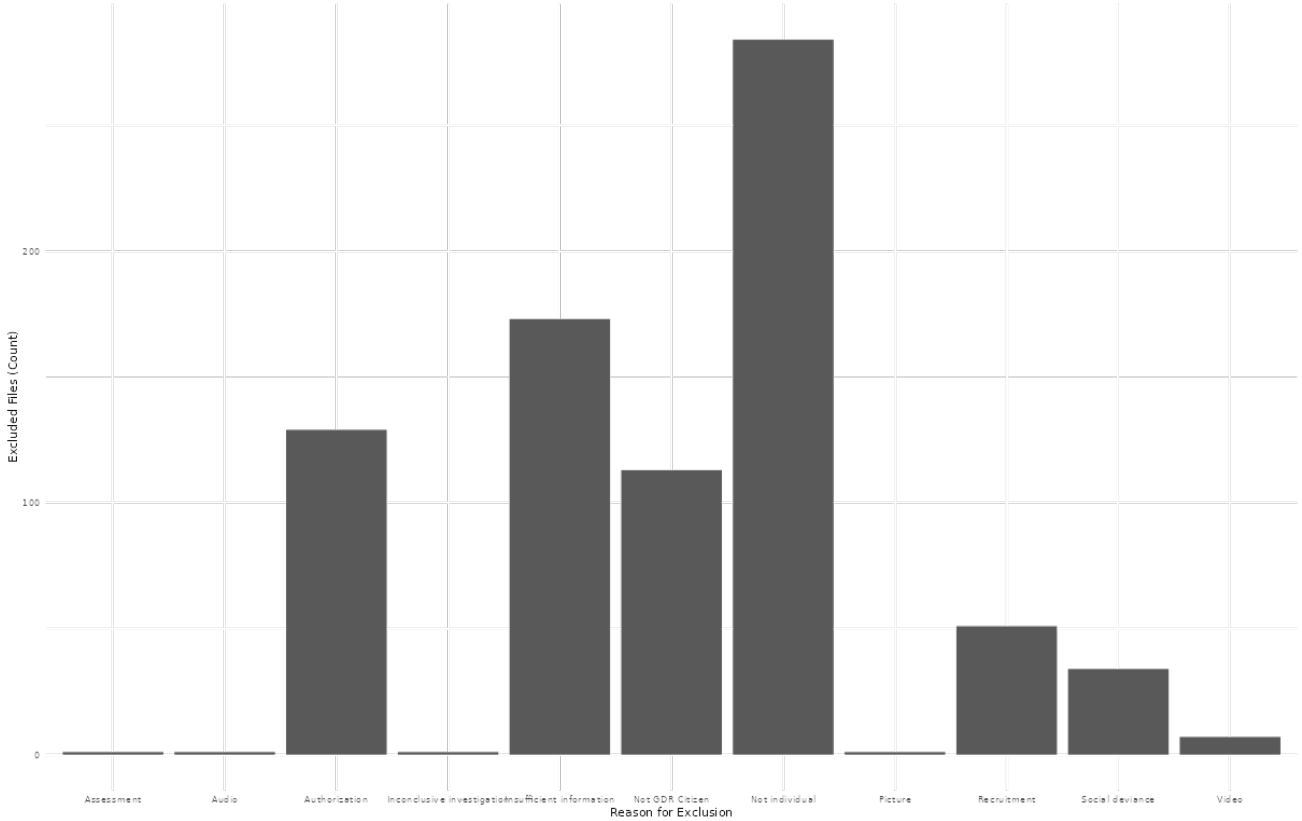
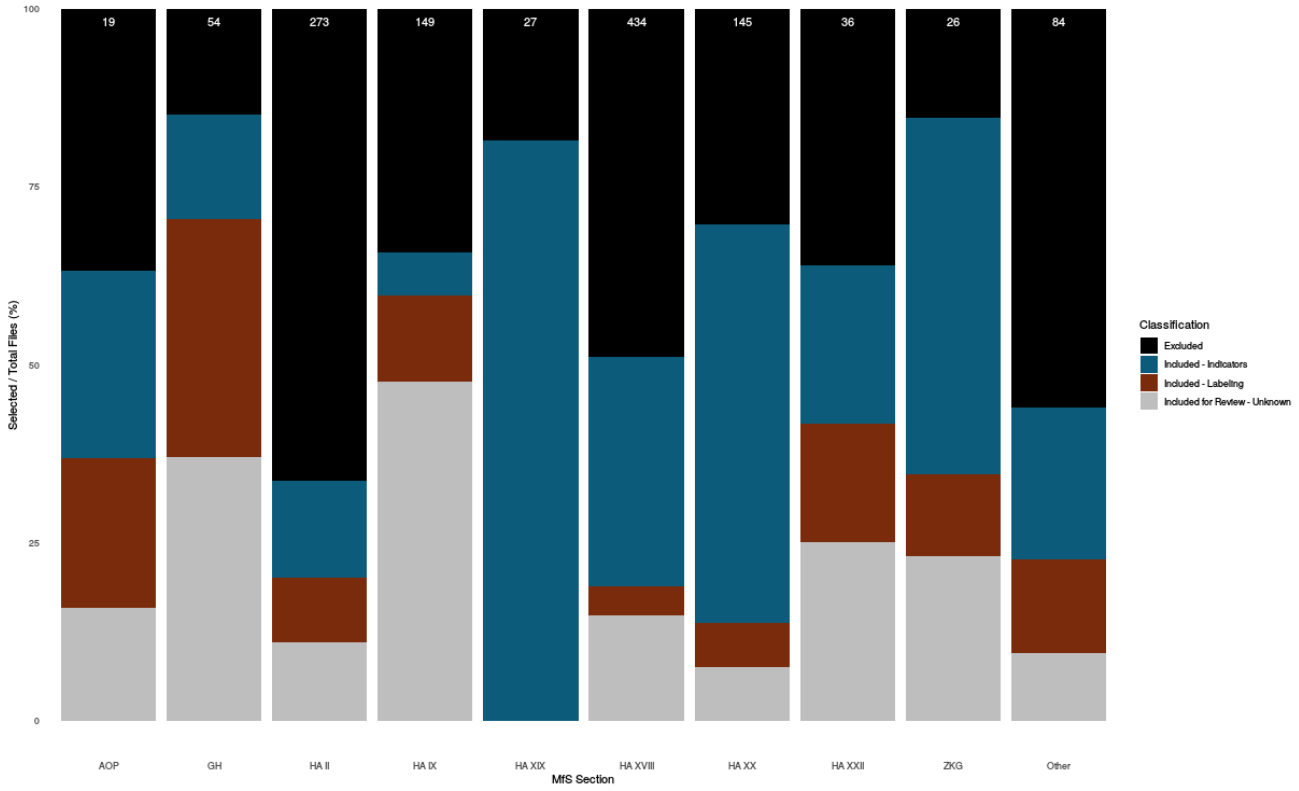


Figure B.3. Stage 2 Sample: Relative File Counts per Section.



Note: Numbers show absolute counts of *included* files. MfS sections with fewer than the third quartile of file counts were summarized in the category 'Other'.

Figure B.4. Coding Example

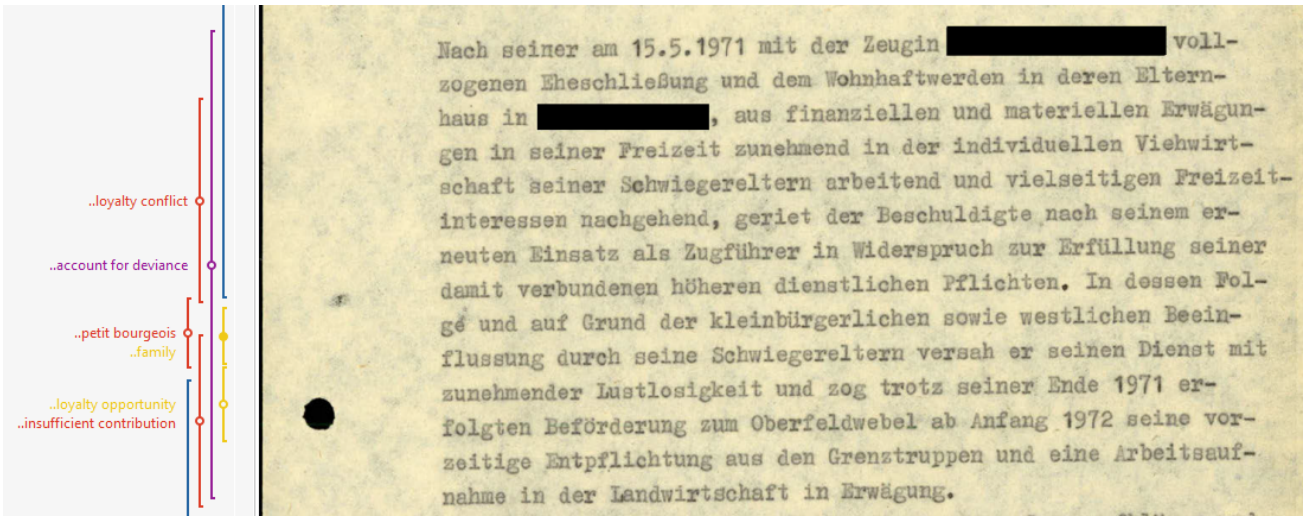
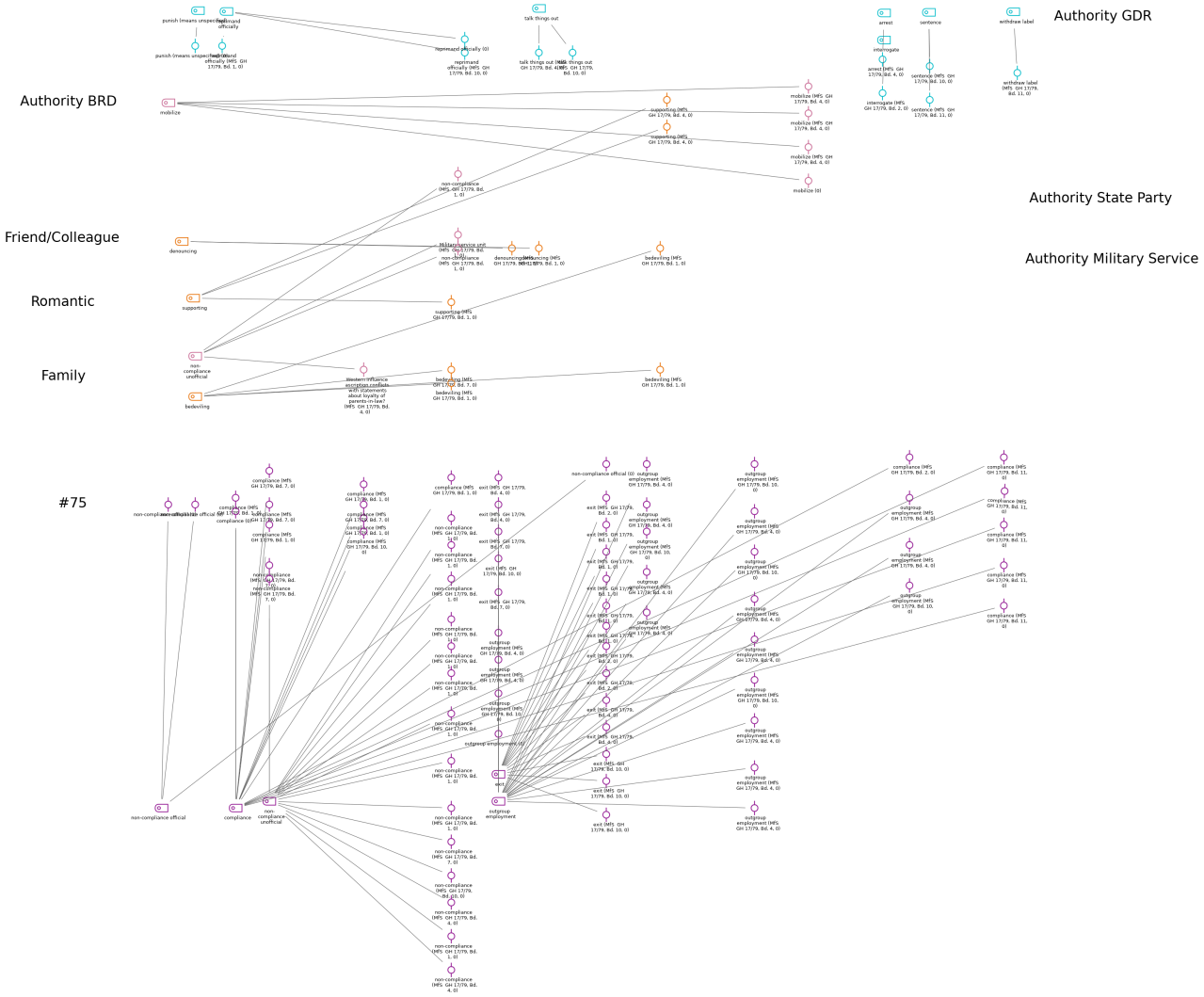


Figure B.5. ‘Deviant Defector’ Example of a Visual Aid for Database Construction



PAPER 3 APPENDICES

C.I. Robustness Checks

All data analysis is done in R (R Core Team 2018), using several packages to evaluate experimental designs and run statistical tests (Blair et al. 2019; Blair et al. 2021; Blair et al. 2022a; Blair et al. 2022b; Coppock 2022; Croissant and Millo 2008; Jackson 2011; Leifeld 2013; Waring et al. 2022; Wickham 2016; Wickham et al. 2019). Most of the discussion at this stage concerns the design of the game, rather than the experimental effects, given the scarcity of session-level data. The following sections analyze the development of the game design through pilot studies, give an overview of alternative treatments that were considered but ultimately dismissed, and show supplementary results tables and figures.

C.I.1. Pilot Summary

As shown in Table C.1 and Table C.2, the current pilot sample is not balanced. In order to move from a conventional group contest to a loyalty conflict, it was necessary to gradually introduce novel design features to understand how it affects participant behavior. This is both for reviewers who might take issue with the deviation from established experiments, and for ethical reasons as committee members might be worried that unknown games might lead to behavior that should not occur in the lab. Three different versions of the 'loyalty game' were tested thus far:

1. Pilot *v1*: Endowment Homogeneity vs. Heterogeneity
   + Defection (opportunity for all participants)
   + 15 Rounds
2. Pilot *v2*: Endowment Heterogeneity (Equality vs. Inequality/1 Poor)
   + Defection (opportunity for two participants)
   + 20 Rounds
3. Pilot *v2.1*: Endowment Heterogeneity (Equality vs. Inequality/2 Poor)
   + Defection (opportunity for two participants)
   + 20 Rounds
   + Scarcity

Version *v1* tested the least invasive condition compared to existing contests: do participants defect or not if everybody can switch? Predictably, participants in the chat condition figured out relatively quickly that they can coordinate on defection and thereby beat the game. Moreover, it became clear that the endowment heterogeneity condition created incentives for participants to state and discuss contributions as a proportion of their endowments, yet there were few exclusionary effects given the high variation across rounds.

Version *v2* therefore dropped the endowment homogeneity condition in favor of an inequality condition, and constrained opportunities to defect. This precluded coordination on defection, yet testing of the *Ineq-Chat* condition showed that the rich participants (3/4) tended to identify and believe low contributors that the game disadvantaged them. Seemingly content with their high endowments, rich

participants refrained from blaming poor participants, and in one case expressed the desire to balance out the unfair endowment distribution.

Version *v2.1* therefore introduced the *Scarcity* condition, which significantly increased winnings relative to endowments such that participants 'care' about the endogenous competition outcome more than the exogenous wealth accumulation mechanism. Specifically in the 'non-scarcity' condition, the nash equilibrium contribution is equal to competition winnings, as in previous group contest experiments that mainly study contribution levels ($\hat{e} = 100, v_i = 100$). By comparison in the current 'scarcity' condition, the nash equilibrium contribution is equal to the endowment ($\hat{e} = 60, v_i = 240$). In addition, the number of poor and rich participants changed between *v2.0* and *v2.1*: previously, only one participant was assigned the 'poor' type, whereas in the current design the group is polarized (two rich vs. two poor participants). This change is meant to reduce the risk of participants identifying their own type (which occurred in several pilot sessions previously), and to increase the probability for endowments to be centered around the nash contribution on average, without deceiving participants.

Some of the decisions-making observations in the pilots may have been due to the timing of the sessions. Shortly after the lab was re-opened (in April 2022), an experiment related to 'Oppression' was run, and due to a shortage of participants it was necessary to include participants of that study. In general, almost all participants indicated that they had previous experience with lab experiments, making strategic coordination more likely.[1] Due to technical issues and conflicting demands from different experiments (other researcher had to run studies for their dissertations last-minute, too), it was not possible to reach the whole participant pool consistently throughout the pilot study period. These issues are expected to be remedied for the main study, which will begin in mid October 2022 with a new participant pool.

C.I.2. Considered Treatments

Stages 1/5: 'Non-Scarcity'

In the current design, the mean endowment is equal to the nash equilibrium contribution that participants should spend on the prize, conditional on all participants playing nash and the endowment constraint: $\hat{e_i} = \hat{c_i}$. The 'non-scarcity' condition would instead increase participant endowments and decrease the prize, such that $\hat{e_i} = 100$ with $e_i \in [60, 80, 100, 120, 140]$, $v_i = 100$ and $\hat{c_i} = 25$. This reduces the pressure for participants to invest in the competition, and increases comparability to existing group contest experiments.

Stage 3: 'Penalty Incentive'

The participant who assigns the most penalty tokens to the player who ends up being penalized receives a reward, denoted by U_i. The reward is equal to the penalty of the penalized player, as per Equation (3.5):

$$U_{iI} = s_j(\sum_{i \neq j}^{n} l_{ij} \geq l_{kj}, \text{for all } k \neq j) \tag{C.1}$$

In cases where multiple participants tie for the most penalty tokens assigned to the penalized player, they split the reward equally. As before, participants are not informed who assigned penalty tokens to whom, and they are not informed who received a reward for assigning penalty tokens. Only the

[1] E.g. from chat messages in the *Inequality-Chat* condition: "yeah I think they like to pit people against each other and see what happens ... Yes, previous experiments in this lab had star participants who were always rich"

participants who receive a reward see the points they receive (and with how many other participants they had to split the reward).

Stage 3: 'Stigmatization'
The stigmatization of the penalized participant is increased by *labeling* them as *disloyal*. As before, whichever participant is penalized as per Equation (3.4) receives a material penalty that is deducted from per-period payoffs. In addition, the participant is socially penalized: in group statistics tables and in the chat, the 'disloyal' label appears next to their identifier. The label persists until another player is penalized. With this mechanism, labeled participants are expected to be suspected of defection in future periods. Moreover, participants who are labeled have an additional incentive to identify others as defectors.

Stages 3/4: 'Identification'
This condition is applied jointly with the opportunity to defect. Players may receive perfect information about the behavior of other group members: they see the *defection* choice of a punished participant from the previous round while the results are displayed.

After Round 10 in addition to introducing the possibility of defection, participants are informed that their defection may be revealed if they are punished. In Round 11 as before, they receive endowments, choose their contributions, potentially deliberate each others contributions in the chat, and assign punishment points. If a participant is punished in Round 11, other group members will see that they cooperated in Round 10. As before, they have the opportunity to defect before viewing the competition results. From Round 12, group members will see whether participants who are punished *cooperated* or *defected* in the previous round.

Participants are expected to reduce defection for fear of being identified, but increase punishment to identify defectors compared to conditions without identification.

C.I.3. Supplementary Figures and Tables

Table C.1. Design Development

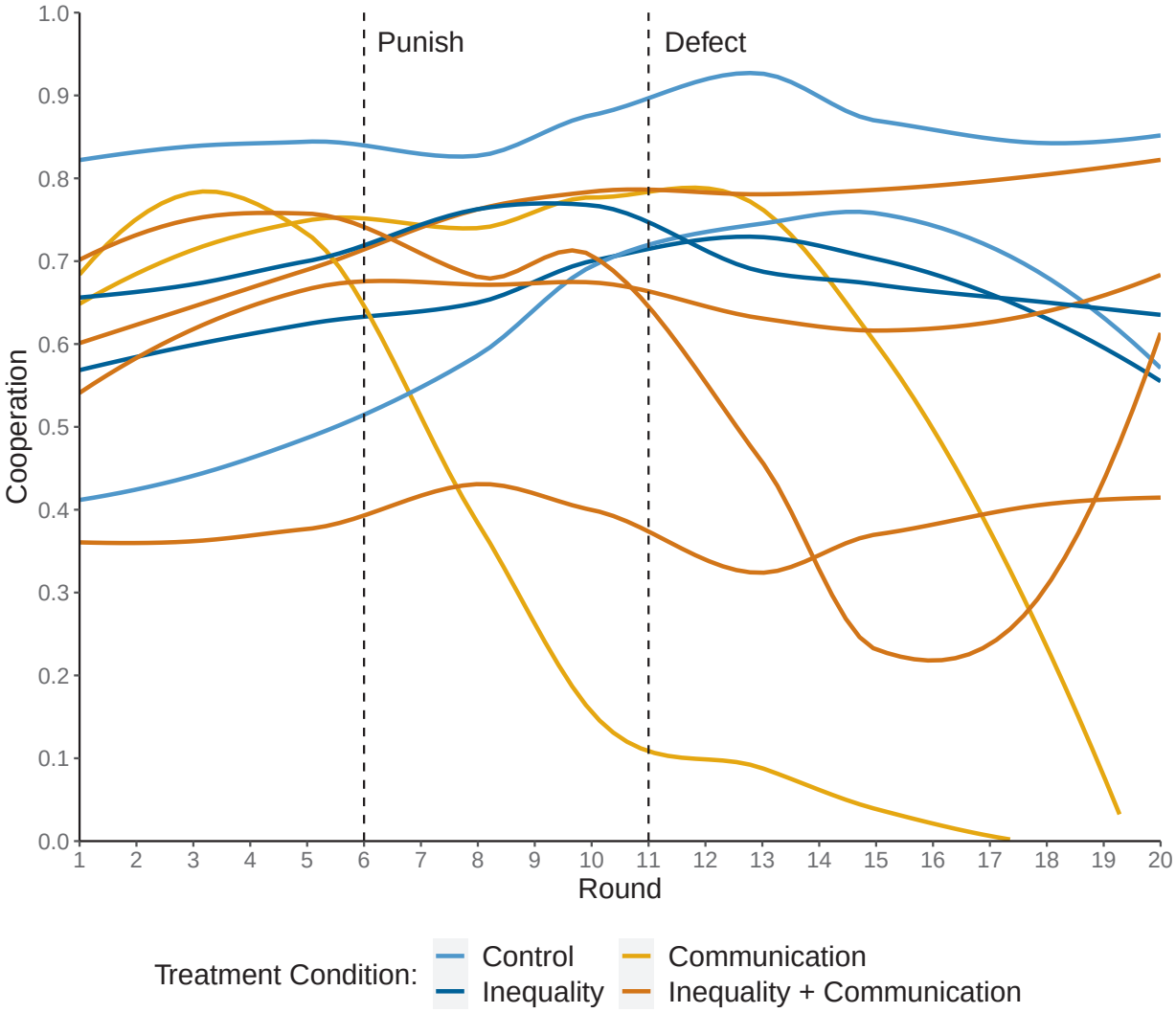
Game Version	**Treatment** (Between-Subject)	**Rounds** (Within-Subject)			**Participants**
		1-5	6-10	11-20	
v1	*Uncertainty*	Contest	+ Punish	+ Side-Switch*	16
	x Communication	Contest	+ Punish	+ Side-Switch*	16
v2	*Uncertainty*	Contest	+ Punish	+ Side-Switch	8
	x Inequality x Communication	Contest	+ Punish	+ Side-Switch	16
v2.1	*Uncertainty x Salience*	Contest	+ Punish	+ Side-Switch	8
	x Communication	Contest	+ Punish	+ Side-Switch	8
	x Communication x Inequality	Contest	+ Punish	+ Side-Switch	8
Sum	10	1440			80

Note: *In the first iteration all participants were allowed to side-switch and the session only lasted 15 rounds. From v2, only two participants are allowed to side-switch per round, and sessions lasted for 20 rounds.

Table C.2. Summary Statistics

Variable	Missing	Level Counts	Mean	SD	Min	Q1	Median	Q3	Max
Participant Type	0	Poor: 200, Rich: 440, Random: 320							
Contribution	0		54.26	34.10	0.00	25.00	50.00	80.00	150.00
Cooperation	0		0.62	0.31	0.00	0.40	0.66	0.92	1.00
Defection	720		0.40	0.49	0.00	0.00	0.00	1.00	1.00
Punished	0		0.10	0.30	0.00	0.00	0.00	0.00	1.00
Punishment	0		2.40	8.93	0.00	0.00	0.00	0.00	75.00
Group Identification	20		6.04	2.87	1.00	3.00	7.00	8.00	10.00
Participant-Rounds: 960 Sessions: 6									

Figure C.1. Cooperation by Group/Session



Cooperation for each group, with colors indicating session-level treatments. Notably, the spread of cooperation over time is relatively high for all but the *Inequality* condition.

Table C.3. Statistical Evaluation of Theoretical Propositions

	DV: Cooperation (OLS)				DV: Cohesion (OLS)	DV: Punishment (OLS-Lin
	Prop. 1	Prop. 2-3	Prop. 4	Prop. 1-4	Prop. 5	Prop. 7-8
Intercept	0.71*	0.72*	0.65*	0.78*	0.27*	0.18
	[0.49; 0.93]	[0.44; 1.00]	[0.54; 0.75]	[0.49; 1.07]	[0.19; 0.35]	[−0.73; 1.08]
Participant Type (Rich)	−0.11			−0.11		−0.12
	[−0.41; 0.18]			[−0.29; 0.07]		[−0.88; 0.63]
Communication		−0.12		−0.01	−0.02	
		[−0.53; 0.29]		[−0.75; 0.72]	[−0.07; 0.03]	
Team Attachment		−0.00		−0.01	−0.00	
		[−0.05; 0.04]		[−0.10; 0.07]	[−0.01; 0.00]	
Penalty Stage ($5 < t < 11$)			0.01	0.04		
			[−0.12; 0.14]	[−0.13; 0.21]		
Round					−0.00	
					[−0.01; 0.00]	
Contribution						−0.00
						[−0.02; 0.02]
Participant Type x Contribution						0.00
						[−0.02; 0.02]
R^2	0.04	0.04	0.00	0.06	0.05	0.04
Adj. R^2	0.03	0.04	−0.00	0.04	0.04	0.04
Num. obs.	640	940	480	310	940	640
RMSE	0.27	0.31	0.29	0.26	0.12	0.29
N Clusters	3	5	5	3	5	3

Significance levels: * = 0.05, ** = 0.01, *** = 0.001. Standard errors in parentheses.

Table C.4. Hypotheses Tests

	DV: Cooperation (OLS-Lin)	DV: Defection (OLS-Lin)	DV: Defection (OLS)		
	H1	H2	H3	H4	H5
Intercept	0.61*		0.62*	0.62*	0.63*
	[0.41; 0.82]		[0.47; 0.78]	[0.47; 0.78]	[0.47; 0.78]
Punished (Lag)					
Punishment (Lag)	0.00				
	[−0.12; 0.13]				
Group Win (Lag)	0.11	−0.33			
	[−0.10; 0.32]	[−0.91; 0.26]			
Punished (Lag) x Punishment (Lag)	−0.00				
	[−0.13; 0.13]				
Punished (Lag) x Group Win (Lag)	−0.05				
	[−0.33; 0.23]				
Punished		0.69*			
		[0.08; 1.30]			
Punishment		−0.05*			
		[−0.08; −0.01]			
Punished x Punishment		0.05*			
		[0.01; 0.09]			
Punished x Group Win (Lag)		0.31			
		[−0.48; 1.09]			
Inequality			−0.15		−0.15
			[−0.37; 0.07]		[−0.37; 0.07]
Communication				−0.18	−0.17
				[−0.40; 0.05]	[−0.39; 0.04]
Inequality x Communication					−0.03
					[−0.31; 0.26]
R^2	0.04	0.19	0.02	0.03	0.07
Adj. R^2	0.03	0.17	0.01	0.02	0.06
Num. obs.	720	240	80	80	240
RMSE	0.32	0.45	0.50	0.50	0.47
N Clusters	5	5			

Significance levels: * = 0.05, ** = 0.01, *** = 0.001. Standard errors in parentheses.

,

C.II. Information Sheet and Consent Form

Information Sheet for LE-0075

Welcome! You are about to participate in a decision-making experiment as part of a team. The experiment is being conducted by Dr. Noah Bacine (CESS Oxford) and Mirko Reul (Graduate Institute Geneva) in conjunction with the Centre for Experimental Social Sciences. In concordance with CESS policy, today's study has received ethics approval from CESS's independent review board.

We ask that you read this form carefully prior to deciding whether to participate in the upcoming laboratory session. Only individuals who complete this form will be allowed to participate. If you decide you do not want to participate, you may withdraw your participation at any time without providing a reason and without penalty by closing your browser. If you come to the session but choose to leave prior to the completion of the study, you will still receive £5 for coming to the session.

Purpose: The purpose of this study is to examine how people make decisions as individuals and within groups under various monetary incentives. The results of this study are intended to be used in aggregate for academic publications.

What Happens During the Study: The study requires you to sit in front of a computer terminal and make a series of decisions that may affect your final earnings and the final earnings of others. As part of the experiment, you may be asked to chat with other participants via an anonymous chat (please keep in mind that using offensive and/or inflammatory language could result in your exclusion from the study). Additionally, you may be asked to evaluate the decisions of others in your group, and others may be asked to evaluate your decisions. After the main part of your experiment, you will be asked to answer some survey questions about your personal characteristics, opinions, and experience during the session. At the conclusion of the study, participants will be asked to leave the laboratory one at a time to receive their payment in private. The entire study is expected to take approximately 2.0 hours.

Participation: The study is expected to take up to 2.0 hours. You always have the option of stopping your participation, and you may leave at any time without providing a reason and without penalty. If your behaviour is disruptive, you may be asked to leave.

Potential Risks: Participation in today's study poses minimal risk. None of the risks are greater than those encountered in daily life.

COVID-19 Considerations: Although COVID-19 has created additional concerns, we have

done our best to minimize these risks as much as possible).To reduce the possible risks associated with COVID-19: We encourage all participants to wear a face covering & sanitize their hands prior to entering the laboratory. We also ask that you help us ensure everyone's safety by not coming to the laboratory if you are experiencing any of the symptoms associated with the COVID-19 virus (for a list of common symptoms, please visit Symptoms of coronavirus (COVID-19) - NHS (www.nhs.uk)). In the event that we learn of a potential exposure that may have occurred during a laboratory session, we will inform all individuals who participated in the session immediately.
If you have any additional questions or concerns regarding our precautions against potential exposure to COVID-19, please contact cess-lab@nuffield.ox.ac.uk.

Benefits: Beyond the monetary compensation you will receive for your participation today, you are aiding the understanding of social behaviour.

Compensation: You will receive a minimum of £5 for coming to the lab on the day of the session before the scheduled start time. This amount will be given to you independent of your decision to participate in the session. Additionally, you will receive a bonus of £5 if you decide to participate and stay for the entirety of the session. Further, during the session you will have the opportunity to earn experimental tokens, which will be converted into cash at the end of the session, using an exchange rate of 125 tokens = £1. The amount you earn will depend on your decisions, the decisions of other participants, and luck. You will be paid in private and in cash at the end of the experiment. Although we cannot tell you exactly how much you will earn for your participation in the session, CESS participants historically earn an additional £10 per hour spent in the lab.

Data Protection & Privacy: During the experiment, you will only be identified using your unique subject ID (the same one you used to sign this form). You will not learn who your team members or your opponents are, neither during nor after the laboratory session. Likewise, neither your team members nor your opponents will learn about your identity.
The only place where your personal information appears is on the registration sheet for the session and the receipt that you will sign at the conclusion of the study. No personal information about you provided by you during this research will be disclosed to others without your written permission, except: if necessary to protect your rights or welfare (for example, if you are injured and need emergency care); or if required by law.

The information collected during the study will be kept private. In concordance with the Data Protection Act of 1998 & 2018, the University of Oxford is the data controller. Responsible members of the University of Oxford and funders may be given access to data for monitoring and/or audit of the study to ensure we are complying with guidelines or as otherwise required by law. The original data will be made anonymous and then stored on CESS's secure server that is only accessible by members of CESS. No one other than members of CESS or responsible members of the University of Oxford will have access to your personal information. The data will be stored in electronic form, password protected. All identifying information will be stored in a locked secure location. A copy of the anonymized data will be provided to the investigators

listed at the top of this form. The data that we collect from you may be transferred to, and stored and processed at, a destination outside the United Kingdom. By submitting your personal data, you agree to this transfer, storing, and processing. Your anonymity will be maintained in all publications or presentations resulting from this study.

Additional Information: If you are interested in receiving additional information about the results of the study, please contact noah.bacine@nuffield.ox.ac.uk or cess-lab@nuffield.ox.ac.uk.

Concerns: If you have any questions or concerns about any aspect of the project, you can contact the primary investigator (mirko.reul@graduateinstitute.ch) or CESS (cess-lab@nuffield.ox.ac.uk), who will do their best to answer your query. The researcher(s) should acknowledge reception of your concern within 10 working days and give you an indication of how they intend to address it. If you fail to receive a response, are dissatisfied with the response you receive, or desire to report an aspect of how the study is being conducted, please contact the relevant Chair of Research Ethics Committee at the University of Oxford:

Chair, Social Sciences & Humanities Inter-Divisional Research Ethics Committee;
Email: ethics@socsci.ox.ac.uk
Address: Research Services, University of Oxford, Wellington Square, Oxford OX1 2JD

The Chair will seek to resolve the matter in a reasonably expeditious manner.

Please confirm the following by marking each box and entering your Unique ID to signify that you have read, understood, and agreed to all of the information provided in this form. Note: Failure to complete this form correctly in advance of the study may result in your exclusion from the session as the ID you provide will be used to check that you have been properly consented prior to being allowed entry.

☐ I confirm that I have read and understood the information for the above study. I have had the opportunity to consider the information provided, ask questions, and have had these answered satisfactorily.

☐ I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason and without any adverse consequences.

☐ I understand that research data collected during the study may be looked at by designated individuals from the University of Oxford and CESS where it is relevant to my taking part in this study. I give permission for these individuals to access my data.

☐ I understand this project has been reviewed by, and received ethics approval through the CESS Ethics Review Committee.

☐ I understand who will have access to the personal data provided, how the data will be stored, and will happen to the data at the end of the project.

☐ I understand how this research will be written up and published.

☐ I understand how to raise a concern or make a complaint.

☐ I understand the risks associated with my participation and am still interested in being part of the study.

☐ I agree to not attend a session if I am experiencing any of the symptoms associated with the COVID-19 virus.

☐ I agree to take part in the study.

By entering my unique ID below, I confirm that I have read and agreed to all of the information presented above:

_____

Date (DD/MM/YYYY)

_____

,

C.III. Instructions

Instructions

Guide: homogenous endowment only, heterogenous endowment, chat only, priming only, penalty incentive only, scarcity only, non scarcity only

Intro

Welcome! You are about to take part in a decision-making experiment as part of a team. The other individuals in this room are also participating in the experiment. From now until the end of the study, please do not talk or communicate with anyone other than one of the experimenters (unless otherwise specified). We ask that you turn your phone off for the duration of today's session. At various points during today's study, you will likely have to wait while others make decisions. When this happens, we ask that you please wait patiently for the experiment to continue.

Today's study takes place over multiple parts. We will provide a description of how each part differs as we progress through the study.

In addition to the £5 you are guaranteed for coming to today's session, and the £5 you are guaranteed for completing today's experiment, you can earn considerably more depending on your decisions, the decisions of others, and luck. Please pay close attention to the instructions we provide as your decisions will determine your payoffs and the payoffs of others in this room. You will be paid in private and in cash at the end of the experiment.

During the experiment, you will have the chance to earn points, which will be converted into cash at the end of today's session using an exchange rate of 125 points = £1. We will round the sum of your points up to the nearest pound. Thus, the more points you earn, the more cash you will receive at the end of the session.

If you have a question at any point during today's study, please raise your hand and an experimenter will come to your station to assist you.

Part 1 Instructions

Each participant in today's session has been randomly assigned to a team with three other participants in this room. You will stay with the same team for the duration of today's study. During this part of the study, your team will be playing against one of the other teams and that team will be the opponent of your team. Please note that none of the participants in today's session will learn the identity of their teammates or opponents during or after today's session.

This part of the experiment consists of 5 rounds. In each round, your team and your opponents will compete for a prize, as will now be explained:

At the start of each round, you will receive 100 points from us which you can use to spend on the competition.

At the start of each round, you will be given points which you can use to spend on the competition. The number of points you receive each round is randomly determined by the computer. The amount you can receive ranges from 20 to

100, in increments of 20. Thus, you can receive 20, 40, 60, 80, or 100 points to start with in any round. The amount you can receive ranges from 60 to 140, in increments of 20. Thus, you can receive 60, 80, 100, 120, or 140 points to start with in any round. The points that you and all other participants receive are randomized each round, and the total number of points that your team receives may differ between rounds. However, the total number of points your team receives in a round will be equal to the total number of points received by the opposing team in that round. Put differently, each individual team member may receive different points, but competing teams will have the same total amount of points to spend on the competition.

You can then use these points to purchase 'competition tokens' for your team. Each competition token costs you 1 point, and you can choose to spend some, none or all of your points on competition tokens. Any points you choose not to invest in competition tokens will simply be added to your point balance and are yours to keep. Likewise, your teammates and your opponents will have the chance to buy tokens in exactly the same way.

After each participant has chosen how many competition tokens to buy, you will have an opportunity to review the number of competition tokens purchased by you and your teammates in that round as well as the total from all previous rounds. You will not receive any information about the competition tokens purchased by your opponents.

After you and your teammates have had a chance to review each other's decisions, you will have the opportunity to discuss with your teammates via a chat box. In the chat box, each member of a group can write messages that all members of the group see. You may use the chat to discuss any of your decisions in previous or current rounds, including each other's competition token purchases. We ask that you refrain from using any inappropriate language. The opportunity to chat lasts for 2 minutes in the first round of each part and for 1 minute in each round afterwards.

Then, the computer will determine whether you or your opponents win the prize as follows:

The computer sums the number of competition tokens purchased by each team which form that group's 'competition fund'. Then the computer runs a raffle in which each team gets a number of tickets equal to the number of tokens in their competition fund. The computer then selects one of those tickets at random and the team that owns that ticket wins the prize.

Thus, the probability of your team winning can also be expressed as:

$$\text{Probability of winning} = \frac{YourTeam's CompetitionFund}{(YourTeam's CompetitionFund + Opponent's CompeitionFund)}$$

This also means that your team's chances of winning the prize increases with the number of competition tokens bought by your team. Conversely, the more tokens your opponents buy, the higher the probability you lose.

If one of the teams doesn't buy any competition tokens, the other team automatically wins the prize. If neither team purchases competition tokens, neither team wins the prize.

Each member of the team who wins the prize receives an additional **240 100 points** to keep. Members of the losing team end the round with the number of points they **didn't** spend on competition tokens. Members of the winning team end the round with the number of points they **didn't** spend on competition tokens **plus 240 100 points**.

After the winner of the prize is determined, you will have a chance to review a summary of the round including your decision, the contributions of your teammates (but not your opponents), and your final earnings for the round prior to the start of the next round. You will also receive a summary of the total competition tokens purchased by each of your teammates across all rounds.

At the end of each round, the points that you kept and any prize you received will be added to your point total. When the next round begins, you will receive a new set of points which you can use to spend on competition tokens (you will not be able to use points from previous rounds to spend on competition tokens).

Part 2 Instructions

In the second part of the experiment, you will remain in the same team as in Part 1 and will continue to compete for a prize against the same opponents over 5 rounds. Unless mentioned otherwise below, the prize competition will occur in the same fashion as in Part 1. However, relative to Part 1, there will now be an additional stage which occurs after each participant has chosen how many competition tokens to purchase but before the winner of the prize is determined:

Once everyone on your team has made their competition token decisions and has had a chance to review their team members' choices, you will have the opportunity to apply up to 10 penalty tokens to each of them. Each penalty token you apply to one of your teammates will reduce your final earnings for that round by 1 point.

The participant who receives the most penalty tokens on your team will be penalized. Their final earnings for the round will be reduced by 3 times the number of penalty tokens they received in that round, and they will be labelled as disloyal until another member of the team is penalized. In the event that multiple team members are tied for the most penalty tokens received, the computer will randomly select one to be penalized.

The participant on your team who assigns the most penalty tokens to the penalized team member will receive the points that the penalized teammate loses (i.e. 3x the received points). In the event that multiple team members are tied for the most penalty tokens assigned to the team member who is penalized, the points will be split equally amongst them.

For example, if you were to receive a total of 10 penalty tokens from your teammates in a round, and that was the most anyone on your team received, then your final earnings for that round would be reduced by 30 points, while the team member(s) who assigned the most penalty tokens to you would be awarded 30 points.

You will still have the opportunity to chat during this part. The chat will occur after everyone has made their competition token decision but before the penalty token stage. You are free to discuss any of your decisions in previous or current rounds, including each other's competition token purchases and assignment of penalty tokens.

After each participant has chosen how many, if any, penalty tokens they wish to assign, the total number of penalty tokens each team member received (but not which team member assigned them) and which team member was penalized will be revealed. Further, the disloyalty label will appear next to the penalized participant on all screens until another player is labelled.

Afterwards, the winner of the prize is determined as in Part 1 and then each participant has a chance to review a summary of the round (as well as the total competition tokens purchased by each of your teammates thus far) before the next round begins.

Part 3 Instructions

In the third part of the experiment, you will remain in the same team as in Part 2 and will continue to compete for a prize against the same opponents over 10 rounds. Unless mentioned otherwise below, the prize competition will occur in the same fashion as in Part 2. However, relative to Part 2, there will now be an additional stage which occurs after the penalty token stage but before the winner of the prize is determined:

After you have reviewed your teammate's contributions, made your penalty token decisions, and the outcome of the penalty token stage has been revealed, two members of your team, selected at random by the computer, will have the opportunity to revisit their competition token purchase. During this stage, the two randomly selected team members will have the opportunity to choose whether they would like to keep their competition tokens in your team's fund or if they would prefer to switch them to your opponents in that round instead.

If they choose to keep their tokens in your team's competition fund, then the prize will be awarded in same manner as in the previous parts of the experiment. However, if they choose to switch, the competition tokens they contributed to your team's competition fund will be removed and applied to the opposing team's competition fund instead. In this case, they will not receive a reward if your team wins the prize, but will receive a reward of **192 80 points** if the opposing team wins instead.

The two team members who were not randomly selected to revisit their choice must keep their competition tokens in your team's competition fund.

While two members of your team have the opportunity to switch to the other team, so too do two members of the opposing team (selected at random); each randomly selected participant has the option to apply their competition tokens to the opposing team rather than their own if that's what they prefer (this also means that two members of each team must keep their contribution with their team).

You will not learn whether you have the opportunity to switch your contribution until after the penalty token assignment stage has been completed. The participants who have the option to switch their contribution decision in each team may be the same or different in each round. Whether you had the opportunity to switch in the previous round will have no impact on the possibility of you having the opportunity to switch in the following round. But in every round, your team and the opposing team each have two participants with the opportunity to switch.

You will still have the opportunity to chat during this part. The chat will occur after everyone has made their competition token decision but before the penalty token stage. You are free to discuss any of your decisions in previous or current rounds, including each other's competition token purchases, assignment of penalty tokens, and switching to the opponent.

After each participant who has the opportunity to do so has made their switching decision, the winner of the prize is determined as in Part 2 and then each participant has a chance to review a summary of the round as well as the total competition tokens purchased by each team member thus far before the next round begins. Which team members had the opportunity to switch and their decision to contribute to your team or to switch will **not** be visible to other participants.

As before, participants only observe their own contribution decision, the number of competition tokens purchased by their teammates (but not which team's fund they were applied to), whether they were penalized, and if they earned a reward from the group competition.

,

C.IV. Survey

Survey Questions (1/2) – Experiment

Please answer the following questions truthfully.
Your answers do not affect your final payout in any way.

1. What do you think the purpose of this experiment is? _____
2. Please tell us how you made your decisions over the course of today's session:
    - How did you choose the number of competition tokens you purchased in each round? _____
    - After round five, participants had the ability to apply penalty tokens to other team members. When, if ever, did you feel it was appropriate to penalize other team members? _____
    - After round ten, participants had the ability to "switch" to the opponent. When, if ever, did you feel it was appropriate to switch? _____
3. Please describe your feelings during the experiment: _____
4. Which of parts of game did you find difficult to understand? Please select the ones that you found difficult to understand, and do not select the ones that you fully understood after reading the instructions.
    - Points received in each round
    - Purchasing competition tokens
    - Applying penalty tokens to other team members
    - Switching to the opponent
    - Competition outcomes
5. Is there anything else you would like to tell us? _____

[SUBMIT AND CONTINUE TO PAGE 2]

Survey Questions (2/2) – Background

Please answer the following questions truthfully. Your answers do not affect your final payout in any way (*fields are mandatory).

6. *Please indicate your age group:
    - 19 or younger
    - 20-29
    - 30-39
    - 40-49
    - 50-59
    - 60-69
    - 70 or older
    - Prefer not to say
7. *How would you describe your gender?
    - Woman
    - Man
    - Non-binary / third gender

- Not listed: _____
- Prefer not to say

8. *Are you a student?
   - Yes
   - No

9. If you are or ever were a university student, in which department do/did you primarily study?
   - _____

10. *Which of the following categories best describes your current economic situation?
    - Lower class
    - Lower - middle class
    - Middle class
    - Middle - upper class
    - Upper class
    - Don't know
    - Prefer not to say

11. In political matters, people talk of "the left" and "the right". How would you place your views on this scale?
    Please use a scale from 1 to 10 where 1 means 'left' and 10 means 'right'.
    - 1 (Left) ... 10 (Right)

12. [GSS Risk]
    In general, how do you see yourself; are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?
    Please use a scale from 1 to 5, where 1 means you are 'completely unwilling to take risks' and 5 means you are 'very willing to take risks'.
    - 1...5

13. [GSS Trust] Do you think most people would try to take advantage of you if they got the chance, or do you think most people can be trusted?
    - Would take advantage of you
    - Would try to be fair
    - Depends
    - Don't know

14. [DOSPERT subset items indicating willingness to deviate from a covenant] For each of the following statements, please indicate the likelihood that you would engage in the described activity or behavior if you were to find yourself in that situation: Please use a scale from 1 to 7, where 1 means 'extremely unlikely' and 7 means 'extremely likely'.
- Admitting that your tastes are different from those of a friend.
- Disagreeing with an authority figure on a major issue.
- Revealing a friend's secret to someone else.
- Speaking your mind about an unpopular issue in a meeting at work.

15. [Group Attachment] How closely attached did you feel to your own team throughout the experiment?

Please use a scale from 1 to 10, where 1 means 'not at all attached' and 10 means 'very attached'.
- 1 (Not at all attached) ... 10 (Very attached)

16. [Social Identification] How did you feel about your team during the experiment? Please use a scale from 1 to 7, where 1 means 'very much disagree' and 7 means 'very much agree'.
    - I enjoyed working with my team
    - I would like to work with this team in the future
    - I identify as a member of this team
    - I see myself as a member of this team
    - I am glad to belong to this team
    - I feel strong ties with other members of this team
    - I feel very different from other members of this team

17. Have you ever participated in any economics or psychology experimental studies before?
    - Yes
    - No

18. How familiar are you with 'social deduction games'?.
    Please use a scale from 1 to 5, where 1 means 'not at all familiar' and 5 means 'very familiar'.
    A social deduction game is any game where one set of players has to identify another set of players, who in turn must try to remain hidden. Examples include 'Ultimate Werewolf', 'Mafia', 'Spyfall', and 'Among Us'.
    - 1 (Not at all familiar) ... 5 (Very familiar)

19. Please check this box to give permission to CESS to contact you about a short follow-up interview connected to this research project: [ ]

BIBLIOGRAPHY

Abbink, Klaus, Jordi Brandts, Benedickt Herrmann, and Henrik Orzen. 2010. "Intergroup Conflict and Intra-Group Punishment in an Experimental Contest Game." *The American Economic Review* 100 (1): 420–447.

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

Abrams, Dominic, Giovanni A. Travaglino, José M. Marques, Isabel Pinto, and John M. Levine. 2018. "Deviance Credit: Tolerance of Deviant Ingroup Leaders Is Mediated by Their Accrual of Prototypicality and Conferral of Their Right to Be Supported." *Journal of Social Issues* 74 (1): 36–55.

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

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

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

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

Albzour, Mai. 2017. "The Deconstruction of the Concept of Normalization within the Context of the Settler-Colonialism in Palestine: The Duality of Acceptance and Rejection." *Jadal Journal of Mada Al-Carmel* 31:1–9.

Albzour, Mai, Sandra Penic, Randa Nasser, and Eva G. T. Green. 2019. "Support for 'Normalization' of Relations between Palestinians and Israelis, and How It Relates to Contact and Resistance in the West Bank." *Journal of Social and Political Psychology* 7 (2): 978–996.

Ali, Nijmeh. 2019. "Active and Transformative Sumud Among Palestinian Activists in Israel." In *Palestine and Rule of Power: Local Dissent vs. International Governance*, ed. by Alaa Tartir and Timothy Seidel, 71–103. Palgrave Macmillan.

Andreoni, James, and Laura K. Gee. 2012. "Gun for Hire: Delegated Enforcement and Peer Punishment in Public Goods Provision." *Journal of Public Economics* 96 (11-12): 1036–1046.

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

Axelrod, Robert. 1986. "An Evolutionary Approach to Norms." *American Political Science Review* 80 (04): 1095–1111. ISSN: 0003-0554, 1537-5943.

B'Tselem. 2021a. "Database on Fatalities." Visited on 09/01/2021. https://statistics.btselem.org/en/intro/fatalities.

— . 2021b. "Statistics on the Death Penalty in the Palestinian Authority and Under Hamas Control in Gaza." Visited on 09/20/2020. https://www.btselem.org/download/death_penalty_verdicts_eng.xls.

Baik, Kyung Hwan. 1993. "Effort Levels in Contests: The Public-Good Prize Case." Ed. by Alan A. Lockard and Gordon Tullock. *Economic Letters* 41:363–367.

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

Balcells, Laia, and Christopher Sullivan. 2018. "New Findings from Conflict Archives: An Introduction and Methodological Framework." *Journal of Peace Research* 55 (2): 137–146.

Baltag, Alexandru. 2002. "A Logic for Suspicious Players: Epistemic Actions and Belief-Updates in Games." *Bulletin of Economic Research* 54 (1): 1–45.

BArch. *MfS, AOP, Nr. 16183/81.*

— . *MfS, AP, Nr. 14791/72.*

— . *MfS, BV Potsdam, Abt. XX, Nr. 790.*

— . *MfS, BV Potsdam, KD KY, Nr. 75.*

— . *MfS, GH, Nr. 107/80.*

— . *MfS, GH, Nr. 16/77.*

— . *MfS, GH, Nr. 17/79.*

— . *MfS, GH, Nr. 337/79.*

— . *MfS, HA I, Nr. 14995.*

— . *MfS, HA II, Nr. 32130.*

— . *MfS, HA II, Nr. 38995.*

— . *MfS, HA IX, Nr. 25283.*

— . *MfS, HA IX, Nr. 25609.*

— . *MfS, HA VII, Nr. 305.*

— . *MfS, HA XVIII, Nr. 25480.*

— . *MfS, HA XVIII, Nr. 28434.*

— . *MfS, HA XVIII, Nr. 37797.*

— . *MfS, HA XVIII, Nr. 38403.*

— . *MfS, HA XVIII, Nr. 6320.*

— . *MfS, HA XX, Nr. 350.*

— . *MfS, HA XX/AKG, Nr. 6215.*

BBC. 2017. "Bowe Bergdahl Pleads Guilty to Desertion." BBC News. Visited on 12/11/2019. https://www.bbc.com/news/world-us-canada-41642168.

Be'er, Yizhar, and Saleh Abdel-Jawad. 1994. *Collaborators in the Occupied Territories: Human Rights Abuses and Violations*. B'Tselem.

Becker, Howard S. 2017. *Evidence*. University of Chicago Press.

— . 1963. *Outsiders*. New York and London: Free Press.

Beetham, David. 2002. *The Legitimation of Power*. 5th ed. Issues in Political Theory. Basingstoke: Macmillan.

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

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

Bhavnani, Ravi. 2006. "Ethnic Norms and Interethnic Violence: Accounting for Mass Participation in the Rwandan Genocide." *Journal of Peace Research* 43 (6): 651–660.

Bhavnani, Ravi, Dan Miodownik, and Hyun Jin Choi. 2011. "Violence and Control in Civil Conflict: Israel, the West Bank, and Gaza." *Comparative Politics* 44 (1): 61–80.

al-Bitawi, Ahmad Hamed. 2016. *Palestinian Agents and Spies: Israel's Third Eye*. Al-Zaytouna Centre for Studies & Consultations.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113 (3): 838–859.

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys, Clara Bicalho, Neal Fultz, and Lily Medina. 2021. *DesignLibrary: Library of Research Designs*.

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys, Aaron Rudkin, and Neal Fultz. 2022a. *Fabricatr: Imagine Your Data before You Collect It*.

Blair, Graeme, Jasper Cooper, Alexander Coppock, Macartan Humphreys, and Luke Sonnet. 2022b. *Estimatr: Fast Estimators for Design-Based Inference*.

Booß, Christian. 2018. "Im Goldenen Käfig: Die Politische Justiz Und Die Anwälte in Der DDR Der Ära Honecker." *Remembrance and Justice* 31 (1): 386–403.

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

Bornstein, Gary. 1992. "The Free-Rider Problem in Intergroup Conflicts Over Step-Level and Continuous Public Goods." *Interpersonal Relations and Group Processes* 62 (4): 597–606.

Braithwaite, John. 1989. *Crime, Shame, and Reintegration*. Cambridge: Cambridge University Press.

Brücher, Lars. 2000. "Das Westfernsehen Und Der Revolutionäre Umbruch in Der DDR Im Herbst 1989." Magisterarbeit, Bielefeld.

BStU. *MfS, BV Rostock, AKG, Nr. 559*.

Budde, Heidrun. 2014. "Politische Fremdbestimmung Durch Gruppen: Stabilisator Des SED-Staates." In *Deutschland Archiv*.

Casari, Marco, and Luigi Luini. 2009. "Cooperation Under Alternative Punishment Institutions: An Experiment." *Journal of Economic Behavior & Organization* 71 (2): 273–282.

Cason, Timothy N., Roman M. Sheremeta, and Jingjing Zhang. 2012. "Communication and Efficiency in Competitive Coordination Games." *Games and Economic Behavior* 76 (1): 26–43.

Castano, Emanuele, Maria-Paola Paladino, Alastair Coull, and Vincent Y. Yzerbyt. 2002. "Protecting the Ingroup Stereotype: Ingroup Identification and the Management of Deviant Ingroup Members." *British Journal of Social Psychology* 41 (3): 365–385.

Cederman, Lars-Erik, Kristian Skrede Gleditsch, and Halvard Buhaug. 2013. *Inequality, Grievances, and Civil War*. New York: Cambridge University Press.

Charness, Gary, Ramón Cobo-Reyes, and Natalia Jiménez. 2014. "Identities, Selection, and Contributions in a Public-Goods Game." *Games and Economic Behavior* 87:322–338.

Chassany, Anne-Sylvaine. 2017. "France: The Permanent State of Emergency." Financial Times. Visited on 04/07/2020. https://www.ft.com/content/f5309ff8-a521-11e7-9e4f-7f5e6a7c98a2.

Chen, Daniel L, Martin Schonger, and Chris Wickens. 2016. "OTree: An Open-Source Platform for Laboratory, Online, and Field Experiments." *Journal of Behavioral and Experimental Finance* 9:88–97.

Cinyabuguma, Matthias, Talbot Page, and Louis Putterman. 2006. "Can Second-Order Punishment Deter Perverse Punishment?" *Experimental Economics* 9 (3): 265–279.

Cohen, Hillel. 2008. *An Army of Shadows: Palestinian Collaborators in the Service of Zionism, 1917-1948*. University of California Press.

— . 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.

— . 2010. "The Matrix of Surveillance in Times of National Conflict." In *Surveillance and Control in Israel/Palestine: Population, Territory and Power*, ed. by Elia Zureik, David Lyon, and Yasmeen Abu-Laban, 99–112. Routledge.

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

Cohen, Stanley. 2011. *Folk Devils and Moral Panics: The Creation of the Mods and Rockers*. 5th ed. Abingdon, Oxon ; New York: Routledge.

Cope, Kevin L, Charles Crabtree, and Yonatan Lupu. 2018. "Beyond Physical Integrity." *Law and Contemporary Problems* 81 (4): 185–195.

Coppock, Alexander. 2022. *Randomizr: Easy-to-use Tools for Common Forms of Random Assignment and Sampling*. Manual.

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

Croissant, Yves, and Giovanni Millo. 2008. "Panel Data Econometrics in R: The Plm Package." *Journal of Statistical Software* 27 (2): 1–43.

Cunningham, David. 2004. *There's Something Happening Here: The New Left, the Klan, and Fbi Counterintelligence*. Berkeley: University of California Press.

Davenport, Christian. 2007. "State Repression and Political Order." *Annual Review of Political Science* 10 (1): 1–23.

— . 2022. "The Art of Keeping the People in Line: Lisa Wedeen's *Ambiguities of Domination* after 20 Years." *PS: Political Science & Politics* 55 (1): 44–47.

— . 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.

Davenport, Christian, and Molly Inman. 2012. "The State of State Repression Research since the 1990s." *Terrorism and Political Violence* 24 (4): 619–634.

DCI Palestine. 2012. *Recruitment and Use of Palestinian Children in Armed Conflict*. Defence for Children International - Palestine.

De Dreu, Carsten, Jörg Gross, Zsombor Méder, Michael Giffin, Eliska Prochazkova, Jonathan Krikeb, and Simon Columbus. 2016. "In-Group Defense, Out-Group Aggression, and Coordination Failures in Intergroup Conflict." *Proceedings of the National Academy of Sciences* 113 (38): 10524–10529.

DeAngelo, Gregory, and Laura K. Gee. 2018. "Peers or Police? Detection and Sanctions in the Provision of Public Goods." *IZA Discussion Paper*, no. 11540.

Dechenaux, Emmanuel, Dan Kovenock, and Roman M. Sheremeta. 2015. "A Survey of Experimental Research on Contests, All-Pay Auctions and Tournaments." *Experimental Economics* 18 (4): 609–669.

Della Porta, Donatella. 1995. *Social Movements, Political Violence, and the State: A Comparative Analysis of Italy and Germany*. Cambridge and New York: Cambridge University Press.

Ditrich, Lara, and Kai Sassenberg. 2016. "It's Either You or Me! Impact of Deviations on Social Exclusion and Leaving." *Group Processes & Intergroup Relations* 19 (5): 630–652.

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

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

Dworschak, Christoph. 2020. "Jumping on the Bandwagon: Differentiation and Security Defection during Conflict." *Journal of Conflict Resolution* 64 (7-8): 1335–1357.

Eck, Kristine, Sophia Gabrielle Levin Hatz, Charles Crabtree, and Atsushi Tago. 2020. "Evade and Deceive? Citizen Responses to Surveillance." *Journal of Politics* Forthcoming.

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

Eisenfeld, Bernd, and Peter Eisenfeld. 1999. "Widerständiges Verhalten in Der DDR 1976-1982." In *Opposition in Der DDR von Den 70er Jahren Bis Zum Zusammenbruch Der SED-Herrschaft*, ed. by Eberhard Kurt, 83–121. Wiesbaden: VS Verlag für Sozialwissenschaften.

Engelmann, Roger, and Frank Joestel. 2016. *Hauptabteilung IX: Untersuchung*. Ed. by BStU. Anatomie Der Staatssicherheit: Geschichte, Struktur Und Methoden. Berlin.

Esteban, Joan, and Debraj Ray. 2008. "Polarization, Fractionalization and Conflict." *Journal of Peace Research* 45 (2): 163–182.

Farrell, Joseph. 1987. "Cheap Talk, Coordination, and Entry." *The RAND Journal of Economics* 18 (1): 34.

Farrington, David P., and Joseph Murray. 2014. "Empirical Tests of Labeling Theory in Criminology." In *Labeling Theory: Empirical Tests*, ed. by David P. Farrington and Joseph Murray, 1–9.

Faytre, Léonard. 2020. "Islamophobia in France: National Report 2019." In *European Islamophobia Report 2019*, ed. by Enes Bayraklı and Farid Hafez. Istanbul: SETA.

Fehr, Ernst, and Simon Gächter. 2000. "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* 90 (4): 980–994.

Fitzpatrick, Sheila. 1996. "Signals from Below: Soviet Letters of Denunciation of the 1930s." *The Journal of Modern History* 68 (4): 831–866.

Franzmann, Gabriele. 2009. *Bevölkerung in Der Ehemaligen DDR, 1946 Bis 1989*.

Freedman, David A. 2008. "Randomization Does Not Justify Logistic Regression." *Statistical Science* 23 (2): 237–249.

Gandhi, Jennifer, and Adam Przeworski. 2007. "Authoritarian Institutions and the Survival of Autocrats." *Comparative Political Studies* 40 (11): 1279–1301.

Garfinkel, Harold. 1967. *Studies in Ethnomethodology*. Englewood Cliffs: Prentice-Hall.

Gates, Scott. 2017. "Membership Matters: Coerced Recruits and Rebel Allegiance." *Journal of Peace Research* 54 (5): 674–686.

Geserick, Rolf. 1989. *40 Jahre Presse, Rundfunk Und Kommunikationspolitik in Der DDR*. München: Minerva Publikation.

Gieseke, Jens. 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*, ed. by Roger Engelmann and Clemens Vollnhals, 531–553. Berlin: Links.

— . 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.

— . 2003. “Die Einheit von Wirtschafts-, Sozial- Und Sicherheitspolitik: Militarisierung Und Überwachung Als Probleme Einer DDR-Sozialgeschichte Der Ära Honecker. Christoph Kleßmann Zum 65. Geburtstag.” *Zeitschrift für Geschichtswissenschaft* 51 (11): 996–1021.

— . 2014. *The History of the Stasi: East Germany’s Secret Police, 1945-1990*. Trans. by David Burnett. Berghahn Books.

Glaeser, Andreas. 2011. *Political Epistemics: The Secret Police, the Opposition, and the End of East German Socialism*. Chicago ; London: University of Chicago Press.

Glaser, Barney G., and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine.

Glocke, Nicole. 2002. “Werner Stiller: Versuch Eines Porträts.” *Zeitschrift des Forschungsverbundes SED-Staat* 11 (11): 102–109.

Goode, Erich. 2015. “The Sociology of Deviance: An Introduction.” In *The Handbook of Deviance*, ed. by Erich Goode, 3–29. Wiley-Blackwell.

Granovetter, Mark. 1978. “Threshold Models of Collective Behavior.” *American Journal of Sociology* 83 (6): 1420–1443.

Grechenig, Kristoffel, Andreas Nicklisch, and Christian Thöni. 2010. “Punishment Despite Reasonable Doubt-A Public Goods Experiment with Sanctions Under Uncertainty: Punishment Despite Reasonable Doubt.” *Journal of Empirical Legal Studies* 7 (4): 847–867.

Grimmer, Reinhard, ed. 2003. *Die Sicherheit: Zur Abwehrarbeit Des MfS*. Berlin: Das Neue Berlin.

Gulen, Fethullah. 2020. “Turkey to Detain 82 Military Officers Over Alleged Gulen Links.” Al Jazeera. Visited on 01/04/2021. https://www.aljazeera.com/news/2020/12/1/turkey-detaining-82-military-personnel-over-suspected-gulen-links.

Gunnthorsdottir, Anna, and Amnon Rapoport. 2006. “Embedding Social Dilemmas in Intergroup Competition Reduces Free-Riding.” *Organizational Behavior and Human Decision Processes* 101 (2): 184–199.

Gurr, Ted Robert. 1968. “Psychological Factors in Civil Violence.” *World Politics* 20 (2): 245–278.

— . 1970. *Why Men Rebel*. Princeton: Princeton University Press.

Gutiérrez-Sanín, Francisco, and Elisabeth Jean Wood. 2017. “What Should We Mean by ’Pattern of Political Violence’? Repertoire, Targeting, Frequency, and Technique.” *Perspectives on Politics* 15 (01): 20–41.

Gwladys, Fouche, Michelle Nichols, Charlotte Greenfield, Mark Bendeich, and Mike Collett-White. 2021. “Taliban Are Rounding up Afghans on Blacklist - Private Intel Report.” Reuters. Visited on 09/29/2021. https://www.reuters.com/world/asia-pacific/taliban-are-rounding-up-afghans-blacklist-private-intel-report-2021-08-19/.

Halbrock, Christian. 2015. "Denunziation, Meldetätitgkeit Und Informationserhebung Im Kapillarsystem Der SED-Diktatur." In *Hinter Vorgehaltener Hand: Studien Zur Historischen Denunziationsforschung*, 39:137–152. Analysen Und Dokumente: Wissenschaftliche Reihe Des Bundesbeauftragten Für Die Unterlagen Des Staatssicherheitsdienstes Der Ehemaligen Deutschen Demokratischen Republik (BStU). Vandenhoeck & Ruprecht.

Handel, Ariel, and Hilla Dayan. 2017. "Multilayered Surveillance in Israel/Palestine: Dialectics of Inclusive Exclusion." *Surveillance & Society* 15 (3/4): 471–476.

Hanisch, Anja. 2012. *Die DDR Im KSZE-Prozess 1972–1985: Zwischen Ostabhängigkeit, Westabgrenzung Und Ausreisebewegung*. Oldenbourg Wissenschaftsverlag.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–362.

Hass, Amira. 2019. "'I'm the Man Who Killed Your Husband': Palestinian Women Recount Shin Bet Interrogations." Haaretz. Visited on 12/03/2019. https://www.haaretz.com/israel-news/.premium-palestinian-women-tell-how-israel-interrogates-them-1.7962123.

Hassan, Mai, Daniel Mattingly, and Elizabeth R Nugent. 2022. "Political Control." *Annual Review of Political Science* 25:155–174.

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

Heckathorn, Douglas D. 1990. "Collective Sanctions and Compliance Norms: A Formal Theory of Group-Mediated Social Control." *American Sociological Review* 55 (3): 366.

— . 1988. "Collective Sanctions and the Creation of Prisoner's Dilemma Norms." *American Journal of Sociology* 94 (3): 535–562. ISSN: 0002-9602.

Henrich, Joseph, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker. 2006. "Costly Punishment Across Human Societies." *Science* 312 (5781): 1767–1770.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment Across Societies." *Science* 319 (5868): 1362–1367.

Hewitt, Steve. 2010. *Snitch! A History of the Modern Intelligence Informer*. New York: Continuum International.

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

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

Hoffmann, Ruth. 2012. *Stasi-Kinder: Aufwachsen Im Überwachungsstaat*. Ullstein eBooks.

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

Horz, Carlo M., and Moritz Marbach. 2022. "Economic Opportunities, Emigration and Exit Prisoners." *British Journal of Political Science* 52 (1): 21–40.

Hoshur, Shohret, Qiao Long, Joshua Lipes, and Luisetta Mudie. 2018. "Xinjiang Authorities Secretly Transferring Uyghur Detainees to Jails Throughout China." Radio Free Asia. Visited on 05/30/2019. https://www.rfa.org/english/news/uyghur/transfer-10022018171100.html.

Human Rights Watch. 2001. *Justice Undermined: Balancing Security and Human Rights in the Palestinian Justice System*, Israel, the Occupied West Bank and Gaza Strip, and the Palestinian Authority Territories Vol. 13 No. 4 (E).

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95.

Hutchison, Paul, Jolanda Jetten, and Roberto Gutierrez. 2011. "Deviant but Desirable: Group Variability and Evaluation of Atypical Group Members." *Journal of Experimental Social Psychology* 47 (6): 1155–1161.

Iannaccone, Laurence R. 1992. "Sacrifice and Stigma: Reducing Free-Riding in Cults, Communes, and Other Collectives." *Journal of Political Economy* 100 (2): 271–291.

Jackson, Christopher H. 2011. "Multi-State Models for Panel Data: The Msm Package for R." *Journal of Statistical Software* 38 (8): 1–29.

Jalal, Rasha Abou. 2015. "Spies' Families Marginalized in Gaza." Al-Monitor. Visited on 05/08/2020. https://www.al-monitor.com/pulse/originals/2015/02/gaza-spies-israel-resistance-execution-families-shame.html.

Jamal, Amaney A. 2022. "Ambiguities of Domination: 20 Years Later and We Are Still Not Getting It Right." *PS: Political Science & Politics* 55 (1): 48–51.

James, Keith, and Russell Cropanzano. 1994. "Dispositional Group Loyalty and Individual Action for the Benefit of an Ingroup: Experimental and Correlational Evidence." *Organizational Behavior and Human Decision Processes* 60:179–205.

JMCC. 1999. "Public Opinion Poll No. 30: On Palestinian - Israeli Peace Index." Jerusalem Media & Communication Centre. Visited on 02/22/2021. http://www.jmcc.org/polls.aspx.

— . 2000. "Public Opinion Poll No. 39, Part Two: Attitudes of the Israeli and Palestinian Publics towards the Peace Process." Jerusalem Media & Communication Centre. Visited on 02/22/2021. http://www.jmcc.org/polls.aspx.

Joester, Frank. 1999. "Verdächtigt Und Beschuldigt: Statistische Erhebungen Zur MfS-Untersuchungstätigkeit 1971-1988." In *Justiz Im Dienste Der Parteiherrschaft: Rechtspraxis Und Staatssicherheit in Der DDR*, ed. by Roger Engelmann and Clemens Vollnhals, 303–328. Berlin: Links.

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

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

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

— . 2003. "The Ontology of 'Political Violence': Action and Identity in Civil Wars." *Perspectives on Politics* 1 (3): 475–494.

Kao, Kristen, and Mara Redlich Revkin. 2022. "Retribution or Reconciliation? Post-Conflict Attitudes Toward Enemy Collaborators." *American Journal of Political Science*.

Katz, Eliakim, Shmuel Nitzan, and Jacob Rosenberg. 1990. "Rent-Seeking for Pure Public Goods." *Public Choice* 65 (1): 49–60. JSTOR: 30025243.

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

Kepley, David. 1984. "Sampling in Archives: A Review." *The American Archivist* 47 (3): 237–242.

Kocher, Matthew Adam, Thomas B. Pepinsky, and Stathis N. Kalyvas. 2011. "Aerial Bombing and Counterinsurgency in the Vietnam War: Bombing and Counterinsurgency in Vietnam." *American Journal of Political Science* 55 (2): 201–218.

Koehler, Kevin, Dorothy Ohl, and Holger Albrecht. 2016. "From Disaffection to Desertion: How Networks Facilitate Military Insubordination in Civil Conflict." *Comparative Politics* 48 (4): 439–457.

Konrad, Kai A. 2009. *Strategy and Dynamics in Contests*. Oxford University Press.

Kowalczuk, Ilko-Sascha. 2013. *Stasi Konkret: Überwachung Und Repression in Der DDR*. Beck.

Krähnke, Uwe, Matthias Finster, Anja Zschirpe, and Philipp Reimann. 2017. *Im Dienst Der Staatssicherheit: Eine Soziologische Studie Über Die Hauptamtlichen Mitarbeiter Des DDR-Geheimdienstes*. Campus Verlag.

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

Lauderdale, Pat. 2015. "Political Deviance." In *The Handbook of Deviance*, ed. by Erich Goode, 521–536. Wiley-Blackwell.

Leifeld, Philip. 2013. "Texreg: Conversion of Statistical Model Output in R to LATEX and HTML Tables." *Journal of Statistical Software* 55 (8).

Lemert, Edwin M. 1951. "Social Pathology: A Systematic Approach to the Theory of Sociopathic Behavior."

Levine, John M., and Richard L. Moreland. 2002. "Group Reactions to Loyalty and Disloyalty." In *Group Cohesion, Trust and Solidarity*, ed. by Shane R. Thye and Edward J. Lawler, 203–228. Emerald Group Publishing Limited.

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

Lieberman, Robert C. 2002. "Ideas, Institutions, and Political Order: Explaining Political Change." *American Political Science Review* 96 (04): 697–712.

Lin, Winston. 2013. "Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique." *The Annals of Applied Statistics* 7 (1).

Lindström, Björn, and Philippe N. Tobler. 2018. "Incidental Ostracism Emerges from Simple Learning Mechanisms." *Nature Human Behaviour* 2 (6): 405–414.

Liu, Howard. 2022. "Dissent Networks, State Repression, and Strategic Clemency for Defection." *Journal of Conflict Resolution* 66 (7-8): 1292–1319.

Livingston, Eric. 1987. *Making Sense of Ethnomethodology*. London ; New York: Routledge & Kegan Paul.

Lohmann, Susanne. 1994. "The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989–91." *World Politics* 47 (01): 42–101.

Lucas, Colin. 1996. "The Theory and Practice of Denunciation in the French Revolution." *The Journal of Modern History* 68 (4): 768–785.

Ludwig, Klaus-Dieter. 2008. "Das Wörterbuch Der Politisch-Operativen Arbeit - Ein Manipulierendes Wörterbuch - Und Das Wörterbuch Der Deutschen Gegenwartssprache - Ein Teilweise Manipuliertes Wörterbuch." In *Verschlüsseln, Verbergen, Verdecken in Öffentlicher Und Institutioneller Kommunikation*, ed. by Steffen Pappert, Melani Schröter, and Ulla Fix, 273–289. Berlin: Erich Schmidt Verlag.

Lyall, Jason, Yuki Shiraito, and Kosuke Imai. 2015. "Coethnic Bias and Wartime Informing." *The Journal of Politics* 77 (3): 833–848.

Maddrell, Paul. 2013. "Im Fadenkreuz Der Stasi: Westliche Spionage in Der DDR. Die Akten Der Hauptabteilung IX." *Vierteljahrshefte für Zeitgeschichte* 61 (2): 141–171.

Mago, Shakun D., Anya C. Samak, and Roman M. Sheremeta. 2016. "Facing Your Opponents: Social Identification and Information Feedback in Contests." *Journal of Conflict Resolution* 60 (3): 459–481.

Mampilly, Zachariah. 2011. *Rebel Rulers: Insurgent Governance and Civilian Life During War*. Ithaca: Cornell University Press.

Marques, José M., and Dario Paez. 1994. "The 'Black Sheep Effect': Social Categorization, Rejection of Ingroup Deviates, and Perception of Group Variability." *European Review of Social Psychology* 5 (1): 37–68.

Matza, David. 2010. *Becoming Deviant*. 2nd ed. New Brunswick, N.J: Transaction Publishers.

McAdam, Dough, Sidney Tarrow, and Charles Tilly. 2001. *Dynamics of Contention*. New York: Cambridge University Press.

McCarter, Matthew W., Anya C. (Savikhin) Samak, and Roman M. Sheremeta. 2013. "Divided Loyalties or Conditional Cooperation? An Experimental Study of Contributions to Multiple Public Goods." *ESI Working Papers* 13 (8).

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

McLauchlin, Theodore, and Álvaro La Parra-Pérez. 2018. "Disloyalty and Logics of Fratricide in Civil War: Executions of Officers in Republican Spain, 1936-1939." *Comparative Political Studies*: 1–31.

Merton, Robert King. 1968. *Social Theory and Social Structure*. Enlarged Edition. New York: Free Press.

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

Morgan, Sarah Blake. 2020. "Bergdahl Lawyers Say Judge's Job Application Posed Conflict." Washington Post. Visited on 04/16/2022. https://www.washingtonpost.com/national/bergdahl-lawyers-say-judges-job-application-posed-conflict/2020/09/18/f9f65350-f9fd-11ea-85f7-5941188a98cd_story.html.

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

Müller-Enbergs, Helmut. 2008. *Inoffizielle Mitarbeiter Des Ministeriums Für Staatssicherheit, Teil 3: Statistiken*. Ed. by BStU. In collab. with Susanne Muhle. Berlin.

Murphy, Maureen Clare. 2018. "Israel Uses Online Blackmail to Recruit Collaborators." The Electronic Intifada. Visited on 05/15/2020. https://electronicintifada.net/content/israel-uses-online-blackmail-recruit-collaborators/23461.

Nechepurenko, Ivan. 2019. "American Held in Russia on Spying Charge Must Stay in Prison." The New York Times. Visited on 12/03/2019. https://www.nytimes.com/2019/10/24/world/europe/paul-whelan-prison-russia-arrest.html.

Nerenberg, Daniel. 2016. "Cooperating with the Enemy: Purpose-Driven Boundary Maintenance in Palestine, 1967-2016." PhD thesis, George Washington University.

Nikiforakis, Nikos, and Hans-Theo Normann. 2008. "A Comparative Statics Analysis of Punishment in Public-Good Experiments." *Experimental Economics* 11 (4): 358–369.

Nitzan, Shmuel. 1991. "Collective Rent Dissipation." *The Economic Journal* 101 (409): 1522.

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

O'Conner, Nigel, and Lazar Simeonov. 2013. "Gay Palestinians Are Being Blackmailed Into Working As Informants." Visited on 06/09/2019. https://www.vice.com/en_uk/article/av8b5j/gay-palestinians-are-being-blackmailed-into-working-as-informants.

Oliver, Pamela E, and Gerald Marwell. 1988. "The Paradox of Group Size in Collective Action: A Theory of the Critical Mass. II." *American Sociological Review*: 1–8.

Oliver, Pamela, Gerald Marwell, and Ruy 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.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press.

Onishi, Norimitsu, and Constant Méheut. 2020. "France's Dragnet for Extremists Sweeps Up Some Schoolchildren, Too." The New York Times. Visited on 01/04/2021. https://www.nytimes.com/2020/11/23/world/europe/france-extremism-children.html.

Ophir, Adi. 2020. "Exit, Voice, Loyalty: The Case of the BDS." *Philosophy & Social Criticism* 46 (1): 25–33.

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

Ostrom, Elinor, James Walker, and Roy Gardner. 1992. "Covenants with and without a Sword: Self-Governance Is Possible." *American Political Science Review* 86 (2): 404–417.

Palestinian Central Bureau of Statistics. 2005. *Labour Force Survey: Annual Report 2004*. Ramallah.

— . 1999. *Population in the Palestinian Territory, 1997-2025*. Ramallah.

Pappert, Steffen. 2008. "Verdecken Und Verschüsseln Durch Fachsprache? Zur Transformation von Alltagssprache in Die Sprache Des MfS." In *Verschlüsseln, Verbergen, Verdecken in Öffentlicher Und Institutioneller Kommunikation*, ed. by Steffen Pappert, Melani Schröter, and Ulla Fix, 291–313. Berlin: Erich Schmidt Verlag.

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

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

Pearlman, Wendy, and Kathleen Gallagher Cunningham. 2012. "Nonstate Actors, Fragmentation, and Conflict Processes." *Journal of Conflict Resolution* 56 (1): 3–15.

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

Pfaff, Steven, and Hyojoung Kim. 2003. "Exit-Voice Dynamics in Collective Action: An Analysis of Emigration and Protest in the East German Revolution." *American Journal of Sociology* 109 (2): 401–444.

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

Pollack, Detlef. 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*, ed. by Detlef Pollack and Dieter Rink, 303–331. Frankfurt and New York: Campus Verlag.

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

Poulsen, Lauge N Skovgaard. 2020. "Loyalty in World Politics." *European Journal of International Relations* 26 (4): 1156–1177.

PSR. 2000. "Public Opinion Poll #1: 27-29 July 2000." Palestinian Center for Policy and Survey Research. Visited on 02/22/2021. http://www.pcpsr.org/sites/default/files/Palestinian%20Public%20Opinion%20Poll%20No%281%29%20with%20Table.pdf.

— . 2001. "Public Opinion Poll #2: 5-9 July 2001." Palestinian Center for Policy and Survey Research. Visited on 02/22/2021. http://www.pcpsr.org/sites/default/files/Palestinian%20Public%20Opinion%20Poll%20No%282%29%20%20with%20Table.pdf.

Pulwer, Sharon, and Gili Cohen. 2016. "Israel Refusing to Help Palestinians Who May Face Death for Selling Land to Jews." Haaretz. Visited on 11/02/2019. https://www.haaretz.com/israel-news/.premium-israel-won-t-help-palestinians-who-sell-land-to-jews-1.5393449.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna.

Rapoport, Amnon, and Gary Bornstein. 1987. "Intergroup Competition for the Provision of Binary Public Goods." *Psychological Review* 94 (3): 291–299.

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

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

Rembold, Elfie. 2003. ""'Dem Eindringen Westlicher Dekadenz Ist Entgegenzuwirken": Jugend Und Die Kultur Des Feindes in Der DDR." In *Fremde Und Fremd-Sein in Der DDR. Zu Historischen Ursachen Der Fremdenfeindlichkeit in Ostdeutschland*, ed. by Jan C. Behrends, Thomas Lindenberger, and Patrice G. Poutrus, 193–214. Berlin: Metropol Verlag.

Richter, Hedwig. 2015. "Die Effizienz Bürokratischer Normalität. Das Ostdeutsche Berichtswesen in Verwaltung, Parteien Und Wirtschaft." In *Hinter Vorgehaltener Hand: Studien Zur Historischen Denunziationsforschung*, 127–136. Vandenhoeck & Ruprecht.

Rink, Dieter. 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*, ed. by Detlef Pollack and Dieter Rink, 54–77. Frankfurt and New York: Campus Verlag.

Riolo, Rick, Michael Cohen, and Robert Axelrod. 2001. "Evolution of Cooperation without Reciprocity." *Nature* 414 (6862): 441–443.

Ritter, Emily Hencken. 2014. "Policy Disputes, Political Survival, and the Onset and Severity of State Repression." *Journal of Conflict Resolution* 58 (1): 143–168.

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

Rosenberg, Matthew. 2016. "Long After Bergdahl's Release, His Hometown Is Still Under Siege." The New York Times. Visited on 04/16/2022. https://www.nytimes.com/2016/02/17/us/long-after-bergdahls-release-his-hometown-is-still-under-siege.html.

Roy, Sara. 1987. "The Gaza Strip: A Case of Economic De-Development." *Journal of Palestine Studies* 17 (1): 56–88.

Salah, Hana. 2019. "Hamas Signals Alarm on Israeli Efforts to Recruit Collaborators." Al-Monitor. Visited on 03/07/2021. https://www.al-monitor.com/pulse/originals/2019/10/hamas-efforts-block-israeli-efforts-trap-gazans-collaborate.html.

Santoro, Wayne A., and Marian 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.

Schlichte, Klaus, and Ulrich Schneckener. 2015. "Armed Groups and the Politics of Legitimacy." *Civil Wars* 17 (4): 409–424.

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

Scott, James C. 1989. "Everyday Forms of Resistance." *The Copenhagen Journal of Asian Studies* 4 (1): 33–62.

— . 1985. *Weapons of the Weak: Everyday Forms of Peasant Resistance*. New Haven and London: Yale University Press.

Sebastian, Kailah. 2019. "Distinguishing Between the Types of Grounded Theory: Classical, Interpretive and Constructivist." *Journal for Social Thought* 3 (1): 1–9.

Seck, Hope Hodge. 2021. "Court Upholds Bowe Bergdahl's Sentence Despite Trump 'Dirty Traitor' Comments." Military.com. Visited on 04/16/2022. https://www.military.com/daily-news/2020/08/28/court-upholds-bowe-bergdahls-sentence-despite-trump-dirty-traitor-comments.html.

Sheremeta, Roman M. 2018. "Behavior in Group Contests: A Review of Experimental Research: Behavior in Group Contests." *Journal of Economic Surveys* 32 (3): 683–704.

— . 2009. "Perfect-Substitutes, Best-Shot, and Weakest-Link Contests Between Groups." *Korean Economic Review* 27:5–32.

Sheremeta, Roman M., and Jingjing Zhang. 2010. "Can Groups Solve the Problem of Over-Bidding in Contests?" *Social Choice and Welfare* 35 (2): 175–197.

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

— . 2014. "Experiments in Criminal Sanctions: Labeling, Defiance, and Restorative Justice." In *Labeling Theory: Empirical Tests*, ed. by David P. Farrington and Joseph Murray, 149–176.

Sherwood, Harriet. 2011. "Palestinian Collaborator: 'I Am a Traitor. I Sold My People. but for What?'" The Guardian. Visited on 06/09/2019. https://www.theguardian.com/world/view-from-jerusalem-with-harriet-sherwood/2011/may/17/israel-palestinian-territories.

Shesterinina, Anastasia. 2016. "Collective Threat Framing and Mobilization in Civil War." *American Political Science Review* 110 (3): 411–427.

Shukla, Anu. 2019. "Son of Jailed British-Iranian: 'My Dad's Never Been Political'." Al Jazeera. Visited on 12/03/2019. https://www.aljazeera.com/news/2019/10/son-jailed-british-iranian-dad-political-191001134233545.html.

Simmel, Georg. 1955. *Conflict and the Web of Group-Affiliations*. New York: Free Press.

— . 1964. *The Sociology of Georg Simmel. Translated, Edited, and with an Introduction by Kurt H. Wolff*. Trans. by Kurt Heinrich Wolff. New York: Collier-Macmillan.

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

Sorek, Tamir. 2010. "The Changing Patterns of Disciplining Palestinian National Memory in Israel." In *Surveillance and Control in Israel/Palestine: Population, Territory and Power*, ed. by Elia Zureik, David Lyon, and Yasmeen Abu-Laban, 113–129. Routledge.

Spiegel. 1961. “Aktion Ochsenkopf.” *Der Spiegel: Politik*.

— . 1993. “Stasi: Wohin Mit Den Spitzeln a. D.?” *Spiegel Online*.

Staniland, Paul. 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, Christoph Valentin. 2022. “The Impact of Domestic Surveillance on Political Imprisonment: Evidence from the German Democratic Republic.” *Journal of Conflict Resolution* 0 (0): 1–28.

Stieglitz, Olaf. 2001. “Sprachen Der Wachsamkeit: Loyalitätskontrolle Und Denunziation in Der DDR Und in Den USA Bis Mitte Der 1950er Jahre.” *Historical Social Research* 26 (2/3): 119–135.

Sullivan, Christopher. 2016a. “Political Repression and the Destruction of Dissident Organizations.” *World Politics* 68 (04): 645–676.

— . 2016b. “Undermining Resistance: Mobilization, Repression, and the Enforcement of Political Order.” *Journal of Conflict Resolution* 60 (7): 1163–1190.

Sullivan, Christopher, and Christian Davenport. 2018. “Resistance Is Mobile: Dynamics of Repression, Challenger Adaptation, and Surveillance in US ’Red Squad’ and Black Nationalist Archives.” *Journal of Peace Research* 55 (2): 175–189.

— . 2017. “The Rebel Alliance Strikes Back: Understanding the Politics of Backlash Mobilization.” *Mobilization: An International Quarterly* 22 (1): 39–56.

Sutter, Matthias, and Christina Strassmair. 2009. “Communication, Cooperation and Collusion in Team Tournaments—an Experimental Study.” *Games and Economic Behavior* 66 (1): 506–525.

Svolik, Milan W. 2012. *The Politics of Authoritarian Rule*. Cambridge University Press.

Sykes, Gresham M., and David Matza. 1957. “Techniques of Neutralization: A Theory of Delinquency.” *American Sociological Review* 22 (6): 664.

Tajfel, Henri, and John C. Turner. 1986. “The Social Identity Theory of Intergroup Behavior.” In *Psychology of Intergroups Relations*, ed. by Stephen Worchel and William G. Austin, 7–24. Chicago: Nelson-Hall.

Tange, O. 2011. “GNU Parallel: The Command-Line Power Tool,” ;Login: The USENIX Magazine (February): 42–47.

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

Tartir, Alaa. 2019. “Criminalizing Resistance: Security Sector Reform and Palestinian Authoritarianism.” In *Palestine and Rule of Power: Local Dissent vs. International Governance*, ed. by Alaa Tartir and Timothy Seidel, 205–226. Palgrave Macmillan.

— . 2015. “The Evolution and Reform of Palestinian Security Forces 1993–2013.” *Stability: International Journal of Security & Development* 4 (1).

Thiranagama, Sharika, and Tobias Kelly. 2010. “Introduction: Specters of Treason.” In *Traitors: Suspicion, Intimacy, and the Ethics of State-Building*, ed. by Sharika Thiranagama and Tobias Kelly, 1–23. Philadelphia: University of Pennsylvania Press.

Thomson, Henry. 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, Charles. 1978. *From Mobilization to Revolution*. Reading: Addison - Wesley.

— . 2003. *The Politics of Collective Violence*. Cambridge: Cambridge University Press.

Travaglino, Giovanni A., Dominic Abrams, Georgina Randsley de Moura, José M. Marques, and Isabel R. Pinto. 2014. "How Groups React to Disloyalty in the Context of Intergroup Competition: Evaluations of Group Deserters and Defectors." *Journal of Experimental Social Psychology* 54:178–187.

Tullock, Gordon. 1980. "Efficient Rent Seeking." In *Towards A Theory of Rent-Seeking in Society*, ed. by James M Buchanan, Robert D. Tollison, and Gordon Tullock, 97–112. College Station: Texas AM University Press.

Tyler, Tom R., and Yuen Huo. 2002. *Trust in the Law: Encouraging Public Cooperation with the Police and Courts*. New York: Russell Sage Foundation.

Van Rossum, Guido, and Fred L Drake Jr. 1995. *Python Reference Manual*. Centrum voor Wiskunde en Informatica Amsterdam.

VERBI Software. 2020. "MAXQDA 2020 Online Manual." Visited on 08/28/2020. https://maxqda.com.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. "Scipy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17:261–272.

Waring, Elin, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu, and Shannon Ellis. 2022. *Skimr: Compact and Flexible Summaries of Data*.

Waskom, Michael L. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6 (60): 3021.

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

Weinstein, Jeremy M. 2007. *Inside Rebellion: The Politics of Insurgent Violence*. New York: Cambridge University Press.

Wes McKinney. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, ed. by Stéfan van der Walt and Jarrod Millman, 56–61.

Wickham-Crowley, Timothy P. 1987. "The Rise (And Sometimes Fall) of Guerrilla Governments in Latin America." *Sociological Forum* 2 (3): 473–499.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686.

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

Wohlrab-Sahr, Monika, Thomas Schmidt-Lux, and Uta Karstein. 2008. "Secularization as Conflict." *Social Compass* 55 (2): 127–139.

Wood, Elisabeth Jean. 2003. *Insurgent Collective Action and Civil War in El Salvador*. New York: Cambridge University Press.

Wu, Huizhong, and Dake Kang. 2022. "Uyghur County in China Has Highest Prison Rate in the World." AP News. Visited on 05/19/2022. https://apnews.com/article/china-prisons-uyghurs-religion-0dd1a31f9be29d32c584543af4698955.

Xinhua. 2019. "Hamas Detains 45 Palestinians Suspected of Collaborating with Israel." Visited on 06/07/2019. http://www.xinhuanet.com/english/2019-01/09/c_137729353.htm.

Yousef, Mosab Hassan. 2010. *Son of Hamas: A Gripping Account of Terror, Betrayal, Political Intrigue, and Unthinkable Choices*. Tyndale House Publishers.

Zdaniuk, Bozena, and John M. Levine. 2001. "Group Loyalty: Impact of Members' Identification and Contributions." *Journal of Experimental Social Psychology* 37 (6): 502–509.

Zeit. 2019. "Staatssicherheit: Öffentlicher Dienst Wird Bis 2030 Auf Stasi-Tätigkeit Überprüft." Die Zeit. Visited on 12/08/2019. https://www.zeit.de/politik/deutschland/2019-05/stasi-ddr-ueberpruefung-mitarbeit-unterlagengesetz-bundeskabinett.

Zimmermann, Dorothe. 2015. "Praktiken Der Denunziation in Der Schweiz: Der Politische Nachrichtendienst Des Schweizerischen Vaterländischen Verbandes, 1930 Bis 1948." In *Hinter Vorgehaltener Hand: Studien Zur Historischen Denunziationsforschung*, 1st ed., ed. by Anita Krätzner, 51–66. Göttingen: Vandenhoeck & Ruprecht.

Zureik, Elia. 2010. "Colonialism, Surveillance, and Population Control: Israel/Palestine." In *Surveillance and Control in Israel/Palestine: Population, Territory and Power*, ed. by Elia Zureik, David Lyon, and Yasmeen Abu-Laban, 3–46. Routledge.