Table of Contents

Democracy's Chance: The Shaping of Public Opinion Through Transitional Justice *

Egor Bronnikov ® Monika Nalepa

November, 2025

Abstract

This paper offers a conceptualization of transitional justice (TJ) success. It reconstructs the TJ process as an interaction between the government, the opposition and an audience of citizens. While the initial TJ message is considered exogenous, the Government can decide on how much to "promote" it for citizens to accept it. At the same time, the opposition decides to what extent it wants to "de-legitimize" the TJ process. We characterize Citizens by personal preferences for accountability that shape their formation of an opinion of the TJ message in presence of the government's and opposition's TJ promotion and de-legitimation activities. The combination of the citizens' preferences and the elite activities described above results in acceptance or rejection of the TJ. We then formalize the conditions under which transitional justice is accepted in two broad public-opinion contexts: a unified society and a polarized one.

Keywords: Transitional justice, democratization, formal theory, dynamic games.
Word count: 10110.

*The ® symbol indicates that the authors are in random order. We are grateful to Kristof Bosmans, Scott Gehlbach, Konstantin Sonin, Elias Tsakas, Alexander Vostroknutov, Xingyu Zhou, Kelebogile Zvobgo for their insightful comments and valuable suggestions, as well as to all participants of the 82nd MPSA Annual Conference at the University of Chicago. Bronnikov: School of Business and Economics, Maastricht University; School of Science, Brīvā Universitāte; egor.bronnikov@maastrichtuniversity.nl. Nalepa: Department of Political Science, University of Chicago; mnalepa@uchicago.edu.

1 Introduction

When democratic polities deliberate over an issues of morality, often some people take a firm stance on one side of the debate while others staunchly oppose it. Some scholars believe that when such polarizing arguments occur among elites they sow divisions among the general public (Iyengar et al., 2019). Other recent empirical research in the United States has failed to see such elite divisions translating into polarization of mass publics (e.g., Orr et al., 2023; Broockman et al., 2023; Westwood et al., 2022).

We show that the adversarial process of promoting one morally controversial policy over another can shape public opinion on morally contentious issues. To put a more substantive interpretation on our contribution, we focus on Transitional Justice (TJ) processes. According to the classical definition of TJ offered by Jon Elster (2004) TJ is a process of dealing with former members and collaborators of an authoritarian regime that is implemented by a national level government of accepted democratic legitimacy. Policies surrounding TJ are among the most morally challenging decisions polities recovering from political violence face. When perpetrators of human rights violations are held accountable or even when the acts they engaged in are disclosed, disagreements emerge whether such acts were justified, whether they were reactions to violence committed by the other side, and whether those directly responsible for the violence were "just obeying orders." Reckoning with past atrocities and responsibility for them comprises the TJ process.

In the last few decades, the world has witnessed the emergence of a TJ norm following regime transitions. A key way through which states transitioning out of an authoritarian regime to a democracy signal to the international community their commitment to democratic institutions, rule of law, and respect for human rights is by holding to account those who committed war crimes, persecuted members of the political opposition and violated human rights of the state's citizens (e.g., Sikkink, 2012; Loyle, 2024). Such procedures encompass criminal trials of perpetrators of abuses against civilians and are conducted regardless of whether the offenders were heads of state or were following

orders from superiors. They have taken place at least since the end of World War II and the setting up of the International Military Tribunal in Nuremberg for prosecuting Nazi crimes. Although Nuremberg is the most famous instance of TJ, TJ encompasses a range of extra-judicial procedures as well, such as lustrations (e.g., Ang and Nalepa, 2019; Aidt et al., 2022), purges (e.g., Nalepa, 2022), and truth commissions (e.g., Zvobgo, 2020). The specific goals of these institutions vary, but are widely believed to alleviate tensions associated with the legacies of violence.

In numerous countries, however, TJ outcomes have sparked controversies (Gibson, 2004), at times even outright backlash (Bates, 2024). James L. Gibson argues that such controversy is an inevitable part of a fair and proportional distribution of blame. In Gibson's view, a truth commission has accomplished its goals when everyone affected by it is equally dissatisfied (Gibson, 2004). This is a sign that the commission's report is even-handed in assigning responsibility for crimes perpetrated by the authoritarian regime or during the civil war. Gibson's remarks came in the context of the Truth and Reconciliation Commission report on South Africa. The report sent ripple effects through South African society as it exposed not only crimes committed by the Apartheid regime but also some gruesome acts of terror committed by the African National Congress (ANC). In his 2004 book, Gibson argued that the truth process facilitated reconciliation by apportioning blame to all sides in the struggle over apartheid (Gibson, 2006).

In contrast to Gibson's portrayal of South Africa, Cyanne Loyle documents how governments in Rwanda and Uganda selectively exposed the truth about human rights violations committed during the civil war in a way that left some victims pleased with the outcomes of the TJ process while leaving others completely disappointed and disillusioned. This suggests that rather than "the more, the better" TJ, particularly in its extrajudicial form, is rather an instance of distributive justice, where the bad that it distributed is blame or culpability (Loyle, 2024).

In this paper we offer a conceptualization of the TJ process to be "successful" when public opinion converges on one interpretation of the past and uncover conditions under

which such success occurs. In order to develop a universal criterion for TJ success we formalize the intuition used in Gibson and Loyle: the broad and equal acceptance of the TJ outcome by citizens. We will show how this acceptance emerges following a promotion campaign led by the government and criticism from the opposition.

We first introduce this conceptualization of TJ success and next, we ask when can such broad acceptance be achieved. We propose a mechanism that illustrates how how a plurality of moral view points can translate into the ultimate acceptance of TJ outcomes that are subject to (government's) promotion or (opposition's) resistance. We also show that a diversity of these moral viewpoints in fact helps in the ultimate acceptance of TJ outcomes. In addition, we use rich empirical material in the form of numerous examples of TJ processes that have been used across the globe in the last five decades to motivate our assumptions about actors, payoffs and beliefs.

2 Assumptions about the TJ Process

Our analysis focuses on the period following the conclusion of a truth commission's work, during which the commission has issued its final report, but its findings—due to the report's volume and complexity—remain largely inaccessible to the general public in their original form. During this stage, citizens hold prior beliefs about the the fairness of the TJ outcome, but must rely on competing elite narratives to interpret it. The government can choose to promote the findings to enhance public acceptance, while the opposition may attempt to delegitimize them. These opposing actions generate a public signal, which citizens observe and use to update their beliefs about the fairness of the TJ outcome. Based on these updated beliefs, citizens then decide whether to accept or reject the TJ process.

Throughout the paper, we will refer to this stage as the *TJ process*. This reflects our view that TJ is not a seamless or uncontested episode, but a sequence of phases: truth documentation, public interpretation, and societal reception. By focusing on the often-overlooked period between the issuance of the final report and the consolidation of

public opinion, we underscore that the TJ process is politically mediated and contingent, not self-executing. Recognizing these steps is essential for understanding when and why TJ efforts succeed or provoke backlash.

Government's promotion vs. opposition's resistance. In transitional settings, governments promote truth commission findings to signal democratic legitimacy, foster national unity, and align with international human rights norms; this promotion is a strategic and costly effort aimed at shaping public belief in the fairness of the TJ outcome, and its success depends on both message credibility and public receptiveness. At the same time, opposition actors resist the truth commission's findings to protect their political future, discredit the process as biased or illegitimate, and mobilize disaffected constituencies; modeled as a costly strategic choice, resistance aims to erode public belief in the fairness of the TJ outcome and to undermine the government's legitimacy by shaping citizen perceptions through elite signaling.

TJ outcome. Throughout the paper, we use the term *TJ outcome* to refer not to the content of the truth commission's report itself, but to the extent to which that report is accepted by the public. Specifically, the TJ outcome reflects the aggregate level of societal acceptance of the commission's findings—whether citizens regard them as legitimate, fair, and deserving of political and moral endorsement. In our model, this outcome is operationalized through individual citizen actions: each citizen chooses a level of acceptance based on their belief about the fairness of the report. Importantly, since citizens do not observe the full report directly, their beliefs—and eventual decisions to accept or reject the TJ outcome—are shaped by signals from political elites: government's promotion and the opposition's resistance. The TJ outcome, in this framework, is therefore not an automatic result of the truth commission's work, but the consequence of strategic elite framing and citizens' independent belief updating.

2.1 Assumptions

Below we list and justify several crucial assumptions of our model.

States of the world. The underlying state of the world is characterized by a single parameter representing the unbiasednesas—or *evenhandedness*—of the truth commission's findings. Evenhandedness is central to the legitimacy of TJ. The core objective of truth commissions is not merely to produce a record of past abuses, but to do so in a balanced manner acceptable to a broad citizenry. This means assigning blame proportionally to responsibility, acknowledging suffering across societal divisions, and avoiding selective memory. James Gibson's research Gibson (2004) on the South African Truth and Reconciliation Commission demonstrates that citizens accepted the TJ outcome not because they agreed with its every detail but because they perceived the commission as treating all sides fairly. In short, θ captures whether the public perceives the TJ process as credible.

Second, among several essential factors that structure truth commission operations, evenhandedness is one of the factors in which they vary to the greatest extent. While most modern truth commissions share procedural features—such as collecting testimonies, establishing timelines, and recommending reparative or punitive measures—they diverge sharply in how equitably they treat different groups. The South African TRC, for instance, named both Apartheid security forces and ANC actors as perpetrators, whereas the Rwandan and Ugandan cases emphasized one-sided narratives Loyle (2024). These discrepancies in fairness shape how TCs are perceived and whether they are embraced or rejected by the public. In this light, evenhandedness emerges as the principal axis of variation in TC practice and outcome reception: validating θ as a faithful representation of this crucial underlying dimension.

Finally, defining θ as the evenhandedness of the truth commission outcome provides both analytical tractability and empirical falsifiability, enabling empirical researchers to estimate its value (in)directly. This combination of formal clarity and empirical accessi-

bility makes evenhandedness a reasonable conceptualization of θ.

Independence of truth commissions. We assume that the truth commission operates independently of political elites. From the perspective of players, the report is exogenous in that is not strategically shaped by the government or opposition.

First, from the stance of empirical evidence, there are numerous instances in which truth commissions operated at arm's length from political actors, offering credible evidence that their outcomes were not directly shaped by government or opposition interference. Prominent examples include Argentina's CONADEP, Peru's CVR, and, most notably, South Africa's Truth and Reconciliation Commission (TRC). The latter produced an independent final report that not only detailed the systemic injustices committed under the Apartheid regime, but also documented specific crimes and the institutional complicity of various state and non-state actors. Crucially, it included a frank accounting of human rights violations committed by members of the African National Congress (ANC), such as the brutal treatment of suspected informers within its ranks (e.g., Mussi and Mussi, 2020; Hoeres and Knabe, 2023). In this context, the TRC's findings, testimonies, and recommendations are best understood as exogenous to the government—not because the process was apolitical, but because the structure and mandate of the TRC ensured that its outputs were grounded in evidence rather than elite manipulation.

Second, from a normative perspective, in democratic and democratizing regimes, it is widely presumed that truth commissions act with a significant degree of independence from political power. Embedding this expectation in the model by treating the TJ outcome as exogenous aligns the theoretical framework with these democratic ideals and reflects the way TJ is legitimated in practice. This assumption is consistent with a vision of TJ adopted within functioning or aspiring democracies—where the outcomes of truth commissions are not dictated by whoever happens to be in power at the moment. The legitimacy of the South African TRC, led by Archbishop Desmond Tutu, rested on its independence from the ANC-led government.

Third, by holding the TJ outcome fixed and independent, the model cleanly separates

two distinct stages of the TJ process: truth production (by the commission) and truth framing (by political elites). This conceptual separation mirrors real-world dynamics, in which truth commissions produce long and complex reports that are later filtered, promoted, or challenged by political actors in the public sphere.

No Direct Access Assumption. Finally, we assume that citizens cannot directly access the truth commission's report, but, due to the report's complexity calling for specialized knowledge required to assess its fairness, must rely on elite interpretations. Truth commission reports often result from years of investigation, thousands of testimonies, documents, and forensic findings. The Peruvian Truth and Reconciliation Commission's report, for instance, spanned more than 8,000 pages, while Argentina's "Nunca Más" report was over 50,000 pages long. Even executive summaries of such reports frequently run into the hundreds of pages and rarely contain a single, easily digestible conclusion that clearly attributes proportional blame or offers a definitive assessment of fairness. The language is typically legalistic or technocratic, making it difficult for untrained readers to interpret.

Moreover, these reports are highly mediated products. Producing a report involves multiple layers of filtering into a carefully curated version of the truth. Commissioners select and then synthesize evidence. For the average citizen, reading the report provides access not to primary truth but to a of it. Without transparency into the underlying selection criteria readers cannot independently assess whether certain groups were privileged in the narrative, whether key events were discounted, or whether interpretive bias affected how the history of abuse was framed. Finally, understanding whether the report portrays the actors and events of the transitional period accurately requires deep contextual knowledge of the country's history, including timelines of conflict, political cleavages, and the roles of specific institutions or factions. In this context, it is reasonable to assume that very few citizens will engage directly with the full report in forming their beliefs about its fairness. Given this, citizens are likely to rely on interpretations of elite-generated signals—governmental promotion or oppositional resistance—to inform

their beliefs about whether the outcome is legitimate. The model in the paper builds on this insight.

3 Model

During transitional periods following authoritarian rule, new democratic governments seek citizen approval of the TJ outcome for any number of reasons. Gaining public endorsement for TJ outcomes allows the government to legitimize its authority domestically Elster (2004); Horne (2017) and internationally Sikkink (2012); Subotić (2011).

Conversely, the opposition will seek citizen disapproval of TJ policies to undermine the legitimacy of the government. For instance if the opposition to the democratic government is consolidated around elites who are successors to the former authoritarian government, disparaging TJ may help such an opposition shield its members from accountability. Fostering disapproval for the TJ narrative may also allow it to maintain ideological relevance and mobilize support.[1]

3.1 States of the world

More formally, let us start by assuming that the state of the world $\theta \in \Theta = [0,1]$ is the fundamental TJ outcome parameter that captures the *fairness* or *even-handedness* of truth commissions. A lower θ indicates that the TJ process was discriminatory by disproportionately targeting specific groups. Conversely, a high(er) θ denotes a fair and even-handed TJ process, in which the TJ mechanisms avoided favoritism or bias. This general formulation of the even-handedness of the TJ process aligns with recent literature on TJ (e.g., Gibson, 2004, 2009; Loyle, 2024). When θ is low, the TJ mechanisms might disproportionately hold certain groups accountable while neglecting others. This will likely lead to societal perceptions that the process is unjust, deepening existing polarization, exacerbating grievances, and deepening interpersonal and institutional trust.

[1]Despite the fact that the opposition may create divisions that hinder the government's ability to promote reconciliation, the opposition need not be anti-democratic in nature.

However, when the values of θ are high, the TJ mechanisms may transparently investigate and acknowledge wrongdoing from all relevant parties, apply justice consistently, and offer balanced consideration of historical events. A high θ suggests a TJ process that strengthens societal trust by demonstrating impartiality and equity in assigning responsibility, which results in sustainable institutional trust (e.g., Horne, 2014, 2017; Fiedler and Mross, 2023).

Finally, we assume that the government and the opposition might have private information about the real TJ outcome θ, while citizens form prior beliefs about θ which are distributed according to a density f with full support on $\Theta = [0, 1]$, reflecting as we discussed before, the assumption that citizens cannot discover the TJ outcome directly, but only in response to the typical scenario, where the government promotes the TJ outcome and the opposition will seek to de-legitimize it.

Previewing our game, it proceeds as follows: First, Nature chooses the true state of the world $\theta \in [0, 1]$. Next, citizens form their prior beliefs π_i^0 simultaneously. This is followed by the government choosing the level of promotion p, and the opposition choosing the level of resistance r simultaneously and setting the signal $s = \theta + p - r$. After observing the signal s, citizens simultaneously update their beliefs about the TJ outcome to posteriors π_i^s. Finally, in the last stage of the game, citizens simultaneously choose whether to accept the TJ outcome, $a_i \in [0, 1]$.

3.2 Players

The Government can engage in education, memorialization, and media campaigns (or even propaganda) to make the TJ outcome appear more favorable or fair in the eyes of the public. This could involve highlighting the benefits or successes of the TJ policies or reinterpreting past events in a manner that casts the outcome in a better light. We model this as an action p taken by the government that set the promotion level of the TJ outcome.

The decision to promote the TJ outcome (i.e., the findings and recommendations

detailed in a truth commission's report) is endogenous and therefore entirely at the discretion of the government. Promoting a TJ outcome could include public endorsement of the report, incorporating its recommendations into policy, ensuring reparations were delivered, and fostering societal dialogue based on its conclusions. Conversely, the government could choose to downplay or selectively implement the report's recommendations. Consider as an example the South African TRC again. After the TRC completed its work. The South African government widely publicized the TRC's work. DEspite this, debates continue over whether the government adequately implemented its recommendations, particularly those regarding reparations and institutional reforms (e.g., Stanley, 2001; Hoeres and Knabe, 2023).

But a government can also refuse its endorsement of a TJ outcome. The response of the Peruvian government to the report of the CVR set up in 2001 to investigate the State, the Shining Path and the Tupac Amaru Revolutionary Movement between 1980 and 2000 is such an example. According to the CVR's final report the principal perpetrator responsible for more that 60 thousand deaths was indeed the Shining Path (a Maoist guerrilla movement), but the CRV attributed to them "only" about 54 percent of the deaths. Meanwhile, the State, military, police, other security forces, political parties, and each of the governments in power during the conflict, most notably the administration of Alberto Fujimori was deemed responsible for the remaining deaths. The report assigned responsibility for the failure to control the excesses of the armed forces to the three presidents, Fernando Belaúnde Terry, Alan García Pérez, and Alberto Fujimori.[2]

The level of promotion, p, in our model can be interpreted as the government's engagement in promoting the TJ outcome: a higher p means the government invests more resources or efforts to convince citizens to accept the TJ outcome. The government aims to maximize the number of citizens accepting the TJ outcome and so it's utility is given

[2]Later in 2009, their successor in the Presidency, Mr Garcia still refused to fund the creation of a museum commemorating victims of the Dirty War, which was a recommendation of the CRV.

by

$$u_G = \int_0^1 a_i \, di - c_G(p) \tag{1}$$

where $c_G(p)$ is the government's cost function dependent on the level of promotion p. The interpretation of eq. (1) is straightforward: this is simply the fraction of citizens who accept the TJ outcome minus the cost of promoting this outcome.

An example able to illustrate that promotion activity is different from the exogenous TJ outcome itself is the Argentinean National Commission on the Disappearance of Persons established in 1983 by President Raul Alfonsin (*Comision Nacional sobre la Desaparicion de Personas*, CONADEP), tasked with investigating the human rights abuses committed during Argentina's "Dirty War" (1976–1983). The dictatorship had orchestrated systematic forced disappearances, torture, and extrajudicial killings, targeting political dissidents and perceived enemies of the state (e.g., Crenzel, 2012). After nine months of collecting testimonies from survivors and families of the disappeared, CONADEP documented 8,961 cases of forced disappearances. Yet the operations of CONADEP, including its investigations, collection of testimonies, and the preparation of the *Nunca Mas* report, were largely independent and aimed at creating an objective record of the dictatorship's crimes. After CONADEP delivered its report, the government had an opportunity to promote the results of the report or to silence or ignore them. President Alfonsin initially supported the report's findings and promoted its recommendations, leading to the prosecution of key military leaders in the Trial of the Juntas. However, subsequent governments faced challenges, including military pushback, which led to laws like the Full Stop Law (1986) and the Law of Due Obedience (1987), effectively limiting prosecutions (e.g., Crawford, 1990).[3]

The potential for the TJ outcomes to invite criticism and elicit efforts to de-legitimize the process brings up the opposition. The Opposition is the second player in our game,

[3]These laws were repealed in the early 2000s, enabling the resumption of trials (e.g., Humphrey and Valverde, 2008).

interpreted as the conservative or the *ancien régime* elite who acts as the restrainer (e.g., Grillo et al., 2023) of the government. The opposition takes an action r that represents the level of resistance or counter-promotion to the TJ outcome. The opposition typically resists change to preserve the status quo, maintain power, and avoid accountability for past actions. It may attempt to delegitimize truth commissions, tribunals, or reparations processes by labeling them as politically biased. They may try to block or weaken legal reforms, engage with misinformation and mobilize public sentiment against the TJ outcome.

For instance, after South Africa established the TRC to address human rights violations committed between 1960 and 1994 (e.g., Wilson, 2001), some groups aligned with the former apartheid regime, particularly segments of the Afrikaner community argued that the commission disproportionately focused on the crimes of the apartheid government while downplaying abuses committed by liberation movements (e.g., Stanley, 2001).

The opposition may also resist efforts to reform judicial systems or pass new laws that could lead to prosecutions of former officials or elites tied to the old regime. or by leveraging its influence in political institutions, it can obstruct the legal foundation of TJ. Consider the example of Chile, where an 1978 Amnesty Law granted immunity to individuals involved in crimes between 1973 and 1978. Efforts to repeal or challenge this law encountered resistance from military and political figures loyal to the Pinochet regime.[4]

The opposition's level of resistance is captured in the model by the parameter r. A higher r represents more efforts devoted to convincing citizens that $\theta = 0$, which ultimately reduce the acceptance of the TJ outcome rate by the citizens. Since the opposition maximizes the number of citizens rejecting the TJ outcome its utility is given by

$$u_R = \int_0^1 (1 - a_i)\, di - c_R(r) \tag{2}$$

[4]See, for example, "Chile: Amnesty law keeps Pinochet's legacy alive", *Amnesty International*, September 11, 2015.

where $c_R(r)$ is the opposition's cost function dependent on the level of resistance r. The interpretation of eq. (2) is straightforward: this is simply the faction of citizens who reject the TJ outcome, net the cost of resisting the TJ outcome.

Furthermore, the *ancien régime* elite or conservative forces may cast TJ efforts in a retributive light. In post-Soviet Russia (particularly under Vladimir Putin), for example, certain elites trained in the former Soviet period downplayed Stalinist repression, opposing accountability for Soviet-era crimes. This rehabilitation of Stalin's image and the suppression of organizations that document Soviet-era atrocities has come to be known as re-Stalinization.

Finally, the (pro-)*ancien régime* opposition may appeal to public fears or concerns, painting TJ as a foreign-imposed process that undermines national sovereignty, arguing that investigating past crimes is unpatriotic to the point of betraying one's nation. For instance, the International Criminal Tribunal for the Former Yugoslavia (ICTY) drew criticisms for disproportionately targeting Serbs while ignoring crimes by other groups. In Republika Srpska, denial campaigns led many Serbs to see TJ as punitive rather than reconciliatory Subotić (2011).

We define a signal provides citizens with information about the true state, θ, how close to even-handed the TJ outcome is:

$$s = \theta + p - r \tag{3}$$

Citizens, not knowing the true state, θ, have only prior beliefs about how fair the TJ outcome is, $\pi_i^0 \in [0,1]$, but after observing the signal s, they can (simultaneously) update their beliefs.

Finally, we turn to the Citizens. Indexed by i, Citizens form a continuum with mass one: $i \in I = [0,1]$. Every citizen i simultaneously takes an action $a_i \in A = [0,1]$, where $a_i = 0$ represents entirely rejecting and $a_i = 1$ unequivocally accepting the TJ outcome. Examples of rejecting a TJ outcome include expressing distrust in the results of truth commissions or rejecting the legitimacy of court decisions on war crimes. Accepting a

TJ outcome could mean agreeing with reparations or supporting verdicts from trials related to human rights abuses, or actively participating in truth commissions. Finally the acts of rejecting or accepting a TJ outcome could manifest themselves in voting for that reject or support TJ policies, respectively.

One illustration of citizens' acceptance of the TJ outcome (i.e., the case when $\int_0^1 a_i di \to 1$) is the Argentinean human rights organization, Mothers and Grandmothers of Plaza de Mayo, which led marches, rallies, and campaigns to demand accountability of the atrocities committed during Argentina's dictatorship (e.g., González Ocantos, 2014), especially following the annulment of amnesty laws in the early 2000s. citizens embraced TJ outcomes by endorsing museum exhibitions, supporting memorials like *Espacio Memoria y Derechos Humanos*.

When choosing an action, citizens may be motivated by forward-looking, reconciliation-oriented goals—such as restoring peace—or backward-looking justice and insist on full accountability, even at the risk of political or social tension. To formalize this, we interpret $\kappa_i \in [0,1]$ as citizen i's preference for accountability versus reconciliation—that is, their individual demand for accountability. High κ_i indicates a strong demand for a TJ process that is unflinching in its pursuit of truth, punishment, and moral reckoning. A low κ_i, by contrast, reflects a prioritization of reconciliation and stability, even if that means accepting symbolic or partial justice. This parameter enters the citizen's decision rule alongside their belief about the actual fairness of the TJ outcome, captured by θ, which we interpret as the degree of evenhandedness in the TJ process.

We will also use a weight $\gamma_i \in [0,1]$ to represent the extent to which citizen i cares about being correct in estimating θ over realizing his actions in accordance with his preferences for the demand for accountability κ_i. In this way, the model captures a core societal tension: citizens may agree on the facts, but disagree on what justice should achieve—fairness across sides versus healing across time. Importantly, citizens with $\gamma_i < 1$ are not necessarily ideologues or partisan; rather, they are individuals who place non-negligible weight on what they believe the TJ process *ought* to have achieved.

That is, they evaluate the outcome not only based on updated beliefs about evenhandedness, but also based on whether the process aligns with their values (accountability or reconciliation). In this sense, γ_i allows the model to represent variation in the degree to which citizens act on principle versus on evidence.

To illustrate how the combination of γ_i and κ_i work consider the South African TRC, which was explicitly designed to balance competing demands: on the one hand moral reckoning with perpetrators of apartheid violence and on the other, the desire for peace and national unity. A citizen with $\gamma_i = 0$ and a high κ_i (strong demand for accountability) would supported the TRC only if they believed it would lead to genuine accountability. If these individuals believed the TRC merely encouraged symbolic acts or allowed perpetrators to evade justice via amnesty, they would reject it, even if it promoted peace. By contrast, a citizen with $\gamma_i = 0$ and a low κ_i (preference for reconciliation) would evaluate the TRC primarily in terms of whether it promoted peace. For such a person, even if perpetrators were exonerated, the process could still be judged as "fair" or desirable—provided it successfully prevented further violence and enabled mutual recognition of harm. In both cases, citizens update and act based on their belief about what the TJ process has achieved ($\mathbb{E}_i[\theta|s]$)—whether it aligned with their ideal of justice as accountability or justice as reconciliation.

Formally, citizen's i utility is a function of his chosen action a_i, his belief about the true state of the world $\mathbb{E}_i[\theta|s]$, his preferences over his demand for accountability κ_i, and the weight $\gamma_i \in [0,1]$:

$$u_i(\cdot) = -\left[\gamma_i (a_i - \mathbb{E}_i[\theta|s])^2 + (1-\gamma_i)(a_i - \kappa_i)^2\right] \tag{4}$$

The first term (weighted by γ_i) in eq. (4) captures citizen's i aspiration to match his actions in the best possible way to the state of the world. One may think of this as *the precision motive*. The second (weighted by $1-\gamma_i$) term models the extent that the citizen's action matches his principled beliefs about what the TJ outcome *should* be. We will call this the *moral motive*. We can think of γ_i as the extent to which i action is correct

regardless of her moral aspirations. That is if $\gamma_i = 1$, citizen i takes into consideration only the suitability of his actions and has no strict preference or normative views on what the TJ policies should have been. On the contrary, if $\gamma_i = 0$, the only concern the citizen i has is the alignment of his chosen actions with his ideal view of the realization of the TJ outcome (we provide an illustrations of this dynamic in the section below).We do not put restrictions on the proportion of citizens with high or low γ.

We are not the first to incorporate such non-pure-consequentialist concerns (or restrictions) into a utility function. For instance Kuran (1998) explored multiple components of utilities, including what he calls "expressive utility" to depict how individuals can gain value from sincere manifestation of their opinions even when those opinions do not lead to changes in states of the world. [5] Beyond this classic example, other research incorporating concern of matching one's actions to moral principles include (e.g., Dewan and Myatt, 2008; Edmond, 2013; Little, 2017).

To summarize, timing of the game is summarized below:

- at $t = 0$, Nature selects the true state of the world, θ, which is unknown to the citizens;[6]
- at $t = 1$, citizens form prior beliefs π_i^0 about θ;
- at $t = 2$, the government strategically sets the level of promotion p, while the opposition simultaneously sets the level of resistance r, influencing the observable signal $s = \theta + p - r$;
- at $t = 3$, citizens observe the signal s and update their beliefs using Bayes' rule, forming their posterior beliefs π_i^s;
- at $t = 4$, given (updated) beliefs, citizens select an action a_i.

[5] Cf. an elaborate review of the influential *Private Truths, Public Lies: The Social Consequences of Preference Falsification* by Frank (1996).

[6] Whether or not the government and opposition observe θ is immaterial because the the TJ outcome is not part of their preferences

4 Analysis

The solution concept we use is Perfect Bayesian Equilibrium (PBE), which is appropriate for a dynamic game unfolding in a setting of imperfect information, where players form and update beliefs based on observed signals.

The PBE ensures that: (i) players' strategies are sequentially rational (the government and opposition choose p and r optimally, anticipating citizens' responses, and citizens choose a_i optimally given their updated beliefs); (ii) beliefs are consistent with Bayes' rule (citizens interpret s in a rational way, incorporating their prior beliefs and the strategic behavior of the government and opposition); and (iii) no player has an incentive to deviate (given that beliefs are updated correctly and strategies are best responses, no single player can improve their outcome by changing their action).

Before turning to solve this game, we note that citizens' optimal action depends on their posterior beliefs. Hence, we begin by analyzing belief formation and updating. First, we assume citizens' beliefs follow a normal distribution; later we relax this assumption. Using the citizens' utility function (eq. 4), we derive their optimal action a_i^*. In subsection 4.1 we assume that citizens' prior beliefs are distributed normally. We determine the government's optimal level of promotion, p^*, and the opposition's optimal level of resistance, r^* (these results are summarized in Proposition 1). In subsection 4.2 we assume that citizens' prior beliefs are distributed bimodally and summarize those results in Proposition 2). Section 5 presents the comparative statics of two cases of prior beliefs distributions.

4.1 Equilibrium with Unimodal Public Opinion

Assume that the prior belief π_i^0 about the real state of the world θ is represented by a normal distribution with mean $\mu_{0,i}$ and variance $\sigma_{0,i}^2$. The signal, denoted by s is expressed as $s = \theta + p - r$, where both p and r follow a normal distribution with means μ_p and μ_r and variances σ_p^2 and σ_r^2. Deriving citizens' posterior beliefs of π_i^s given s using the standard Bayes' rule we obtain a distribution that is also normal with mean

$\lambda_i\left[s-\left(\mu_p-\mu_r\right)\right]+(1-\lambda_i)\mu_{0,i}$ and variance $(1-\lambda_i)\sigma^2_{0,i}$, where $\lambda_i=\frac{\sigma_p^{-2}+\sigma_r^{-2}}{\sigma_p^{-2}+\sigma_r^{-2}+\sigma_{0,i}^{-2}}$. (The formal proposition and the proof are in the Appendix (p. 44)).

We see that the posterior mean—i.e., $\lambda_i\left[s-\left(\mu_p-\mu_r\right)\right]+(1-\lambda_i)\mu_{0,i}$—is a weighted average of the observed signal s and the prior mean $\mu_{0,i}$. The weight λ_i reflects the precision of the signal relative to the prior belief. That is when $\lambda_i=0$, a citizen does not use information coming from the signal s but relies to his initial (prior) beliefs (i.e. his posterior beliefs are identical to priors); when $\lambda_i=1$, a citizen fully relies on information from the signal, s disregarding the prior beliefs completely. More specifically, λ_i increases with the precision of the signal (i.e., as $1/\sigma_p^2+1/\sigma_r^2$ becomes large relative to $1/\sigma^2_{0,i}$) and decreases with the precision of the prior (i.e., as $1/\sigma^2_{0,i}$ becomes large relative to $1/\sigma_p^2+1/\sigma_r^2$).[7]

The posterior variance reflects the updated uncertainty about θ after observing the signal. The posterior variance is reduced from the prior variance $\sigma^2_{0,i}$ based on the precision of the observed signal. The factor $1-\lambda_i$ represents the portion of the prior variance that is retained after incorporating the information from the signal.

We provide elaborations on the dynamics of citizens' public opinion (at the population level) in Appendix E.

Returning to the citizens' problem, we assume that citizens maximize their expected utility, that is, each citizen maximizes eq. (4). This gives the optimal action a_i^*:

$$a_i^*=\gamma_i\mathbb{E}_i\left[\theta|s\right]+(1-\gamma_i)\kappa_i \tag{5}$$

where $\mathbb{E}_i\left[\theta|s\right]$ is a mean of the posterior belief of citizen i.

Their optimal action is the weighted (by moral motive) average of the precision motive (dependent on the expected state of the world) and the moral motive

First let us denote the sum of all actions undertaken by all citizens as the aggregate

[7]It is worth underscoring that because the citizen does not know the true value of θ, the TJ outcome, but only have prior beliefs about its value, those beliefs follow a distribution, hence alowing us to speak about the prior's precision.

acceptance level expressed as

$$\int_0^1 a_i^* \, di = \int_0^1 (\gamma_i \mathbb{E}_i[\theta|s] + (1-\gamma_i)\kappa_i) \, di \tag{6}$$

Assuming the weight γ_i and demand for accountability κ_i are independent, we can re-write eq. (6) as

$$\int_0^1 a_i^* \, di = \mathbb{E}[\gamma_i](\mathbb{E}[\mathbb{E}_i[\theta \mid s]] - \mathbb{E}[\kappa_i]) + \mathbb{E}[\kappa_i]$$

where $\mathbb{E}[\mathbb{E}_i[\theta \mid s]] = \mathbb{E}[\lambda_i] \cdot \big(s - (\mu_p - \mu_r)\big) + (1 - \mathbb{E}[\lambda_i]) \cdot \mathbb{E}[\mu_{0,i}]$.

This means that the aggregate of optimal actions is the difference between the expected TJ outcome (state of the world) ($\mathbb{E}[\mathbb{E}_i[\theta \mid s]]$) and the average societal demand for accountability ($\mathbb{E}[\kappa_i]$), which is weighted by the expected (averaged) importance of the citizen's precision motive ($\mathbb{E}[\gamma_i]$).

Having derived posterior beliefs of citizens conditional on the signal, above we can substitute them into the government's utility given by eq. (1). After assuming that the government's cost function of TJ promotion $c_G(\cdot)$ is quadratic in p (i.e., $c_G(p) = \frac{c_g}{2}p^2$ where c_g is just a constant) we solve for the optimal level of government's promotion (see formal details in Appendix B.1):

$$p^* = \frac{1}{c_g}\mathbb{E}[\lambda_i]\mathbb{E}[\gamma_i] \tag{7}$$

where $\mathbb{E}[\lambda_i] = \mathbb{E}\left[\frac{\sigma_p^{-2}+\sigma_r^{-2}}{\sigma_p^{-2}+\sigma_r^{-2}+\sigma_{0,i}^{-2}}\right]$.

The government's optimal promotion level p^* balances the benefits of influencing public acceptance of the TJ outcome with the costs of promotion. p^* depends on these costs, c_g, citizens' prior beliefs rigidity and the extent of citizens' motivation to match the real state of the world. Recall, that the term $\mathbb{E}[\gamma_i]$ represents the average citizen's motivation to align their actions with the true state θ rather than their preferences for accountability κ_i. For instance, a higher $\mathbb{E}[\gamma_i]$ means that, on average, citizens care

more about the accuracy with which the TJ outcome matches θ than about their personal normative ideals κ_i. When citizens place a high weight on accuracy of matching with θ, the government's promotion efforts p have a greater impact on citizens' acceptance rate, incentivizing a higher promotion level.

The parameter $\mathbb{E}[\lambda_i]$ reflects how much weight citizens on average place on the signal s (influenced by p and r) versus their initial beliefs $\mathbb{E}\left[\mu_{0,i}\right]$: a higher $\mathbb{E}[\lambda_i]$ indicates that citizens on average are highly receptive to new information from the signal s. $\mathbb{E}[\lambda_i]$ increases when initial uncertainty is high and when σ_p^2 and σ_r^2, indicating credible promotion or resistance efforts, are relatively small. In short, the higher $\mathbb{E}[\lambda_i]$ the more influential the government's promotional efforts. Low $\mathbb{E}[\lambda_i]$ indicates citizens rely more on their priors and are less on promotion efforts. This happens if prior variance is low or when the variances σ_p^2 and σ_r^2 are high, diminishing the credibility of p and r.

We proceed analogously to solve the opposition's problem susubstituting into the opposition's utility citizens' posterior beliefs and the signal and assuming the opposition's cost of resisting TJ is quadratic to find the opposition's optimal level of resistance. We relegated these calculations to Appendix B.1.

We substitute into the opposition's utility from eq. (2) citizens' posterior beliefs and the signal. Similarly as in the case of the government, we assume the opposition's cost of resisting TJ $c_R(\cdot)$ is quadratic in r, i.e., $c_R(r) = \frac{c_r}{2} r^2$ where c_r is just a constant (see formal details in Appendix B.1). The optimal level of opposition's resistance is then

$$r^* = \frac{1}{c_r} \mathbb{E}[\lambda_i] \mathbb{E}[\gamma_i] \tag{8}$$

where $\mathbb{E}[\lambda_i] = \mathbb{E}\left[\frac{\sigma_p^{-2}+\sigma_r^{-2}}{\sigma_p^{-2}+\sigma_r^{-2}+\sigma_{0,i}^{-2}}\right]$.

Here too, the optimal level r^* balances the advantages of public acceptance of the TJ outcome with associated costs. The exact balance depends on how costly protesting the TJ outcome is, c_r, the rigidity of citizens' priors how important the precision motive is for the citizens. In this expression too, the parameter $\mathbb{E}[\lambda_i]$ quantifies the extent to

which citizens rely on the signal relative to their initial beliefs. Our first proposition summarizes the equilibrium for the case when citizens beliefs are normal:

Proposition 1. *When citizens' prior beliefs π_i^0 about the TJ outcome θ are distributed* normally*, there exists a unique (Perfect Bayesian) equilibrium, in which the government and opposition choose unique levels of promotion $p^* = \frac{1}{c_g}\mathbb{E}[\lambda_i]\mathbb{E}[\gamma_i]$ and resistance $r^* = \frac{1}{c_r}\mathbb{E}[\lambda_i]\mathbb{E}[\gamma_i]$, the citizens update their beliefs to posterior π^s and choose optimal actions $a_i^* = \gamma_i\mathbb{E}_i[\theta|s] + (1-\gamma_i)\kappa_i$.*

The upshot of this is that when a majority of the public does not arrive at the TJ process with strong convictions about TJ, elites on both sides will invest a lot in interpreting the outcome to align with their side. This equilibrium however, assumes that the the priors of the citizens regarding the actual TJ outcome, θ, are uni-modally (normally) distributed. What if they are not? What if, instead, public opinion is polarized over the outcome of the TJ process? In the next subsection, case when citizens' prior beliefs are bimodally distributed.

4.2 Equilibrium with Polarized Public Opinion

Relaxing the assumption about normally distributed public opinion allows post-conflict or post-authoritarian publics to be polarized. This assumption approximates the actual settings in South Africa, where following creation of the TRC, the agreement over amnesty was understood differently by the Afrikaner Whites than the ANC. The former believed that the TRC would recommend amnesty for all Apartheid crimes and attribute some blame for the conflict to the ANC, whereas supporters of the ANC believed that only Apartheid era crimes would be exposed, with amnesty restricted to rank-and-file apartheid soldiers, who fully revealed the chain of command. The way we will model this situation is as a bimodal distribution of public opinion, as if there were "two public opinions" with two peaks of expectations regarding what the transitional outcome will be. Conceptually, we will be following the steps from the previous section: After solving the game for PBE, we summarize the results of this subsection in Proposition 1.

Assume that the prior belief π_i^0 about the real state of the world θ is captured by a distribution that combines two normal distributions: $\pi_i^0(\theta) = \alpha_i \cdot \mathcal{N}(\mu_{0,i,L}, \sigma^2_{0,i,L}) + (1 - \alpha_i) \cdot \mathcal{N}(\mu_{0,i,R}, \sigma^2_{0,i,R})$, where $0 < \alpha_i < 1$, and the subscripts L and R correspond to the two modes of the prior. Compared to the previous section, the signal here remains the same. After deriving citizens' posterior beliefs using Bayes' rule (see Observation 3 for formal details), we see that $\pi_i^s(\theta)$ is a combination of the two posterior components:

$$\pi_i^s(\theta) = \alpha^s_{L,i} \cdot \mathcal{N}(\mu_{L,i|s}, \sigma^2_{L,i|s}) + \alpha^s_{R,i} \cdot \mathcal{N}(\mu_{R,i|s}, \sigma^2_{R,i|s}) \tag{9}$$

Bimodality captures the simplest case of (Bayesian) updating under polarization. Unlike the case of a single normal prior where citizens converge toward a single updated belief, the population consists of two groups, each updating its beliefs separately. The posterior belief $\pi_i^s(\theta)$ consists of $\alpha^s_{L,i}$–a fraction of the population that updates their belief based on their prior mode, which is $\mu_{0,i,L}$–and $\alpha^s_{R,i}$– a fraction of the population that updates based on their prior mode, which is $\mu_{0,i,R}$). Notably, instead of belief convergence, there is relative belief persistence or even divergence, depending on how much the groups weigh the signal. When the signal is ambiguous, prior beliefs dominate and polarization persists. When the signal is clear and precise, it has the potential to move both groups in the same direction, reducing polarization.

For each citizen, the posterior mean is given by $\mathbb{E}_i[\theta|s]$. This is the mean of the posterior belief described in (9), which is a weighted average of the two group-specific posterior means $\mu_{L,i|s}$ and $\mu_{R,i|s}$. If $\alpha^s_{L,i} \approx \alpha^s_{R,i}$, the posterior remains polarized. If one group updates more significantly than the other, the overall population belief shifts more toward that group's posterior mean. This implies the following. When polarization is strong, belief updating remains asymmetric, meaning different segments of the population react differently to the same signal. If the signal aligns more with one group's expectations, that group updates beliefs more strongly, increasing its influence over the aggregate belief.

The posterior variance is given by $\mathbb{V}_i[\theta|s]$. But the best way to understand developing

polarization is using the term $(\mu_{L,i|s} - \mu_{R,i|s})^2$, which represents the squared difference between the two updated means. If $\mu_{L,i|s}$ and $\mu_{R,i|s}$ are far apart, posterior variance remains large, indicating sustained polarization. This implies if polarization is deep, the distance between $\mu_{L,i|s}$ and $\mu_{R,i|s}$ remains large, leading to greater overall uncertainty in the population. If the signal moves both groups toward a common belief, variance decreases, meaning society becomes more unified in belief.

Appendix E contains further detail on the population-level dynamics of citizens' public opinion.

As before, we assume that the government's cost promoting of TJ $c_G(\cdot)$, is quadratic in p, i.e., $c_G(p) = \frac{c_g}{2}p^2$ where c_g is just a constant (see formal details in Appendix B.2). Proceeding as before, we solve for the optimal level of government's promotion:

$$p^{**} = \frac{1}{c_g}\mathbb{E}\left[\widetilde{\lambda}_i\right]\mathbb{E}[\gamma_i] \tag{10}$$

where $\mathbb{E}\left[\widetilde{\lambda}_i\right] = \mathbb{E}\left[\frac{\alpha^s_{L,i}\cdot\sigma^2_{L,i|s}+\alpha^s_{R,i}\cdot\sigma^2_{R,i|s}}{\sigma_p^2+\sigma_r^2}\right] = \mathbb{E}\left[\frac{\alpha^s_{L,i}\cdot\sigma^{-2}_{0,i,L}+\alpha^s_{R,i}\cdot\sigma^{-2}_{0,i,R}+2(\sigma_p^2+\sigma_r^2)^{-1}}{\sigma_p^2+\sigma_r^2}\right]$.

Note, however, that whereas in the case of the normal belief distribution, the parameter $\mathbb{E}[\lambda_i]$ represented the weight that citizens assigned to new information relative to their prior beliefs (higher precision in the signal (σ_p^2, σ_r^2) increased $\mathbb{E}[\lambda_i]$, leading to stronger belief updating, on average), in the bimodal case, average belief updating is governed by $\mathbb{E}\left[\widetilde{\lambda}_i\right]$. This parameter accounts for two distinct belief clusters, weighting their individual variances and updating patterns. Unlike $\mathbb{E}[\lambda_i]$, which assumed uniform average belief updating, $\mathbb{E}\left[\widetilde{\lambda}_i\right]$ reflects the average weight of polarization, where different groups process the same signal differently.

In the case with a normal distribution of beliefs, $\mathbb{E}[\lambda_i]$ determined how much citizens, on average, adjust their beliefs based on new information. When $\mathbb{E}[\lambda_i]$ is high, both the government and opposition intensify their efforts in shaping public opinion through promotion (p) and resistance (r). In the bimodal belief case, $\mathbb{E}\left[\widetilde{\lambda}_i\right]$ introduces additional complexity. If $\mathbb{E}\left[\widetilde{\lambda}_i\right]$ is high, both belief groups update strongly, but possibly in

opposite directions, reinforcing polarization. If $\mathbb{E}\left[\widetilde{\lambda}_i\right]$ is low, citizens remain entrenched in their priors, making persuasion strategies ineffective. The government and opposition must target different groups separately rather than assuming a uniform response to promotion or resistance. Thus, while $\mathbb{E}[\lambda_i]$ leads to a unified shift in beliefs, $\mathbb{E}\left[\widetilde{\lambda}_i\right]$ can result in belief divergence, complicating strategic interventions.

Assuming that the cost of TJ resistance $c_R(\cdot)$ is quadratic in r, as we have previously (see formal details in Appendix B.2), the optimal level of resistance is:

$$r^{**} = \frac{1}{c_r}\mathbb{E}\left[\widetilde{\lambda}_i\right]\mathbb{E}[\gamma_i] \tag{11}$$

where $\mathbb{E}\left[\widetilde{\lambda}_i\right] = \mathbb{E}\left[\frac{\alpha^s_{L,i}\cdot\sigma^2_{L,i|s}+\alpha^s_{R,i}\cdot\sigma^2_{R,i|s}}{\sigma^2_p+\sigma^2_r}\right] = \mathbb{E}\left[\frac{\alpha^s_{L,i}\cdot\sigma^{-2}_{0,i,L}+\alpha^s_{R,i}\cdot\sigma^{-2}_{0,i,R}+2(\sigma^2_p+\sigma^2_r)^{-1}}{\sigma^2_p+\sigma^2_r}\right]$.

Overall, the equilibrium for the case when citizens beliefs are normal is summarized in proposition 2.

Proposition 2. *When citizens' prior beliefs π^0_i about the TJ outcome θ are distributed* bimodally*, there exists a unique (Perfect Bayesian) equilibrium where the government and opposition choose levels of promotion $p^{**} = \frac{1}{c_g}\mathbb{E}\left[\widetilde{\lambda}_i\right]\mathbb{E}[\gamma_i]$ and resistance $r^{**} = \frac{1}{c_r}\mathbb{E}\left[\widetilde{\lambda}_i\right]\mathbb{E}[\gamma_i]$, and the citizens update their beliefs to posterior π^s_i and choose optimal actions $a^*_i = \gamma_i\mathbb{E}_i[\theta|s]+(1-\gamma_i)\kappa_i$,*

The prior is described in Observation 3 and proved in Appendix C.

When citizens' prior beliefs π^0_i about the TJ outcome θ are distributed *bimodally* important differences emerge compared to the unimodal case. With normally distributed priors, elite signaling effectively shifts public opinion and incentivizes persuasion. Under bimodal priors, instead, entrenched polarization dampens responsiveness, making strategic communication less meaningful and elite investment less worthwhile. Overall, the takeaways from considering citizens' beliefs following a bimodal distribution are that, first, polarization weakens the efficiency of government promotion. Second, in a polarized society, promotional efforts, however strong, may be ignored or interpreted differently by different groups,. Third, in a consensual society, precise and credible gov-

ernment messaging can shift public opinion more effectively. We discuss this in detail in the next section.

5 Comparative Statics

To understand how public opinion responds to elite persuasion and citizen heterogeneity, we compare Citizens' aggregated actions across two belief structures. When citizens' prior beliefs are normally distributed, the aggregated actions take the form:

$$\int_0^1 a_i^* \, di = \mathbb{E}[\gamma_i]\Big(\mathbb{E}[\lambda_i]\big[s - \big(\mu_p - \mu_r\big)\big] + (1 - [\lambda_i])\mathbb{E}\big[\mu_{0,i}\big] - \mathbb{E}[\kappa_i]\Big) + \mathbb{E}[\kappa_i]$$

where $\mathbb{E}[\lambda_i] = \mathbb{E}\left[\frac{\sigma_p^{-2}+\sigma_r^{-2}}{\sigma_p^{-2}+\sigma_r^{-2}+\sigma_{0,i}^{-2}}\right]$ captures the average responsiveness of beliefs to the signal s, depending on the relative precision of the government's promotion, opposition's resistance, and citizens' average priors. In contrast, when prior beliefs are bimodally distributed—reflecting a polarized society—the aggregate action becomes:

$$\int_0^1 a_i^* \, di = \mathbb{E}[\gamma_i]\left(\mathbb{E}\left[\alpha_{L,i}^s \cdot \sigma_{L,i|s}^2 \cdot \frac{\mu_{0,i,L}}{\sigma_{0,i,L}^2} + \alpha_{R,i}^s \cdot \sigma_{R,i|s}^2 \cdot \frac{\mu_{0,i,R}}{\sigma_{0,i,R}^2}\right] + \big(s - (\mu_p - \mu_r)\big)\,\mathbb{E}\big[\widetilde{\lambda}_i\big] - \mathbb{E}[\kappa_i]\right) + \mathbb{E}[\kappa_i]$$

where the posterior mean belief reflects a mixture of belief clusters with respective weights, and the term $\mathbb{E}\big[\widetilde{\lambda}_i\big] = \mathbb{E}\left[\frac{\alpha_{L,i}^s \cdot \sigma_{L,i|s}^2 + \alpha_{R,i}^s \cdot \sigma_{R,i|s}^2}{\sigma_p^2+\sigma_r^2}\right] = \mathbb{E}\left[\frac{\alpha_{L,i}^s \cdot \sigma_{0,i,L}^{-2} + \alpha_{R,i}^s \cdot \sigma_{0,i,R}^{-2} + 2(\sigma_p^2+\sigma_r^2)^{-1}}{\sigma_p^2+\sigma_r^2}\right]$ captures the aggregate sensitivity of these belief groups to the signal s.

Qualitatively, the two expressions share the same structure: citizen actions combine preferences for accountability ($\mathbb{E}[\kappa_i]$) and responsiveness to elite cues, filtered through belief updating. In the unimodal case, updating is smooth and weighted by a single scalar $\mathbb{E}[\lambda_i]$; in the bimodal case, belief polarization leads to heterogeneous responsiveness, where each belief group updates differently and contributes separately to the aggregated actions.

To explore how citizens' aggregated actions respond to perceptions of elite credibility, we differentiate the expression for aggregate action with respect to the perceived

promotional biases of the government. Taking the derivative with respect to μ_p (the government's promotional bias), we obtain $\frac{\partial}{\partial \mu_p} \int_0^1 a_i^* \, di = -\mathbb{E}[\lambda_i] \cdot \mathbb{E}[\gamma_i]$ for normally distributed prior beliefs (unimodal public opinion), and $\frac{\partial}{\partial \mu_p} \int_0^1 a_i^* \, di = -\mathbb{E}\left[\widetilde{\lambda}_i\right] \cdot \mathbb{E}[\gamma_i]$ for bimodally distributed prior beliefs (polarized beliefs).

The negative derivative with respect to μ_p reveals a *backfire* effect in government communication. As citizens expect stronger promotional bias from the government (higher μ_p), they discount the observed signal s more heavily, interpreting it as inflated. This decreases the adjusted posterior belief about the true TJ outcome and, consequently, reduces its aggregate acceptance.

When prior beliefs are normally distributed (i.e., society is relatively ideologically unified), this effect is proportional to $\mathbb{E}[\lambda_i]$, the average weight citizens assign to new information. When beliefs are bimodally distributed (reflecting a polarized public), the analogous effect is governed by $\mathbb{E}\left[\widetilde{\lambda}_i\right]$, which aggregates the responsiveness of ideologically distinct groups to elite messaging. In both cases, the effect is scaled by $\mathbb{E}[\gamma_i]$, which captures how much citizens prioritize epistemic accuracy over principled or ideological alignment.

Turning to the opposition, taking the derivative with respect to μ_r yields $\frac{\partial}{\partial \mu_r} \int_0^1 a_i^* \, di = \mathbb{E}[\lambda_i] \cdot \mathbb{E}[\gamma_i]$ and μ_r yields $\frac{\partial}{\partial \mu_r} \int_0^1 a_i^* \, di = \mathbb{E}\left[\widetilde{\lambda}_i\right] \cdot \mathbb{E}[\gamma_i]$ for unimodally and bimodally distributed prior beliefs, respectively. The positive derivative with respect to μ_r highlights the legitimizing effect of opposition resistance. When citizens anticipate that the opposition is attempting to discredit the TJ process (higher μ_r), but nevertheless observe a favorable signal s, they infer that the underlying truth must be stronger than it appears. Resistance is understood as a suppressive force; therefore, the observed signal is interpreted as "surviving" an adversarial attempt to distort it, which enhances its credibility.

5.1 Comparative Statics on Citizens' Acceptance

When citizens' prior beliefs follow a normal distribution, the extent to which opposition resistance increases the credibility of the TJ process depends on the average weight of

signal or how much people rely on new information compared to their initial moral commitements ($\mathbb{E}[\lambda_i]$).

Figure 1 visualizes how citizens' aggregate support for the TJ outcome responds to changes in their perceptions of government promotion and opposition resistance, under two types of public opinion: normally distributed (unimodal, denoted by N) and bimodally distributed (polarized, denoted by B) priors.

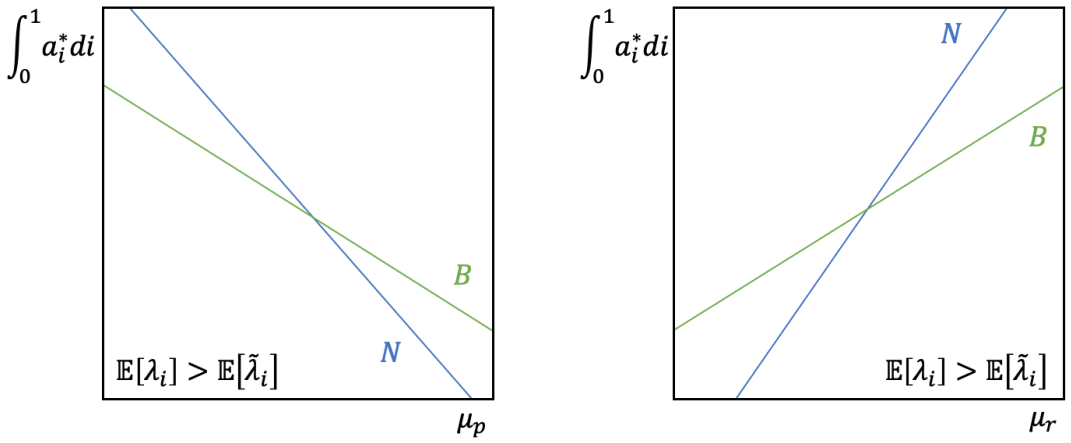

Figure 1: Effect of perceived government promotion (left panel) and opposition resistance (right panel) on citizens' aggregate action. The blue line (N) represents the case of normally distributed prior beliefs (unimodal public opinion), while the green line (B) represents the case of bimodally distributed prior beliefs (polarized public opinion). This figure illustrates the case where $\mathbb{E}[\lambda_i] > \mathbb{E}[\widetilde{\lambda}_i]$; if $\mathbb{E}[\lambda_i] < \mathbb{E}[\widetilde{\lambda}_i]$, the relative steepness of the curves is symmetrically reversed.

The left panel plots how aggregate support changes as a function of how citizens perceive bias in promotional activities of the government, μ_p. Both lines are downward-sloping, reflecting that higher μ_p leads citizens to discount the signal s, thus lowering support for the TJ outcome. However, the blue line (N) is steeper than the green line (B), indicating that when beliefs are unimodal and public opinion converges, citizens are more sensitive to perceived government bias. In contrast, polarization reduces the responsiveness of public opinion by lowering the effective average weight placed on new information, denoted by $\mathbb{E}[\widetilde{\lambda}_i]$; as a result, opinion formation becomes more rigid.

The right panel plots aggregate support against perceived opposition resistance, μ_r. Here, the lines slope upward, since greater resistance is interpreted as an attempt to

suppress a potentially truthful signal. Again, the unimodal case (blue) shows a steeper slope: when society is unified, citizens more sharply increase their support in response to increased opposition resistance. In the polarized case (green), the responsiveness is weaker, consistent with a lower learning rate $\mathbb{E}\left[\widetilde{\lambda}_i\right]$. Taken together, the two panels illustrate the analytical result when $\mathbb{E}[\lambda_i] > \mathbb{E}\left[\widetilde{\lambda}_i\right]$, that is when the average weight assigned by citizens to new information is higher in the unimodal society than in the polarized society. [8]

Overall, the government is expected to affect the signal by inflating the TJ outcome -while the opposition is expected to deflate it. As a result, when citizens perceive the government as excessively biased, its attempts at persuasion lose effectiveness. The signal is discounted as spin, especially in ideologically fragmented societies where prior beliefs already diverge. This leads to declining public support for the TJ outcome as the perceived manipulation overwhelms the informational content of the signal. Interestingly, however, an increase in expected resistance to TJ from the opposition can enhance the credibility of the TJ message. When citizens believe that the signal has survived an adversarial attempt to discredit it, they are more likely to interpret it as truthful. This legitimizing effect is particularly strong when citizens remain open to persuasion and care about accuracy. In sum, while promotion efforts face diminishing returns under perceived bias, resistance can backfire against the opposition, inadvertently reinforcing public belief in the fairness of the TJ outcome.

[8] This is equivalent to the following condition:

$$\mathbb{E}\left[\alpha^{s}_{L,i} \cdot \sigma^{-2}_{0,i,L} + \alpha^{s}_{R,i} \cdot \sigma^{-2}_{0,i,R} + 2\left(\sigma_p^{-2} + \sigma_r^{-2}\right)\right] \cdot \mathbb{E}\left[\sigma_p^{-2} + \sigma_r^{-2} + \sigma^{-2}_{0,i}\right] < 1.$$

Although, the condition might look complicated, the interpretation is rather straightforward: When citizens in the polarized society hold, on average, very rigid beliefs, and the general public is, on average, relatively uncertain, then only a very precise signal (low noise in both promotion and resistance) can make citizens in the unimodal case more responsive than the polarized one. In other words, societies with unimodal belief structures respond, on average, more strongly—whether positively or negatively—to elite bias, while polarized societies exhibit a muted response to the same informational cues.

5.2 Effectiveness of Government Promotion

We now turn to the discussion of a government's optimal level of promotion p^{**} and the opposition's optimal level of resistance r^{**} in a polarized society. In contrast, when beliefs are bimodal (split between ideological camps), the response is shaped by $\mathbb{E}[\widetilde{\lambda}_i]$, which captures the typical updating behavior across these groups. In both cases, the overall impact is stronger when citizens prioritize accuracy over moral commitments (formally, when $\mathbb{E}[\gamma_i]$ is high).

In the case when beliefs are distributed normally (left panel in Figure 2), citizens update their beliefs smoothly in response to government promotion. Government promotion is effective when:

$$\sigma_p^2 < \frac{1}{3}\sigma_r^2 + \frac{1}{3}\mathbb{E}\left[\sigma_{0,i}^2\right]. \tag{12}$$

If σ_p^2 (the variance or consistency) of government's promotion—is too high, citizens discount the message, relying instead on their prior beliefs. As σ_p^2 increases beyond the threshold (given by the right-hand side of eq. 12), p^* decreases sharply, making additional promotion ineffective. That is, in societies where public opinion is unimodal, new information updates the public's perceptions in a continuous way. If government promotion is precise, it significantly shifts opinions. However, if it is noisy, citizens ignore it and stick to their priors.

In the case when beliefs are distributed bimodally (right panel in Figure 2), polarization weakens belief updating because citizens belong to two separate belief clusters (L and R). The effectiveness of government promotion follows a different threshold (given by the right-hand side of eq. 13):

$$\sigma_p^2 > \frac{1}{3}\sigma_r^2, \tag{13}$$

which represents *inefficient* level of promotion in the normal distribution case. Unlike in the normal case, the average variance of prior beliefs, $\mathbb{E}\left[\sigma_{0,i}^2\right]$, is not part of the threshold condition. This suggests that belief in polarized societies dominates initial

uncertainty. If σ_p^2 is too high, not only does government promotion become ineffective, but polarization may deepen as each group interprets the message differently. That is,

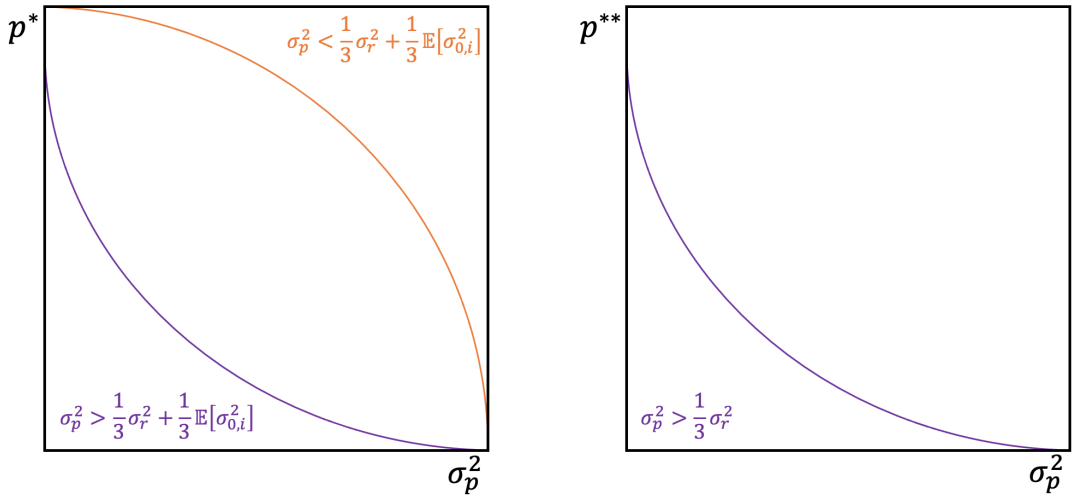

Figure 2: Efficiency of government's promotion: when citizens beliefs are distributed normally (p^*, left panel) and when citizens beliefs are distributed bimodally (p^{**}, right panel).

in a polarized society, the same promotional effort affects different groups differently. If government promotion is too far-reaching, it risks backfiring, reinforcing pre-existing biases rather than shifting opinions. Unlike the normal case, effectiveness depends not just on clarity, but on overcoming entrenched biases.

We discuss the opposition's level of resistance in Appendix B.3.2 as the underlying logic mirrors the case of government promotion.

6 Conclusion

When democratic societies confront morally charged issues, what role do elites play in shaping public opinion? While contentious debates often accompany citizen polarization, we still do not know what mechanisms, if any tie the two together. While some scholars argue that elite divisions contribute to increased polarization among the mass public, recent empirical research has demonstrated that such elite polarization does not necessarily translate into broader public division (Westwood et al., 2022). Our paper il-

lustrates how contentious processes surrounding morally complex policies, particularly in the realm of TJ, can play a crucial role in shaping public opinion. As we explore the intricacies of TJ, specifically its role in reconciling societies recovering from authoritarianism and political violence, we uncover how conflicting views on justice, accountability, and the past contribute to moral reckoning. This paper offers an in-depth analysis of what constitutes a "successful" TJ process and identifies the conditions under which such success is more likely to occur.

We propose a model where a TJ institution such as a truth commission is set up and its work proceeds interdependently of elite influence. In anticipation of the outcomes, citizens form expectation with some portion of citizens aware of their principled beliefs about where such a TJ process should arrive. Once the TJ institution produces an outcome, elites engage in its promotion and contestation. In response to such debates the citizens decide to accept or reject the TJ process.

Our paper reconstructs this dynamic under two scenarios, one where public expectations concerning the TJ outcome are aligned and one where they are not. We make predictions about how both expectations of citizens regarding the TJ outcome and how citizens TJ ideals shape the promotion activities that elites engage in. We find that as the share of principled citizens increases, elites will spend less efforts on promoting or contesting TJ outcomes. We also find that societies that emerge from conflict more polarized will depress the efforts that elites put into promotion and contestation of TJ outcomes. This is because in polarized societies such efforts are less efficient. Crucial to our contribution is that the success of TJ lies not in the quantity or completeness of measures implemented, but in the broader societal acceptance of the process. By focusing on the evenhandedness in the distribution of blame and responsibility, we highlight that public acceptance of TJ outcomes is rooted in how the process is framed and contested.

References

Adler, Nanci, "Reconciliation with–or Rehabilitation of–the Soviet Past?," *Memory Studies*, 2012, *5* (3), 327–338.

__ , *The Gulag survivor: beyond the Soviet system*, Routledge, 2017.

Aidt, Toke, Jean Lacroix, and Pierre-Guillaume Méon, "The Origins of Elite Persistence: Evidence from Political Purges in Post-World War II France," *CESifo Working Paper*, 2022.

Ang, Milena and Monika Nalepa, "Can transitional justice improve the quality of representation in new democracies?," *World Politics*, 2019, *71* (4), 631–666.

Bates, Genevieve, "Backlash and Beyond: Three Perspectives on the Politics of International Justice," 2024.

Berman, Zoë Elizabeth, "Ubunyarwanda and the evolution of transitional justice in post-genocide Rwanda: "To generalize is not fresh"," *African Studies Review*, 2023, *66* (3), 777–800.

Bornkamm, Paul Christoph, *Rwanda's Gacaca courts: between retribution and reparation*, Oxford University Press, 2012.

Broockman, David E, Joshua L Kalla, and Sean J Westwood, "Does affective polarization undermine democratic norms or accountability? Maybe not," *American Journal of Political Science*, 2023, *67* (3), 808–828.

Chamarbagwala, Rubiana and Hilcías E Morán, "The human capital consequences of civil war: Evidence from Guatemala," *Journal of development economics*, 2011, *94* (1), 41–61.

Clark, Phil, *The Gacaca courts, post-genocide justice and reconciliation in Rwanda: Justice without lawyers*, Cambridge University Press, 2010.

Correa, Cristián, "Reparations in Peru: From recommendations to implementation," Technical Report, International Center for Transitional Justice 2013. [HTML].

Crawford, Kathryn Lee, "Due obedience and the rights of victims: Argentina's transition to democracy," *Hum. Rts. Q.*, 1990, *12*, 17.

Crenzel, Emilio, *The memory of the Argentina disappearances: The political history of nunca más*, Routledge, 2012.

Dewan, Torun and David P Myatt, "The qualities of leadership: Direction, communication, and obfuscation," *American Political Science Review*, 2008, *102* (3), 351–368.

Edmond, Chris, "Information manipulation, coordination, and regime change," *Review of Economic studies*, 2013, *80* (4), 1422–1458.

Elster, Jon, *Closing the books: Transitional justice in historical perspective*, Cambridge University Press, 2004.

Fiedler, Charlotte and Karina Mross, "Dealing With the Past for a Peaceful Future? Analysing the Effect of Transitional Justice Instruments on Trust in Postconflict Societies," *International Journal of Transitional Justice*, 2023, *17* (2), 303–321.

Frank, Robert H, "The Political Economy of Preference Falsification: Timur Kuran's Private Truths, Public Lies," *Journal of Economic Literature*, 1996, *34* (1), 115–123.

Gibson, James L, "Overcoming apartheid: Can truth reconcile a divided nation?," *Politikon*, 2004, *31* (2), 129–155.

__ , "The contributions of truth to reconciliation: Lessons from South Africa," *Journal of Conflict Resolution*, 2006, *50* (3), 409–432.

__ , *Overcoming historical injustices: Land reconciliation in South Africa*, Cambridge University Press, 2009.

Glaeser, Edward L, Giacomo AM Ponzetto, and Andrei Shleifer, "Why does democracy need education?," *Journal of Economic Growth*, 2007, *12*, 77–99.

Grillo, Edoardo, Zhaotian Luo, Monika Nalepa, and Carlo Prato, "Theories of Democratic Backsliding," *Annual Review of Political Science*, 2023, *27*.

Hoeres, Peter and Hubertus Knabe, *After Dictatorship: Instruments of Transitional Justice in Post-Authoritarian Systems*, De Gruyter, 2023.

Horne, Cynthia M, "Lustration, transitional justice, and social trust in post-Communist countries. Repairing or wresting the ties that bind?," *Europe-Asia Studies*, 2014, *66* (2), 225–254.

__ , *Building trust and democracy: Transitional justice in post-communist countries*, Oxford University Press, 2017.

Humphrey, Michael and Estela Valverde, "Human rights politics and injustice: Transitional justice in Argentina and South Africa," *The International Journal of Transitional Justice*, 2008, *2* (1), 83–105.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood, "The Origins and Consequences of Affective Polarization in the United States," *Annual Review of Political Science*, 2019, *22* (Volume 22, 2019), 129–146.

Kitchen, Martin and Lauren Faulkner Rossi, *A history of modern Germany: 1800 to the present*, John Wiley & Sons, 2023.

Kuran, Timur, *Private truths, public lies: The social consequences of preference falsification*, Harvard University Press, 1998.

Little, Andrew T, "Propaganda and credulity," *Games and Economic Behavior*, 2017, *102*, 224–232.

Loyle, Cyanne E., *Escaping Justice. Impunity for State Crimes in the Age of Accountability*, Oxford University Press, 2024.

Ludi, Regula, *Reparations for Nazi victims in postwar Europe*, Cambridge University Press, 2012.

Mayer, Peter and Frank Ziegele, "Competition, autonomy and new thinking: Transformation of higher education in Federal Germany," *Higher Education Management and Policy*, 2009, *21* (2), 1–20.

Mitton, Kieran, *Rebels in a rotten state: understanding atrocity in the Sierra Leone civil war*, Oxford University Press, 2015.

Mussi, Francesca and Francesca Mussi, "Introduction: The South African TRC and Its Narrative Legacies," *Literary Legacies of the South African TRC: Fictional Journeys into Trauma, Truth, and Reconciliation*, 2020, pp. 1–39.

Nalepa, Monika, *After Authoritarianism* Political Economy of Institutions and Decisions, Cambridge University Press, 2022.

Neundorf, Anja, Johannes Gerschewski, Natasha Lindstaedt, Roman-Gabriel Olar, and Rosalind Shorrocks, "Autocratic and democratic imprinting: The long-term impact of education on democratic support," *Working paper*, 2023.

Niezen, Ronald, "Speaking for the dead: The memorial politics of genocide in Namibia and Germany," *International Journal of Heritage Studies*, 2018, *24* (5), 547–567.

Ocantos, Ezequiel González, "Persuade them or oust them: Crafting judicial change and transitional justice in Argentina," *Comparative Politics*, 2014, *46* (4), 479–498.

Orr, Lilla V, Anthony Fowler, and Gregory A Huber, "Is affective polarization driven by identity, loyalty, or substance?," *American Journal of Political Science*, 2023, *67* (4), 948–962.

Öztürk, Aykut, Anja Neundorf, Steven Finkel, and Ericka Rascón Ramírez, "(When) Does online civic education work: Evidence from a cross-national experiment," *Working paper*, 2023.

Romeike, Sanya, *Transitional Justice in Germany after 1945 and after 1990*, International Nuremberg Principles Academy, 2016.

Ross, Amy, "The creation and conduct of the Guatemalan Commission for Historical Clarification," *Geoforum*, 2006, *37* (1), 69–81.

Schabas, William A, "Genocide trials and gacaca courts," *Journal of International Criminal Justice*, 2005, *3* (4), 879–895.

Schroeder, Steven M, *To Forget it All and Begin Anew: Reconciliation in Occupied Germany, 1944-1954*, University of Toronto Press, 2013.

Schwartz, Rachel A and Scott Straus, "What drives violence against civilians in civil war? Evidence from Guatemala's conflict archives," *Journal of Peace Research*, 2018, *55* (2), 222–235.

Segura, Rafael Barrantes, "Reparations and displacement in Peru. Case Studies on Transitional Justice and Displacement," Technical Report, International Center for Transitional Justice 2012. [HTML].

Sikkink, Kathryn, *The Justice Cascade: How Human Rights Prosecutions Are Changing World Politics* The Norton Series in World Politics, Norton, 2012.

Stanley, Elizabeth, "Evaluating the truth and reconciliation commission," *The Journal of Modern African Studies*, 2001, *39* (3), 525–546.

Stasavage, David, "Democracy and education spending in Africa," *American journal of political science*, 2005, *49* (2), 343–358.

Subotić, Jelena, *Hijacked justice: Dealing with the past in the Balkans*, Cornell University Press, 2011.

Swart, Mia, "Name changes as symbolic reparation after transition: The examples of Germany and South Africa," *German law journal*, 2008, *9* (2), 105–121.

Tomuschat, Christian, "Clarification commission in Guatemala," *Human Rights Quarterly*, 2001, *23* (2), 233–258.

Tovy, Jacob, *Israel and the Question of Reparations from Germany: Post-Holocaust Reckonings (1949–1953)*, Walter de Gruyter GmbH & Co KG, 2023.

Villamil, Francisco and Laia Balcells, "Do TJ policies cause backlash? Evidence from street name changes in Spain," *Research & Politics*, 2021, *8* (4), 20531680211058550.

Westwood, Sean J, Justin Grimmer, Matthew Tyler, and Clayton Nall, "Current research overstates American support for political violence," *Proceedings of the National Academy of Sciences*, 2022, *119* (12), e2116870119.

Wilke, Christiane, "Remembering complexity? Memorials for Nazi victims in Berlin," *International Journal of Transitional Justice*, 2013, *7* (1), 136–156.

Wilson, Richard, *The politics of truth and reconciliation in South Africa: Legitimizing the post-apartheid state*, Cambridge University Press, 2001.

Zvobgo, Kelebogile, "Demanding truth: The global transitional justice network and the creation of truth commissions," *International Studies Quarterly*, 2020, *64* (3), 609–625.

Appendix

Table of Contents

A Formal Details Summary

Table 1: Key Symbols and Their Interpretations

Symbol	Interpretation
θ	True state of the world; reflects the fairness or even-handedness of the TJ process.
p	Level of TJ promotion chosen by the government.
r	Level of TJ resistance or de-legitimation chosen by the opposition.
s	Signal observed by citizens: $s = \theta + p - r$.
π_i^0	Citizen i's prior belief about θ.
π_i^s	Posterior belief of citizen i about θ after observing signal s.
a_i	Action chosen by citizen i, representing degree of acceptance of the TJ outcome.
κ_i	Citizen i's normative demand for accountability (vs. reconciliation).
γ_i	Weight citizen i places on matching action to perceived state of the world (i.e., precision motive).
$u_G(p)$	Government utility: fraction of citizens accepting TJ outcome minus cost of promotion.
$u_R(r)$	Opposition utility: fraction of citizens rejecting TJ outcome minus cost of resistance.
$c_G(p)$	Cost function for government promoting TJ.
$c_R(r)$	Cost function for opposition resisting TJ.
λ_i	Bayesian update weight determining influence of signal vs. prior on posterior belief.

B Formal Details and Equilibrium Derivations

B.1 Equilibrium with Unimodal Public Opinion

First let us denote the sum of all actions undertaken by all citizens as the aggregate acceptance level expressed as

$$\int_0^1 a_i^* \, di = \int_0^1 (\gamma_i \mathbb{E}_i[\theta|s] + (1-\gamma_i)\kappa_i) \, di \tag{14}$$

Assuming the weight γ_i and demand for accountability κ_i are independent, we can re-write eq. (6) as

$$\int_0^1 a_i^* \, di = \mathbb{E}[\gamma_i](\mathbb{E}[\mathbb{E}_i[\theta \mid s]] - \mathbb{E}[\kappa_i]) + \mathbb{E}[\kappa_i]$$

where $\mathbb{E}[\mathbb{E}_i[\theta \mid s]] = \mathbb{E}[\lambda_i] \cdot \big(s - (\mu_p - \mu_r)\big) + (1 - \mathbb{E}[\lambda_i]) \cdot \mathbb{E}[\mu_{0,i}]$.

Having derived posterior beliefs of citizens conditional on the signal, above we can substitute them into the government's utility given by eq. (1):

$$\begin{aligned} u_G &= \mathbb{E}[\gamma_i](\mathbb{E}[\mathbb{E}_i[\theta \mid s]] - \mathbb{E}[\kappa_i]) + \mathbb{E}[\kappa_i] - c_G(p) \\ &= \mathbb{E}[\gamma_i]\Big(\big[\mathbb{E}[\lambda_i]\big[[\theta + p - r] - \big(\mu_p - \mu_r\big)\big] + (1 - \mathbb{E}[\lambda_i])\mathbb{E}\big[\mu_{0,i}\big]\big] - \mathbb{E}[\kappa_i]\Big) + \mathbb{E}[\kappa_i] - c_G(p) \end{aligned}$$

After assuming that the government's cost function of TJ promotion $c_G(\cdot)$ is quadratic in p (i.e., $c_G(p) = \frac{c_g}{2}p^2$ where c_g is just a constant) we solve for the optimal level of government's promotion:

$$p^* = \frac{1}{c_g}\mathbb{E}[\lambda_i]\mathbb{E}[\gamma_i]$$

where $\mathbb{E}[\lambda_i] = \mathbb{E}\left[\frac{\sigma_p^{-2} + \sigma_r^{-2}}{\sigma_p^{-2} + \sigma_r^{-2} + \sigma_{0,i}^{-2}}\right]$.

We substitute into the opposition's utility from eq. (2) citizens' posterior beliefs and

the signal:

$$
\begin{aligned}
u_R &= 1 - \mathbb{E}[\gamma_i](\mathbb{E}[\mathbb{E}_i[\theta \mid s]] - \mathbb{E}[\kappa_i]) + \mathbb{E}[\kappa_i] - c_R(r) \\
&= 1 - \mathbb{E}[\gamma_i]\Big(\Big[\mathbb{E}[\lambda_i]\big[[\theta + p - r] - \big(\mu_p - \mu_r\big)\big] + (1 - \mathbb{E}[\lambda_i])\mathbb{E}\big[\mu_{0,i}\big]\Big] - \mathbb{E}[\kappa_i]\Big) + \mathbb{E}[\kappa_i] - c_R(r)
\end{aligned}
$$

And similarly as in the case of the government, we assume the opposition's cost of resisting TJ $c_R(\cdot)$ is quadratic in r, i.e., $c_R(r) = \frac{c_r}{2} r^2$ where c_r is just a constant. The optimal level of opposition's resistance is then

$$
r^* = \frac{1}{c_r}\mathbb{E}[\lambda_i]\mathbb{E}[\gamma_i]
$$

where $\mathbb{E}[\lambda_i] = \mathbb{E}\left[\frac{\sigma_p^{-2} + \sigma_r^{-2}}{\sigma_p^{-2} + \sigma_r^{-2} + \sigma_{0,i}^{-2}}\right]$.

B.2 Equilibrium with Polarized Public Opinion

Proceeding as before, the government's utility from eq. (1), can be written as:

$$
\begin{aligned}
u_G &= \mathbb{E}[\gamma_i](\mathbb{E}[\mathbb{E}[\theta|s]] - \mathbb{E}[\kappa_i]) + \mathbb{E}[\kappa_i] - c_G(p) \\
&= \mathbb{E}[\gamma_i]\left(\frac{\theta + p - r - (\mu_p - \mu_r)}{\sigma_p^2 + \sigma_r^2} \cdot \mathbb{E}\Big[\alpha_L^s \cdot \sigma_{L|s}^2 + \alpha_R^s \cdot \sigma_{R|s}^2\Big] + \text{const} - \mathbb{E}[\kappa_i]\right) + \mathbb{E}[\kappa_i] - c_G(p)
\end{aligned}
$$

Again, we assume that the government's cost promoting of TJ $c_G(\cdot)$, is quadratic in p, i.e., $c_G(p) = \frac{c_g}{2} p^2$ where c_g is just a constant. Proceeding as before, we solve for the optimal level of government's promotion:

$$
p^{**} = \frac{1}{c_g}\mathbb{E}\big[\widetilde{\lambda}_i\big]\mathbb{E}[\gamma_i]
$$

where $\mathbb{E}\big[\widetilde{\lambda}_i\big] = \mathbb{E}\left[\frac{\alpha_{L,i}^s \cdot \sigma_{L,i|s}^2 + \alpha_{R,i}^s \cdot \sigma_{R,i|s}^2}{\sigma_p^2 + \sigma_r^2}\right] = \mathbb{E}\left[\frac{\alpha_{L,i}^s \cdot \sigma_{0,i,L}^{-2} + \alpha_{R,i}^s \cdot \sigma_{0,i,R}^{-2} + 2(\sigma_p^2 + \sigma_r^2)^{-1}}{\sigma_p^2 + \sigma_r^2}\right]$.

Finally, we consider the Opposition's problem. With the opposition's utility given by

eq. (2), it turns into

$$
\begin{aligned}
u_R &= 1 - \mathbb{E}[\gamma_i](\mathbb{E}[\mathbb{E}[\theta|s]] - \mathbb{E}[\kappa_i]) + \mathbb{E}[\kappa_i] - c_R(r) \\
&= 1 - \mathbb{E}[\gamma_i]\left(\frac{\theta + p - r - (\mu_p - \mu_r)}{\sigma_p^2 + \sigma_r^2} \cdot \mathbb{E}\left[\alpha_L^s \cdot \sigma_{L|s}^2 + \alpha_R^s \cdot \sigma_{R|s}^2\right] + \text{const} - \mathbb{E}[\kappa_i]\right) + \mathbb{E}[\kappa_i] - c_R(r)
\end{aligned}
$$

Assuming that the cost of TJ resistance $c_R(\cdot)$ is quadratic in r, as we have previously, the optimal level of resistance is:

$$
r^{**} = \frac{1}{c_r}\mathbb{E}\left[\widetilde{\lambda}_i\right]\mathbb{E}[\gamma_i]
$$

where $\mathbb{E}\left[\widetilde{\lambda}_i\right] = \mathbb{E}\left[\frac{\alpha_{L,i}^s \cdot \sigma_{L,i|s}^2 + \alpha_{R,i}^s \cdot \sigma_{R,i|s}^2}{\sigma_p^2 + \sigma_r^2}\right] = \mathbb{E}\left[\frac{\alpha_{L,i}^s \cdot \sigma_{0,i,L}^{-2} + \alpha_{R,i}^s \cdot \sigma_{0,i,R}^{-2} + 2(\sigma_p^2 + \sigma_r^2)^{-1}}{\sigma_p^2 + \sigma_r^2}\right]$.

B.3 Comparative Statics

B.3.1 Aggregated Action

When citizens' prior beliefs are normally distributed, the aggregated actions take the form:

$$
\int_0^1 a_i^* \, di = \mathbb{E}[\gamma_i]\left(\mathbb{E}[\lambda_i]\left[s - \left(\mu_p - \mu_r\right)\right] + (1 - [\lambda_i])\mathbb{E}\left[\mu_{0,i}\right] - \mathbb{E}[\kappa_i]\right) + \mathbb{E}[\kappa_i]
$$

where $\mathbb{E}[\lambda_i] = \mathbb{E}\left[\frac{\sigma_p^{-2} + \sigma_r^{-2}}{\sigma_p^{-2} + \sigma_r^{-2} + \sigma_{0,i}^{-2}}\right]$ captures the average responsiveness of beliefs to the signal s, depending on the relative precision of the government's promotion, opposition's resistance, and citizens' average priors. In contrast, when prior beliefs are bimodally distributed—reflecting a polarized society—the aggregate action becomes:

$$
\int_0^1 a_i^* \, di = \mathbb{E}[\gamma_i]\left(\mathbb{E}\left[\alpha_{L,i}^s \cdot \sigma_{L,i|s}^2 \cdot \frac{\mu_{0,i,L}}{\sigma_{0,i,L}^2} + \alpha_{R,i}^s \cdot \sigma_{R,i|s}^2 \cdot \frac{\mu_{0,i,R}}{\sigma_{0,i,R}^2}\right] + \left(s - (\mu_p - \mu_r)\right)\mathbb{E}\left[\widetilde{\lambda}_i\right] - \mathbb{E}[\kappa_i]\right) + \mathbb{E}[\kappa_i]
$$

where the posterior mean belief reflects a mixture of belief clusters with respective weights, and the term $\mathbb{E}\left[\widetilde{\lambda}_i\right] = \mathbb{E}\left[\frac{\alpha_{L,i}^s \cdot \sigma_{L,i|s}^2 + \alpha_{R,i}^s \cdot \sigma_{R,i|s}^2}{\sigma_p^2 + \sigma_r^2}\right] = \mathbb{E}\left[\frac{\alpha_{L,i}^s \cdot \sigma_{0,i,L}^{-2} + \alpha_{R,i}^s \cdot \sigma_{0,i,R}^{-2} + 2(\sigma_p^2 + \sigma_r^2)^{-1}}{\sigma_p^2 + \sigma_r^2}\right]$ cap-

tures the aggregate sensitivity of these belief groups to the signal s.

B.3.2 Effectiveness of Opposition Resistance

Now let us briefly discuss the opposition's level of resistance r^{**}, since the reasoning is analogous to the effectiveness of government promotion.

When beliefs are distributed normally (left panel in Figure 3), citizens update their beliefs smoothly in response to opposition's resistance. The opposition's resistance is effective when:

$$\sigma_r^2 < \frac{1}{3}\sigma_p^2 + \mathbb{E}\left[\sigma_{0,i}^2\right]. \tag{15}$$

If σ_r^2 is too high, citizens discount the message, relying instead on their prior beliefs. As σ_r^2 increases beyond the threshold, p^* decreases sharply, making additional promotion ineffective. That is, in a society where beliefs follow a normal distribution, new information updates public perception continuously. If opposition resistance is consistent, it can significantly shift opinions. However, if it is noisy (σ_r^2 is too high), citizens ignore it and stick to their priors.

Conversely, when beliefs are distributed bimodally (right panel of Figure 3), polarization weakens belief updating because citizens belong to two separate belief clusters (L and R). The effectiveness of opposition resistance follows a different threshold (defined on the right-hand side of eq. 16):

$$\sigma_r^2 > \frac{1}{3}\sigma_p^2, \tag{16}$$

which resembles the inefficient level of resistance (the purple line in Figure 3) in the normal distribution case. In contrast to the case with Normally distributed citizens' beliefs, the variance of the initial belief, $\mathbb{E}[\sigma_{0,i}^2]$, does not factor into the threshold condition (eq. 16). This indicates that polarization-driven rigidity outweighs the role of initial uncertainty. When σ_r^2 becomes sufficiently large, resistance not only loses its ability to counteract misinformation but can actually intensify polarization, as each side interprets the

act of resistance in its own way.

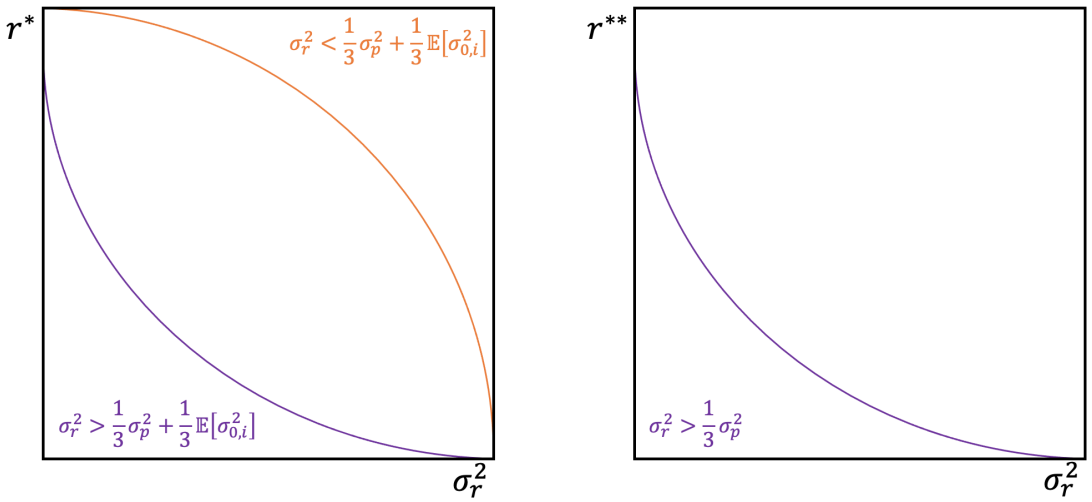

Figure 3: Efficiency of opposition's resistance: when citizens beliefs are distributed normally (r^*, left panel) and when citizens beliefs are distributed bimodally (r^{**}, right panel).

Similarly to what we saw with government promotion efforts, in a polarized society, the same resistance effort affects different groups differently. If the opposition is too aggressive or lacking in credibility, its resistance risks backfiring, reinforcing pre-existing beliefs rather than undermining government messaging. Unlike the unimodal case, effectiveness depends not just on intensity, but on overcoming entrenched beliefs, making successful resistance inherently more challenging.

C Formal Details and Proofs

Observation 1. *Suppose the prior belief follows a normal distribution* $\pi_i^0 \sim \mathcal{N}\left(\mu_{0,i}, \sigma_{0,i}^2\right)$. *Suppose also that there a signal* $s = \theta + p - r$, *where both* p *and* r *follow a normal distribution such that* $p \sim \mathcal{N}\left(\mu_p, \sigma_p^2\right)$ *and* $r \sim \mathcal{N}\left(\mu_r, \sigma_r^2\right)$. *Then the posterior belief is* $\pi^s \sim \mathcal{N}\left(\lambda\left[s - \left(\mu_p - \mu_r\right)\right] + (1-\lambda)\mu_{0,i}, (1-\lambda)\sigma_{0,i}^2\right)$ *where* $\lambda = \frac{\sigma_p^{-2}+\sigma_r^{-2}}{\sigma_p^{-2}+\sigma_r^{-2}+\sigma_{0,i}^{-2}}$.

Proof of Observation 1. To derive the posterior belief, we proceed in three steps.

We use Bayesian updating to compute the posterior distribution $\pi^s(\theta)$ given the observed signal s.

Step 1: Likelihood Function. Given the signal structure:

$$s = \theta + p - r,$$

and knowing that p and r are normally distributed as:

$$p \sim \mathcal{N}\left(\mu_p, \sigma_p^2\right), \quad r \sim \mathcal{N}\left(\mu_r, \sigma_r^2\right),$$

we express the conditional distribution of s given θ as:

$$s|\theta \sim \mathcal{N}\left(\theta + (\mu_p - \mu_r), \sigma_p^2 + \sigma_r^2\right).$$

That is, the likelihood function for s given θ is:

$$f(s|\theta) = \frac{1}{\sqrt{2\pi(\sigma_p^2 + \sigma_r^2)}} \exp\left(-\frac{(s - (\theta + \mu_p - \mu_r))^2}{2(\sigma_p^2 + \sigma_r^2)}\right).$$

Step 2: Prior Distribution of θ. By assumption, the prior belief about θ is normally distributed:

$$\pi^0 \sim \mathcal{N}(\mu_{0,i}, \sigma_{0,i}^2).$$

Thus, the prior density function is:

$$f(\theta) = \frac{1}{\sqrt{2\pi\sigma_{0,i}^2}} \exp\left(-\frac{(\theta-\mu_{0,i})^2}{2\sigma_{0,i}^2}\right).$$

Step 3: Posterior Distribution using Bayes' Rule. Using Bayes' theorem, the posterior distribution is proportional to the product of the prior and likelihood functions:

$$\begin{aligned} f(\theta|s) &\propto f(s|\theta)f(\theta) \\ &= \exp\left(-\frac{(s-(\theta+\mu_p-\mu_r))^2}{2(\sigma_p^2+\sigma_r^2)}\right) \cdot \exp\left(-\frac{(\theta-\mu_{0,i})^2}{2\sigma_{0,i}^2}\right). \end{aligned}$$

Rewriting the exponent terms:

$$-\frac{(s-(\theta+\mu_p-\mu_r))^2}{2(\sigma_p^2+\sigma_r^2)} - \frac{(\theta-\mu_{0,i})^2}{2\sigma_{0,i}^2}.$$

Expanding the squared terms:

$$-\frac{(s-\theta-\mu_p+\mu_r)^2}{2(\sigma_p^2+\sigma_r^2)} - \frac{(\theta-\mu_{0,i})^2}{2\sigma_{0,i}^2}.$$

Grouping terms in θ:

$$-\frac{s^2-2s\theta-2s(\mu_p-\mu_r)+\theta^2+2\theta(\mu_p-\mu_r)+(\mu_p-\mu_r)^2}{2(\sigma_p^2+\sigma_r^2)} - \frac{(\theta-\mu_{0,i})^2}{2\sigma_{0,i}^2}.$$

Completing the square in θ and simplifying, we obtain a normal posterior distribution:

$$\theta|s \sim \mathcal{N}\left(\lambda_i\left[s-(\mu_p-\mu_r)\right]+(1-\lambda_i)\mu_{0,i}, (1-\lambda_i)\sigma_{0,i}^2\right),$$

where:

$$\lambda_i = \frac{\sigma_p^{-2}+\sigma_r^{-2}}{\sigma_p^{-2}+\sigma_r^{-2}+\sigma_{0,i}^{-2}}.$$

□

Observation 2. *If each belief is normally distributed, i.e.,* $\pi_1 \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$, $\pi_2 \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$, $\cdots$, $\pi_{I-1} \sim \mathcal{N}\left(\mu_{I-1}, \sigma_{I-1}^2\right)$, $\pi_I \sim \mathcal{N}\left(\mu_I, \sigma_I^2\right)$, *then the public opinion* $\mathscr{X}$ *given by the sum of the beliefs all citizens in the society is also normally distributed, i.e.,* $\sum_{i=1}^{I} \pi_i = \mathscr{X} \sim \mathcal{N}\left(\mu_{\mathscr{X}}, \sigma_{\mathscr{X}}^2\right)$ *where* $\mu_{\mathscr{X}} = \sum_{i=1}^{I} \mu_i$ *and* $\sigma_{\mathscr{X}}^2 = \sum_{i=1}^{I} \sigma_i^2$.[9, 10]

In Observation 3 we depict the updating of a generic prior belief to simplify the notation: without the loss of generality we omit indexes of i, since all citizens update their beliefs simultaneously and, hence, independently).

Observation 3. *Suppose the prior belief about* θ *is a bimodal distribution given by a mixture of two normal distributions:* $\pi^0(\theta) = \alpha \cdot \mathcal{N}(\mu_{0,L}, \sigma_{0,L}^2) + (1-\alpha) \cdot \mathcal{N}(\mu_{0,R}, \sigma_{0,R}^2)$, *where* $0 < \alpha < 1$, *and the subscripts* L *and* R *correspond to the two modes of the prior. Suppose also that there a signal* $s = \theta + p - r$, *where both* p *and* r *follow a normal distribution such that* $p \sim \mathcal{N}\left(\mu_p, \sigma_p^2\right)$ *and* $r \sim \mathcal{N}\left(\mu_r, \sigma_r^2\right)$. *Then the posterior belief* $\pi^s(\theta)$ *is the mixture of the two posterior components:* $\pi^s(\theta) = \alpha_L^s \cdot \mathcal{N}(\mu_{L|s}, \sigma_{L|s}^2) + \alpha_R^s \cdot \mathcal{N}(\mu_{R|s}, \sigma_{R|s}^2)$ *with the mean* $\mathbb{E}[\theta \mid s] = \alpha_L^s \cdot \mu_{L|s} + \alpha_R^s \cdot \mu_{R|s}$ *and the variance* $\mathbb{V}[\theta \mid s] = \alpha_L^s \left[\sigma_{L|s}^2 + \left(\alpha_R^s\right)^2 \left(\mu_{L|s} - \mu_{R|s}\right)^2\right] + \alpha_R^s \left[\sigma_{R|s}^2 + \left(\alpha_L^s\right)^2 \left(\mu_{R|s} - \mu_{L|s}\right)^2\right]$ *where where* $\mu_{L|s} = \sigma_{L|s}^2 \left(\frac{\mu_{0,L}}{\sigma_{0,L}^2} + \frac{s-(\mu_p-\mu_r)}{\sigma_p^2+\sigma_r^2}\right)$, $\mu_{R|s} = \sigma_{R|s}^2 \left(\frac{\mu_{0,R}}{\sigma_{0,R}^2} + \frac{s-(\mu_p-\mu_r)}{\sigma_p^2+\sigma_r^2}\right)$, $\alpha_L^s = \frac{\widetilde{\alpha}_L}{\widetilde{\alpha}_L+\widetilde{\alpha}_R}$, $\alpha_R^s = \frac{\widetilde{\alpha}_R}{\widetilde{\alpha}_L+\widetilde{\alpha}_R}$, $\widetilde{\alpha}_L = \alpha \cdot \mathcal{N}\left(s; \mu_{0,L} + (\mu_p - \mu_r), \sigma_{0,L}^2 + \sigma_p^2 + \sigma_r^2\right)$, *and* $\widetilde{\alpha}_R = (1-\alpha) \cdot \mathcal{N}\left(s; \mu_{0,R} + (\mu_p - \mu_r), \sigma_{0,R}^2 + \sigma_p^2 + \sigma_r^2\right)$.

Proof of Observation 3. To derive the posterior belief for the case of bimodal prior beliefs, we proceed in three steps.

The prior belief about θ is a bimodal distribution given by a mixture of two normal distributions:

$$\pi^0(\theta) = \alpha \cdot \mathcal{N}(\mu_{0,L}, \sigma_{0,L}^2) + (1-\alpha) \cdot \mathcal{N}(\mu_{0,R}, \sigma_{0,R}^2),$$

[9] To see this, note that $\mathbb{E}[\mathscr{X}] = \mathbb{E}\left[\sum_{i=1}^{I} \pi_i\right] = \sum_{i=1}^{I} \mathbb{E}[\pi_i] = \sum_{i=1}^{I} \mu_i$ and $\mathbb{V}[\mathscr{X}] = \mathbb{V}\left[\sum_{i=1}^{I} \pi_i\right] = \sum_{i=1}^{I} \mathbb{V}[\pi_i] = \sum_{i=1}^{I} \sigma_i^2$.

[10] Following Observations 1 and 2, we denote prior and posterior public opinion by $\sum_{i=1}^{I} \pi_i^0 = \mathscr{X}^0 \sim \mathcal{N}\left(\mu_{0,\mathscr{X}}, \sigma_{0,\mathscr{X}}^2\right)$ and $\sum_{i=1}^{I} \pi_i^s = \mathscr{X}^s \sim \mathcal{N}\left(\mu_{s,\mathscr{X}}, \sigma_{s,\mathscr{X}}^2\right)$ respectively.

where $0 < \alpha < 1$, and the subscripts L and R correspond to the two modes of the prior.

The observed signal is

$$s = \theta + p - r,$$

where $p \sim \mathcal{N}(\mu_p, \sigma_p^2)$ and $r \sim \mathcal{N}(\mu_r, \sigma_r^2)$. This implies

$$s \mid \theta \sim \mathcal{N}\big(\theta + (\mu_p - \mu_r), \sigma_p^2 + \sigma_r^2\big).$$

Our goal is to derive the posterior belief $\pi^s(\theta)$, as well as the posterior mean and variance.

Step 1: Posterior Belief for Each Prior Component. For each component of the prior, the posterior belief is updated independently. Starting with $\pi_L^0(\theta) \sim \mathcal{N}(\mu_{0,L}, \sigma_{0,L}^2)$, the posterior given s is

$$\pi_L^s(\theta) \sim \mathcal{N}(\mu_{L|s}, \sigma_{L|s}^2),$$

where:

The posterior variance is determined by the precisions:

$$\frac{1}{\sigma_{L|s}^2} = \frac{1}{\sigma_{0,L}^2} + \frac{1}{\sigma_p^2 + \sigma_r^2}.$$

The posterior mean is:

$$\mu_{L|s} = \sigma_{L|s}^2 \left(\frac{\mu_{0,L}}{\sigma_{0,L}^2} + \frac{s - (\mu_p - \mu_r)}{\sigma_p^2 + \sigma_r^2} \right).$$

Similarly, for the second component $\pi_R^0(\theta) \sim \mathcal{N}(\mu_{0,R}, \sigma_{0,R}^2)$, the posterior is

$$\pi_R^s(\theta) \sim \mathcal{N}(\mu_{R|s}, \sigma_{R|s}^2),$$

where:

The posterior variance is:

$$\frac{1}{\sigma^2_{R|s}} = \frac{1}{\sigma^2_{0,R}} + \frac{1}{\sigma^2_p + \sigma^2_r}.$$

The posterior mean is:

$$\mu_{R|s} = \sigma^2_{R|s}\left(\frac{\mu_{0,R}}{\sigma^2_{0,R}} + \frac{s - (\mu_p - \mu_r)}{\sigma^2_p + \sigma^2_r}\right).$$

Step 2: Posterior Mixture Weights. The posterior mixture weights are updated using Bayes's rule. The unnormalized weight for the L-component is proportional to the prior weight α times the marginal likelihood of observing s under the L-component prior:

$$\widetilde{\alpha}_L = \alpha \cdot \mathcal{N}\big(s; \mu_{0,L} + (\mu_p - \mu_r), \sigma^2_{0,L} + \sigma^2_p + \sigma^2_r\big).$$

Similarly, for the R-component:

$$\widetilde{\alpha}_R = (1-\alpha) \cdot \mathcal{N}\big(s; \mu_{0,R} + (\mu_p - \mu_r), \sigma^2_{0,R} + \sigma^2_p + \sigma^2_r\big).$$

The normalized posterior weights are then:

$$\alpha^s_L = \frac{\widetilde{\alpha}_L}{\widetilde{\alpha}_L + \widetilde{\alpha}_R}, \quad \alpha^s_R = \frac{\widetilde{\alpha}_R}{\widetilde{\alpha}_L + \widetilde{\alpha}_R}.$$

Step 3: Posterior Mixture Distribution. The posterior belief $\pi^s(\theta)$ is the mixture of the two posterior components:

$$\pi^s(\theta) = \alpha^s_L \cdot \mathcal{N}(\mu_{L|s}, \sigma^2_{L|s}) + \alpha^s_R \cdot \mathcal{N}(\mu_{R|s}, \sigma^2_{R|s}).$$

Step 4: Posterior Mean and Variance. The posterior mean is given by the mixture

mean:

$$\mathbb{E}[\theta \mid s] = \alpha_L^s \cdot \mu_{L|s} + \alpha_R^s \cdot \mu_{R|s}.$$

The posterior variance is given by:

$$\mathrm{Var}(\theta \mid s) = \alpha_L^s\big(\sigma_{L|s}^2 + \mu_{L|s}^2\big) + \alpha_R^s\big(\sigma_{R|s}^2 + \mu_{R|s}^2\big) - \big(\mathbb{E}[\theta \mid s]\big)^2.$$

Equivalently, using the decomposition into within-component and between-component variance:

$$\mathrm{Var}(\theta \mid s) = \alpha_L^s \cdot \sigma_{L|s}^2 + \alpha_R^s \cdot \sigma_{R|s}^2 + \alpha_L^s \cdot (\mu_{L|s} - \mathbb{E}[\theta \mid s])^2 + \alpha_R^s \cdot (\mu_{R|s} - \mathbb{E}[\theta \mid s])^2.$$

□

Observation 4. *If each belief is distributed bimodally, i.e.,* $\pi_1 \sim \alpha_i \mathcal{N}(\mu_{i,L}, \sigma_{i,L}^2) + (1 - \alpha_i)\mathcal{N}(\mu_{i,R}, \sigma_{i,R}^2)$*,* $\cdots$*,* $\pi_I \sim \alpha_i \mathcal{N}(\mu_{i,L}, \sigma_{i,L}^2) + (1-\alpha_i)\mathcal{N}(\mu_{i,R}, \sigma_{i,R}^2)$*, then the public opinion* $\mathscr{X}$ *given by the sum of the beliefs all citizens in the society is either unimodally or multimodally distributed, with expectation* $\mathbb{E}[\mathscr{X}] = \sum_{i=1}^N \big(\alpha_i \mu_{i,L} + (1-\alpha_i)\mu_{i,R}\big)$ *and variance* $\mathbb{V}[\mathscr{X}] = \sum_{i=1}^N \Big(\alpha_i \sigma_{i,L}^2 + (1-\alpha_i)\sigma_{i,R}^2 + \alpha_i(1-\alpha_i)(\mu_{i,L} - \mu_{i,R})^2\Big)$*.*[11]

[11] Since the $\mathbb{E}[\pi_i] = \alpha_i \mu_{i,L} + (1-\alpha_i)\mu_{i,R}$, summing over all I independent variables, the expected value of $\mathscr{X}$ is $\mathbb{E}[\mathscr{X}] = \mathbb{E}\Big[\sum_{i=1}^I X_i\Big] = \sum_{i=1}^I \mathbb{E}[\pi_i] = \sum_{i=1}^I \big(\alpha_i \mu_{i,L} + (1-\alpha_i)\mu_{i,R}\big)$. Since the $\mathbb{V}[\pi_i] = \alpha_i \sigma_{i,L}^2 + (1-\alpha_i)\sigma_{i,R}^2 + \alpha_i(1-\alpha_i)(\mu_{i,L} - \mu_{i,R})^2$, the variance of their sum is given by $\mathbb{V}[\mathscr{X}] = \sum_{i=1}^I \mathbb{V}[X_i] = \mathbb{V}[\mathscr{X}] = \sum_{i=1}^I \Big(\alpha_i \sigma_{i,L}^2 + (1-\alpha_i)\sigma_{i,R}^2 + \alpha_i(1-\alpha_i)(\mu_{i,L} - \mu_{i,R})^2\Big)$.

D Examples of Truth Commissions

Here we present several examples on short-, medium-, and long-term Truth Commissions to illustrate the statement we made earlier in the paper, namely, that the preparation of a report by such a commission typically takes years.

Table 2: Examples Truth Commissions Durations.

Case	Dates	Duration
Morocco[12]	December 2004 – November 2005	$<$ 1 yr.
Serbia and Montenegro[13]	February 2002 – February 2003	1 yr.
Ghana[14]	January 2003 – October 2004	1 yr. $<$
Chile[15]	September 2003 – June 2005	$<$ 2 yr.
(ex-)East Germany[16]	May 1992 – June 1994	$\sim$ 2 yr.
Paraguay[17]	June 2004 – August 2008	4 yr. $<$
South Korea[18]	December 2005 – December 2010	5 yr.
South Africa[19]	December 1995 – 2002	6 yr.
Ethiopia[20]	1993 – 2007(?)	14(?) yr.

Note: The table presents examples of truth commissions and their durations. The duration varies significantly, ranging from less than a year (e.g., Morocco) to over a decade (e.g., Ethiopia). Among the shortest time spans for a truth and reconciliation commission to prepare its final report, one can recall the cases of Equity and Reconciliation Commission in Morocco (less than 1 year), Yugoslav Truth and Reconciliation Commission in Serbia and Montenegro (1 year), and National Reconciliation Commission in Ghana (more than 1 year); for medium-length cases, he National Commission on Political Imprisonment and Torture in Chile (less than 2 years), the Study Commission for Working Through the History and the Consequences of the SED Dictatorship in(ex-)East Germany (about 2 years), t and the Truth and Justice Commission in Paraguay (about 4 years) can serve as examples; and among the longest, one can mention the cases of the Committee on Clearing Up Part Incidents for Truth and Reconciliation in South Korea (5 years), the Commission of Truth and Reconciliation in South Africa (approximately 7 years), and the Special Prosecution Process in Ethiopia by the Office of the Special Prosecutor in Ethiopia (about 14 years(?)). We summarize these examples in Table 2

.

[12]We refer to the Equity and Reconciliation Commission; see, e.g., "Truth Commission: Morocco", *United States Institute of Peace*, December 1, 2004.

[13]We refer to Yugoslav Truth and Reconciliation Commission; see, e.g., "Truth Commission: Serbia and Montenegro", *United States Institute of Peace*, February 1, 2002.

[14]We refer to the National Reconciliation Commission; see, e.g., "Truth Commission: Ghana", *United States Institute of Peace*, January 14, 2003.

[15]We refer to the National Commission on Political Imprisonment and Torture; see, e.g., "Commission of Inquiry: Chile 03", *United States Institute of Peace*, September 1, 2003.

[16]We refer to the Study Commission for Working Through the History and the Consequences of the SED Dictatorship in Germany; see, e.g., "Truth Commission: Germany 92", *United States Institute of Peace*, May 1, 1992.

[17]We refer to the Truth and Justice Commission; see, e.g., "Truth Commission: Paraguay", *United States Institute of Peace*, June 1, 2004.

[18]We refer to the Committee on Clearing Up Part Incidents for Truth and Reconciliation; see, e.g., "Truth Commission: South Korea 2005", *United States Institute of Peace*, April 18, 2012.

[19]We refer to the Commission of Truth and Reconciliation; see, e.g., "Truth Commission: South Africa", *United States Institute of Peace*, December 1, 1995.

[20]We refer to the Special Prosecution Process in Ethiopia by the Office of the Special Prosecutor; see, e.g., "Special Prosecutor's Office: Ethiopia", *United States Institute of Peace*, January 1, 1993.

E Public Opinion

E.1 Unimodal Public Opinion

Here we demonstrate how the aggregation of citizen beliefs formed above creates society's public opinion of the TJ outcome, (what we call the *perceived TJ outcome* denoted $\mathscr{X}$.

If all beliefs are normally distributed random variables, then the sum of these random variables is also going to be normally distributed

Thus, a continuous random variable $\mathscr{X} \in [0,1]$ represents the perceived TJ outcome (formally, this is stated in Observation 2 in Appendix C). This observation provides a convenient and intuitive interpretation of the mean $\mu_{\mathscr{X}}$: $\mu_{\mathscr{X}}$ represents the average perceived fairness of the TJ outcome in the population, and variance $\sigma^2_{\mathscr{X}}$ captures the heterogeneity of these perceptions in the population or the diversity of public opinion.

Furthermore, it is straightforward to see that if the variance of an individual belief increases by a small amount, then the variance of the public opinion of the whole population, will increases also. The same reasoning applies if the variance of an individual belief decreases. [21]

A broad set of opinions on the TJ outcome presents a challenge to both society and policymakers. In the context of the TJ outcome, this is best understood by contrasting a broad public opinion with a narrow one (corresponding to the uniform distribution). In the context of TJ, a good example of 'broad' public opinion (i.e., high $\sigma^2_{\mathscr{X}}$) is the disagreement in Post-Soviet Russia over accountability for the political violence of Stalinism. Some believed the Secret Speech delivered by Khrushchev in February of 1956 did justice to the past, others felt that state institutions should go considerably further to address the Stalinist past. Although the efforts of the 1990s—rightfully in our opinion—

[21] An explicit way to see it is to consider the effect of the change of j's variance of the belief distribution keeping all other $I-1$ beliefs constant: $\mathbb{V}[\mathscr{X}] = \sum_{i=1}^{I-1} \sigma_i^2 + \sigma_j^2$; then if a change in σ_j^2 is positive ($\Delta\sigma_j^2 > 0$) then $\mathbb{V}[\mathscr{X}]$ will increase, and if a change in σ_j^2 is negative ($\Delta\sigma_j^2 < 0$) then $\mathbb{V}[\mathscr{X}]$ will decrease. In other words, if—*ceteris paribus*—the more precise j's worldview is, the less the diversity or heterogeneity of public opinion of the society there is.

are often regarded as ambivalent attempts to deal with the Stalinist past (see, e.g., Adler, 2012, 2017), they undoubtedly went further than the Secret Speech.

A vivid example of 'narrow' public opinion in the context of the TJ outcome (i.e., $\sigma^2_{\mathscr{X}}$ is low) is the contemporary German society's largely unified stance towards atrocities committed during the Nazi era, including bans of Nazi symbols and Holocaust denial. The acceptance of strict anti-Nazi policies is reinforced by the educational system, where instruction on the Holocaust is a component of the school curriculum. [22] This institutional acknowledgment of responsibility has contributed to a widely shared view within German society (e.g., Stasavage, 2005; Glaeser et al., 2007; Neundorf et al., 2023; Öztürk et al., 2023).

E.2 Polarized Public Opinion

As previously, it is worth noticing that, the summation of the (either prior or posterior) beliefs of all citizens creates the society's aggregate public opinion on the TJ outcome $\mathscr{X}$ (see Observation 4 in Appendix C). The dynamics of public opinion formation in the bimodal case follow a pattern. First, higher dispersion within each group (increasing $\sigma^2_{0,i,L}$ or $\sigma^2_{0,i,R}$) leads to greater within-group diversity with not effect on overall polarization. That is, when the variance within each belief cluster increases, individuals within each group hold more heterogeneous views, but the overall distance between the two clusters remains the same. This means that while people within the same ideological camp may disagree on details, the societal-level polarization is not reduced, as the two groups still remain separate in their core beliefs.

Second, greater separation between modes (increasing distance $|\mu_{0,i,L} - \mu_{0,i,R}|$) is associated with higher polarization. Compromise becomes less likely, as citizens increasingly entrenched in their separate belief clusters interpret the same facts in different ways. At the same time, situations where the modes converge (decreasing distance

[22] see, e.g., Federal Agency for Civic Education, which provides educational resources on Germany's Nazi past). This education promotes a consistent and state-endorsed narrative that emphasizes responsibility, remembrance, and the moral lessons learned from that period.

$|\mu_{0,i,L} - \mu_{0,i,R}|$) resemble a normal distribution, representing public opinion with reduced polarization. We can interpret the convergence of two belief clusters as one where citizens across groups start sharing interpretations of the TJ outcome.

Finally, the dominance of one mode (shifting α_i closer to 1 or 0) results in public opinion reflecting a strong societal consensus, akin to the unimodal case. When one belief cluster gains dominance, the majority of the population aligns with a single dominant narrative, effectively reducing polarization. In this case, public opinion starts resembling a unimodal (though not necessarily normal) distribution, where a broad societal consensus emerges, leaving the minority belief group marginalized.

We can illustrate the dynamics of polarization of opinions with two examples. For a case where public opinion on TJ is relatively moderate and concentrated (polarization is low and beliefs are distributed to form a single peak, making $|\mu_{0,i,L} - \mu_{0,i,R}|$ small) consider post-communist Poland. There, despite divergences in ideas on how to organize the democratic state, the previous authoritarian regime was interpreted as an instance of Soviet occupation and ,consequently, very few were willing to defend the previous regime. This allowed for a broad consensus on how to address past injustices. A majority of the population shared the "Gruba Kreska" ideology, which in a nutshell meant that establishing a forward-looking system respecting rule of law was more important than waging a retributive justice campaign. At the same time, some individuals would have preferred more punitive measures against communist riot policemen who committed acts of torture and even murder, but they were either in a small minority or the weight they placed on these normative considerations was negligible.

As an example of a case of highly polarized public opinion, where society remains deeply divided into two conflicting belief clusters (i.e., the distance $|\mu_{0,i,L} - \mu_{0,i,R}|$ is large) consider post-Franco Spain, where the "Pact of Forgetting" created a long-lasting division in public perceptions of TJ. On one side, many conservative groups thought that Spain's transition would indefinitely avoid revisiting past crimes, believing resurrecting historical grievances would destabilize democracy. On the other side victims' families

expected an eventual reckoning with Franco's legacy. They did not think the Pact of Forgetting would imply banning exhumations of mass graves and anticipated a public acknowledgment of repression (Villamil and Balcells, 2021). The persistence of these two distinct, entrenched narratives—one expecting silence and amnesty, the other truth and memorization—corresponds to a bimodal belief distribution where the peaks are far apart, with little space for consensus.

F Further Examples of Government's Promotion and Opposition Resistance

F.1 Examples of Government's Promotion

Here we provide several additional examples on government's promotion.

The South African example helps to distinguish exogenous and endogenous components of the TJ outcome. Since the TRC's operations, hearings, and the preparation of its final report (ideally) should be independent processes aimed at uncovering the truth, the result is exogenous to any outcomes preferred by the government. In example, these activities and the resulting documentation of apartheid-era crimes were exogenous to the government in the sense that the truth commission operated with a degree of autonomy: the findings, testimonies, and recommendations in the final report represented the commission's unbiased output, grounded in the evidence gathered.

Example 1. *Established in 1995 under the leadership of Archbishop Desmond Tutu to address the atrocities committed during the apartheid era (1948–1994) (e.g., Wilson, 2001), the Commission held public hearings that were broadcast nationally, allowing for wide engagement. Its operations were divided into three committees: first, the Human Rights Violations Committee, which investigated abuses such as torture, killings, and other gross violations; second, the Amnesty Committee, which offered amnesty to perpetrators who fully disclosed their crimes and demonstrated political motivation; and third, the Reparation and Rehabilitation Committee, which sought to address the needs of victims through recommendations for reparations and support (e.g., Hoeres and Knabe, 2023). Key for motivating our assumption about exogeneity is that the TRC produced an independent final report, detailing the systemic injustices of apartheid, specific crimes, and the complicity of various institutions. The report also included recommendations for reparations and institutional reforms to prevent future violations. It became a critical historical document, shaping South Africa's collective understanding of its past (e.g., Mussi and Mussi, 2020; Hoeres and Knabe, 2023).*

Moreover, the government can set up reparation programs to provide financial compensation, healthcare, housing, education, or employment opportunities to victims of human rights abuses. Reparations are a tangible acknowledgment of the suffering endured and are aimed at restoring dignity to the victims. In addition to financial compensation, the government can promote symbolic reparations such as public apologies, memorials, and days of remembrance to honor victims and acknowledge past wrongs. Let us consider examples (2) and (3) which present the approaches undertaken by Peru and Germany to TJ, combining material reparations with symbolic acts of remembrance and apology to address historical injustices and promote reconciliation

Example 2 (Peruvian reparations to victims of the civil war)**.** *Following two decades of internal armed conflict from 1980 to 2000, which resulted in approximately 70,000 deaths and widespread human rights violations, in 2001 Peru established the Truth and Reconciliation Commission (*Comisión de la Verdad y Reconciliación*, CVR)*[23]*. The CVR's mandate was to investigate the violence and recommend measures for reconciliation. In 2005, based on the CVR's recommendations, the Peruvian government implemented the Comprehensive Reparations Program (*Programa Integral de Reparaciones*). This program aimed to provide financial compensation, healthcare, education, and housing support to victims and their families, acknowledging their suffering and aiding in restoring their dignity (e.g., Segura, 2012; Correa, 2013).*

Example 3 (German reparations to Israel)**.** *In the aftermath of World War II and the Holocaust, West Germany took significant steps to acknowledge and atone for the atrocities committed during the Nazi era (e.g., Kitchen and Rossi, 2023). In 1952, the Luxembourg Agreement was signed between West Germany and Israel, wherein Germany agreed to pay reparations to Holocaust survivors and the State of Israel (e.g., Ludi, 2012). These reparations included financial compensation to individuals and support for the resettlement of Jewish refugees (e.g., Romeike, 2016; Niezen, 2018; Tovy, 2023). Beyond financial reparations, Germany has engaged in extensive efforts to memorialize the victims of the Holocaust*

[23]CVR reports estimates between 61,007 and 77,552; see, for instance, public report on the Peru case: "Truth Commissions Digital Collection: Peru", *United States Institute of Peace*, July 13, 2001.

and confront its Nazi past. This includes the establishment of numerous memorials and museums, such as the Memorial to the Murdered Jews of Europe in Berlin, which serves as a central place of remembrance and education (e.g., Swart, 2008; Schroeder, 2013; Wilke, 2013; Niezen, 2018). Additionally, Germany has issued public apologies and enacted laws to combat Holocaust denial and hate speech, reflecting a commitment to acknowledging past wrongs and preventing their recurrence (e.g., Romeike, 2016; Kitchen and Rossi, 2023; Tovy, 2023).

At the same time, the government can undertake judicial reforms to restore the rule of law, including the independence of the judiciary, and ensure that courts are equipped to handle cases of past abuses. Strengthening judicial oversight mechanisms can help restore public trust in state institutions. Governments can also promote TJ by reforming security forces (police, military, intelligence) that were involved in past abuses. This might include vetting personnel, restructuring command chains, and instituting human rights training to prevent future abuses. To illustrate this government's strategy, let us consider the case of reforms in El Salvador (4) and Chile (5).

Example 4 (Reforms in El Salvador)**.** *After more than a decade of brutal civil war (1980–1992), El Salvador signed the Chapultepec Peace Accords in 1992, marking a significant turning point for the country. The accords laid the foundation for wide-ranging reforms to address the systemic issues that had fueled the conflict and to prevent future violence. Among the key reforms were the military restructuring and creation of the national civil police (*Policía Nacional Civil*, PNC). The military was restructured to focus exclusively on external defense, removing its involvement in internal security, which had historically led to widespread human rights abuses. The accords mandated a reduction in the size of the armed forces and emphasized the depoliticization and professionalization of military personnel. A new civilian police force was established to replace the repressive, militarized security forces that had contributed to the conflict. The PNC was tasked with maintaining public order while respecting human rights and operating under civilian oversight. The reform aimed to demilitarize law enforcement and build a professional police force trusted by the public.*

Example 5 (Reforms in Chile)**.** *Following the end of Augusto Pinochet's 17-year military dictatorship in 1990, Chile undertook judicial reforms to address the complicity of the judiciary in covering up crimes committed during the regime and to restore the rule of law. Three important domains can be highlighted: the removal of complicit judges, the establishment of judicial independence and justice for victims. During the Pinochet era, many judges were accused of turning a blind eye to or directly participating in covering up human rights violations, including disappearances, torture, and extrajudicial killings. Post-transition, efforts were made to identify and remove corrupt judges who had failed to uphold justice during the dictatorship. This process was vital in restoring public confidence in the judiciary. Reforms were enacted to strengthen judicial independence and eliminate political interference. Measures included restructuring the Supreme Court and increasing the accountability of judges. These reforms aimed to ensure that the judiciary could operate impartially, free from the control of authoritarian institutions. In the years following the transition, Chile's reformed judiciary became an essential vehicle for addressing the atrocities of the Pinochet regime. High-profile cases against former regime officials, including Pinochet himself, were brought to court, signaling a renewed commitment to justice and accountability.*

Alongside this, the government can promote the construction of national monuments, museums, or memorials to honor victims of past violence and educate future generations about historical injustices. These spaces serve as symbols of collective memory and commitment to "never again" scenarios. Furthermore, the government can officially commemorate specific days of remembrance to acknowledge past atrocities. Public apologies from political leaders are also a way of recognizing state responsibility and offering symbolic reparations. Let us consider two examples (6) and (7) from South Africa and Lithuanian respectively.

Example 6 (South African Museum of Apartheid)**.** *Opened in 2001, the Apartheid Museum in South Africa documents and educates the public about apartheid, a system of racial segregation that shaped the nation's 20th-century history. The museum narrates apartheid's rise, enforcement, and fall through exhibitions on systemic injustices, forced removals, and*

resistance movements led by figures like Nelson Mandela. Featuring photographs, testimonies, and archival materials, it honors the resilience of anti-apartheid activists while fostering a commitment to justice, equality, and reconciliation, serving as both a memorial and an educational hub.[24].

Example 7 (Lithuanian Museum of Occupations and Freedom Fights)**.** *In Lithuania, the government has established initiatives like the Museum of Occupations and Freedom Fights (formerly known as the KGB Museum) to acknowledge and educate the public about the deportations and resistance movements during Soviet occupation, particularly focusing on the forced deportation of tens of thousands of Lithuanians to Siberia and other remote regions of the USSR. These efforts include exhibitions detailing the suffering endured by deportees and the resistance efforts by partisan fighters, serving as both a memorial and an educational tool to confront the injustices of the past and preserve historical memory*[25].

Furthermore, the government can initiate national dialogues aimed at fostering reconciliation between victims and perpetrators. This could involve forums for conflict resolution, peacebuilding, and the promotion of shared narratives about the past. Governments can actively involve victims in the TJ process, ensuring their voices are heard and respected. This participation can happen in truth commissions, reparations programs, or legal proceedings, demonstrating a commitment to addressing their concerns. Examples (8) and (9) can shed light on such governments' strategies.

Example 8. *Apologies for pogroms. [TBC]*

Example 9 (Platform for victims' voices in Sierra Leone)**.** *In the aftermath of its brutal civil war (1991–2002) (e.g., Mitton, 2015), Sierra Leone established the Truth and Reconciliation Commission (TRC) to address the atrocities committed during the conflict*[26]. *The TRC provided a platform for victims, including women and children who were often marginalized in traditional justice processes, to share their experiences. This inclusive approach was crucial for national healing and reconciliation. The Commission's final report,*

[24]See the official Apartheid Museum website:[HTML].
[25]See the official Museum of Occupations and Freedom Fights website:[HTML].
[26]See the official Sierra Leone Truth and Reconciliation Commission website:[HTML].

published in 2004, offered a comprehensive account of the conflict and recommended measures to prevent future violations.

Finally, the government can include TJ themes in national education systems. Teaching students about past atrocities, human rights, and the rule of law helps prevent the repetition of past abuses and encourages civic responsibility. Governments can also organize public awareness campaigns to educate citizens about the importance of TJ, its goals, and how it contributes to societal healing and rebuilding. Let us consider two examples (10) and (11) of how Germany and Rwanda combine education with public awareness campaigns to address historical injustices and promote societal healing.

Example 10 (Education about National-Socialism in Germany)**.** *In Germany, education about the Holocaust and Nazi atrocities is a mandatory component of the school curriculum. This initiative aims to promote historical consciousness and a commitment to human rights among students. Each of Germany's 16 federal states has autonomy over its educational curriculum, leading to variations in how and to what extent schools teach about the Holocaust (e.g., Mayer and Ziegele, 2009). However, the overarching goal remains consistent: to ensure that students understand the historical events and their implications for contemporary society*[27]*. This educational approach is supported by state funding for field trips to concentration camps and museums, such as the Auschwitz Memorial and Museum Auschwitz-Birkenau*[28]*. These experiences provide students with tangible connections to history, reinforcing the lessons learned in the classroom.*

Example 11 ('I am Rwandan' public campaign in Rwanda)**.** *In post-genocide Rwanda, the government has implemented public education initiatives to promote unity and reconciliation. One such program is 'I am Rwandan' ('Ndi Umunyarwanda'), which aims to build a national identity based on trust and dignity*[29]*. This program provides forums for Rwan-*

[27] See, for instance, "Teaching the Holocaust to Germany's Generation Z", *Deutsche Welle News*, October 5, 2021; "Teaching about Nazis and the Holocaust in German schools", *Deutsche Welle News*, November 11, 2023.

[28] See, for instance, "Auschwitz visit expands students' Holocaust understanding", *Deutsche Welle News*, January 27, 2024.

[29] See, for instance, the note on the program at the official website of the Genocide Archive of Rwanda [HTML].

dans to discuss the causes and consequences of the genocide, as well as what it means to be Rwandan. By fostering open dialogue, 'Ndi Umunyarwanda' seeks to strengthen unity and reconciliation among Rwandans (e.g., Berman, 2023).

F.2 Examples of Opposition's Resistance

Here we provide several additional examples on opposition's resistance.

Moreover, the opposition or *ancien régime* elite may attempt to co-opt TJ bodies by placing their allies in key positions, ensuring that investigations are limited or skewed. This could involve influencing the appointment of judges, prosecutors, or commissioners. Another way to obstruct TJ is by cutting or limiting financial resources needed for the effective functioning of truth commissions, reparations programs, or courts. By controlling the budget, they can weaken the capacity of these institutions to operate. Let illustrate these with two examples from Rwanda (12) and Guatemala (13).

Example 12. *In the aftermath of the 1994 genocide, Rwanda faced the monumental task of prosecuting a vast number of individuals accused of participating in the recent atrocities. In order to address this, the government established the Gacaca courts (Inkiko Gacaca)—a community-based justice system inspired by traditional conflict resolution mechanisms. Aimed to expedite trials, uncover the truth, and foster reconciliation, these courts tried over 1.2 million cases between 2002 and 2012 (e.g., Schabas, 2005; Bornkamm, 2012). Despite their achievements, the Gacaca courts faced significant criticisms: allegations arose that the government manipulated these courts to target political opponents and protect allies of the ruling Rwandan Patriotic Front (e.g., Clark, 2010; Bornkamm, 2012). Lacking legal representation for defendants, the Gacaca courts deteriorated its legitimacy: a number of international organizations (e.g., Human Rights Watch[30]) reported instances where trials were flawed, with authorities influencing proceedings to settle personal vendettas or suppress dissent. The exclusion of crimes committed by Rwandan Patriotic Front soldiers from the Gacaca's jurisdiction further limited the scope of justice, leaving certain victims*

[30] See, for example, "Justice Compromised: The Legacy of Rwanda's Community-Based Gacaca Courts", *Human Rights Watch*, May 31, 2011.

without redress (e.g., Clark, 2010; Bornkamm, 2012). While not viewed as such by everyone, the CEH's mandate has been considered rather constrained in its impact due to the prohibition against naming individual perpetrators or recommending prosecutions, a limitation often attributed to political pressures aimed at protecting influential figures from accountability[31].

Example 13. *Following a 36-year internal armed conflict that ended in 1996, Guatemala established the Commission for Historical Clarification (La Comisión para el Esclarecimiento Histórico, CEH) to investigate human rights violations and acts of violence (e.g., Chamarbagwala and Morán, 2011; Schwartz and Straus, 2018). Operating from 1997 to 1999, the commission documented over 42,000 victims and attributed 93% of human rights violations to state forces*[32]*. The CEH faced substantial challenges, particularly from conservative factions that sought to undermine its effectiveness (e.g., Tomuschat, 2001). These groups reportedly withheld support and funding, limiting the commission's capacity to conduct thorough investigations. The CEH's mandate, which prohibited naming individual perpetrators or recommending prosecutions, further constrained its impact (e.g., Ross, 2006). This limitation is attributed to political pressures aimed at protecting influential figures from accountability.*

Furthermore, the *ancien régime* elite or conservative forces may push to revise or sanitize the historical narrative of the conflict or authoritarian rule, presenting their actions as legitimate or justified. They may portray themselves as defenders of national integrity or stability, casting TJ efforts as retributive. They may construct a counter-narrative that glorifies the past regime and discredits efforts at historical reckoning. Monuments, public honors, and celebratory historical accounts may serve to resist TJ by cultivating nostalgia for the old order. We illustrate these with examples (14) and (15) from contemporary Turkey and Russia that demonstrate how nationalist and conservative forces actively resist (elements of) TJ efforts by revising historical narratives, suppressing acknowledgment of past atrocities, and framing such initiatives as threats to national identity or

[31] See, for instance, "Guatemala: Historical Background", *ICTJ Case Studies*.
[32] See, for instance, "Truth Commission: Guatemala", *United States Institute of Peace*, February 1, 1997.

sovereignty.

Example 14 (Cleansing Turkey's Dark Past)**.** *In Turkey, nationalist factions have historically resisted efforts to acknowledge the Armenian Genocide, often framing TJ initiatives as attacks on Turkish identity or sovereignty. This resistance is deeply rooted in the state's official narrative, which denies the characterization of the 1915 events as genocide. The Turkish government acknowledges the deaths of Armenians during World War I but disputes that these were part of a systematic genocide, attributing the casualties to the broader context of war and internal strife*[33]*. This denial is reinforced through various means, including legal actions against those who publicly recognize the genocide. For instance, in October 2024, Turkey's radio and television watchdog, RTUK, canceled the terrestrial broadcast license of Open Radio (Acik Radyo) after a guest referred to the 1915 events as genocide during a broadcast. RTUK deemed the remarks as incitement to hatred, illustrating the state's intolerance toward genocide acknowledgment*[34]*. Additionally, public commemorations of the Armenian Genocide have faced governmental suppression. Between 2010 and 2019, the April 24 Commemoration Platform organized public events in Istanbul to commemorate the genocide. However, in 2022 and 2023, these commemorations were banned outright by the Istanbul governor's office, reflecting ongoing resistance to public acknowledgment of historical injustices*[35]*.*

Example 15 (Erasing Russia's Dark Past)**.** *In post-Soviet Russia, certain elites have downplayed or justified Stalinist repression, opposing efforts to critically engage with Soviet-era crimes. Under President Vladimir Putin's administration, there has been a notable shift toward re-Stalinization, characterized by the rehabilitation of Stalin's image and the suppression of organizations that document Soviet-era atrocities. This approach serves to foster national pride and continuity with the Soviet past, while marginalizing critical perspectives*

[33] See, for example, "The Armenian Allegation of Genocide: The issue and the facts", *Republic of Türkiye Ministry of Foreign Affairs*.

[34] See, for example, "Turkey radio station closed after suspension over Armenian genocide remark", *Reuters*, October 17, 2024.

[35] See, for example, "Confronting the Armenian Genocide in Turkey: A TJ Approach", *Kurdish Peace Institute*, April 24, 2023.

on historical repression. A recent manifestation of this trend is the revocation of rehabilitations granted to Stalin's victims. In 2021, Russian authorities ordered the closure of Memorial, a human rights group internationally recognized for its work in documenting Stalinist purges and preserving the memory of political repression. This decision marked a turning point in the state's campaign to suppress critical engagement with Soviet history, signaling an intention to silence independent historical narratives[36]*. Another example of state resistance to historical reckoning is the removal of plaques from the "Last Address" (Posledniy Adres) initiative which conceptually resembels Stolperstein. The "Last Address" project, started in 2014, aimed to commemorate victims of political repression by placing small plaques on the buildings where they had lived before their arrests. The plaques provide brief details about the individual's name, date of birth, and the circumstances of their death during Stalin's purges. In recent years, these memorial plaques have started disappearing from buildings in Moscow, a reflection of growing hostility toward efforts to preserve the memory of repression* [37]*. Furthermore, The Gulag History Museum, an institution dedicated to preserving the memory of Soviet political repression, was abruptly closed on November 14, 2024, "indefinitely." The official reason given was fire safety concerns, but this move coincides with increasing state intolerance toward initiatives that critically examine the Soviet past. The museum had gained recognition for its innovative approaches to engaging with the tragic history of repression, but such efforts are increasingly at odds with state narratives*[38]*. Alongside this, in December 2024, Russian authorities began rescinding official rehabilitations of individuals previously cleared of charges during Stalin's regime, particularly targeting those labeled as "traitors" or "Nazi collaborators." This move aligns with efforts to control historical narratives and suppress dissent, drawing parallels between contemporary policies and Stalinist practices*[39]*.*

---

[36] See, for example, "Memorial: Russia's civil rights group uncovering an uncomfortable past", *BBC*, October 7, 2022.

[37] See, for example, "The "Last Address" plaques a project to perpetuate the memory of victims of political repression, began to disappear from houses in Moscow.", *Meduza*, June 8, 2023.

[38] See, for example, "Moscow's Gulag Museum forced to close indefinitely amid rumoured Kremlin pressure", *Novaya Gazeta Europe*, November 15, 2024.

[39] See, for example, "Putin Decides That Stalin's Victims Were Guilty After All", *BBC*, December 5, 2024.