American Political Science Review (2022) 1–17
doi:10.1017/S0003055422001162

Strategic Reporting: A Formal Model of Biases in Conflict Data

MICHAEL GIBILISCO *California Institute of Technology, United States*
JESSICA STEINBERG *Indiana University, United States*

During violent conflict, governments may acknowledge their use of illegitimate violence (e.g., noncombatant casualties) even though such violence can depress civilian support. Why would they do so? We model the strategic incentives affecting government disclosures of illegitimate violence in the face of potential NGO investigations, where disclosures, investigations, and support are endogenous. We highlight implications for the analysis of conflict data generated from government and NGO reports and for the emergence of government transparency. Underreporting bias in government disclosures positively correlates with underreporting bias in NGO reports. Furthermore, governments exhibit greater underreporting bias relative to NGOs when NGOs face higher investigative costs. We also illustrate why it is difficult to estimate negative effects of illegitimate violence on support using government data: with large true effects, governments have incentives to conceal such violence, leading to strategic attenuation bias. Finally, there is a U-shaped relationship between NGO investigative costs and government payoffs.

In the chaos of violent conflict, it can be difficult to determine whether combatants abuse civilians and which side perpetrated the abuse. State security forces can present themselves in plain clothes; rebels hide their identities. Air strikes present particularly difficult attribution problems. When one warring party is deemed responsible for illegitimate violence, such as indiscriminate violence, mass rape, or the destruction of critical infrastructure, it can lose the support of civilians who care about the armed actor's battlefield behavior (Benmelech, Berrebi, and Klor 2015; Condra and Shapiro 2012; Lyall, Blair, and Imai 2013). This support is critical for leaders: support among the selectorate ensures political survival (Bueno de Mesquita et al. 1999; Prorok 2016; Weeks 2012), and support among civilians in the conflict zone entails battlefield resources such as recruits, information, or supplies (Kalyvas 2006; Shaver and Shapiro 2021). The result is that, for warring governments in particular, there are strong incentives to conceal unpopular collateral damage.

Despite this, governments can, and sometimes do, preemptively disclose illegitimate violence. The Obama administration, for example, acknowledged many civilian causalities resulting from its use of drone strikes in a targeted-killings program outside active military theater. Likewise, in its conflict with Naxalite rebels, the Indian government started publishing a list of violent encounters in the South Asian Terror Portal, where it originally recorded whether government or rebel forces perpetrated the violence and whether the violence led to noncombatant causalities.[1] Critical to both these examples is the presence of watchdog NGOs. The Bureau of Investigative Journalism began systematically documenting drone-strike civilian casualties in 2011, four years before the Obama administrations published its own list. Likewise, in India, a collection of watchdog NGOs, including the Asian Centre for Human Rights and Forum for Fact Finding, Documentation, and Advocacy, began investigating the veracity of the government's list in 2007. In both cases, these reports differed substantially from the governments' accounts.

Given the incentives to conceal illegitimate violence, how and why does government transparency about this type of violence arise? What are the implications for conflict research that uses government and NGO reports as data?

Scholars often rely on government- or NGO-provided data to study the microfoundations of violence. Yet data from different sources can paint significantly different pictures about the extent of illegitimate violence in a conflict. Given the choice between government- and NGO-provided data, it is not clear which should be more accurate a priori. Furthermore, we do not know how relevant background variables (e.g., government popularity or NGO investigative costs) map onto data quality measures like underreporting bias or affect the ability of researchers to estimate parameters of interest (e.g., the degree to which illegitimate violence suppresses support). It is difficult to assess these considerations empirically because we cannot compare the observed data to some *true* account of the conflict. Consequently, we adopt a formal approach in this paper.

Michael Gibilisco, Assistant Professor, Division of Humanities & Social Sciences, California Institute of Technology, United States, michael.gibilisco@caltech.edu.
Jessica Steinberg, Associate Professor, International Studies, Hamilton Lugar School of Global and International Studies, Indiana University, United States, steinbjf@indiana.edu.

Received: July 21, 2021; revised: March 31, 2022; accepted: October 14, 2022.

[1] Some have argued that the South Asia Terror Portal is media sourced, but at the time of our exploration it was founded, developed, and run by KPS Gill, head of government counterinsurgency at the time.

Specifically, we model how governments and NGOs strategically report illegitimate violence. In the model, a government discloses the legitimacy of violence, where illegitimate violence captures acts that potential supporters would find distasteful (e.g., noncombatant casualties). A watchdog NGO allocates costly effort to investigate the veracity of the government's disclosure, receiving an additional benefit when it exposes a cover-up. Both the government's disclosure and the NGO's investigation affect third-party support for the government, which the government seeks to maximize. The third party represents an actor or group on whose support the government relies, either for wartime information or for support among the selectorate (e.g., civilians in conflict zones or voters in democracies). Optimal disclosures, investigations, and support are described by three equilibria: a truthful equilibrium in which the government correctly discloses the state of violence, a never-admit-fault equilibrium in which the government never discloses illegitimate violence, and a partially truthful equilibrium in which the government mixes between disclosing and concealing illegitimate violence. To motivate our theoretical framework, we draw on two disparate examples: the Naxalite insurgency in India and the US drone-strike program.

Our main contribution is to use the model and its equilibrium characterization to derive implications for conflict research that uses data coded from government disclosures or NGO reports. First, we explore underreporting bias—that is, the difference between the baseline frequency of illegitimate violence and the frequency that a source reports illegitimate violence in equilibrium. Both government and NGO reports have underreporting bias, but the causes differ. Whereas the government has incentives to conceal illegitimate violence, the NGOs may fail to produce tangible results when they invest limited investigative effort. We characterize when government disclosures will have less underreporting bias than NGO reports in equilibrium and vice versa. We show that as NGOs face higher investigative costs, both the NGO and government data will underreport illegitimate violence, but the bias in government data will be more extreme. When investigative costs are large, NGOs invest less effort and are therefore unlikely to expose cover-ups. In exactly this situation, governments have large incentives to conceal illegitimate violence. An implication is that the underreporting bias in both data sources is positively correlated across cases; researchers cannot simply trade a bad information source for a good one.

Second, we study when conflict researchers can correctly identify the third-party's distaste for illegitimate violence. The analysis is motivated by empirical work that estimates the degree to which noncombatant casualities depress popular support for counterinsurgent forces (Lyall, Blair, and Imai 2013; Lyall, Shiraito, and Imai 2015; Shaver and Shapiro 2021). We show that when the government is truthful in equilibrium, studying variation in support after different types of government disclosures correctly estimates the third-party's distaste of illegitimate violence. In contrast, when the government conceals illegitimate violence with positive probability in equilibrium, an identical design underestimates the true effect of illegitimate violence on third-party support. This attenuation bias arises because uninformed third-party observers, anticipating government cover-ups in equilibrium, temper support after seeing a report of legitimate violence relative to the truthful baseline. The magnitude of the bias increases as the true distaste for illegitimate violence increases, and the bias exists even when government reports and third-party support are observed without measurement error.

Third, we explore when governments have incentives to manipulate the environment in which civil society operates. For instance, governments might weaken or strengthen transparency institutions such as freedom of information (FOI) laws or press protections more broadly (Colaresi 2012; Egorov, Guriev, and Sonin 2009; Grigorescu 2003; Lorentzen 2014), which affects the investigative costs of NGOs. We find that decreasing NGO investigative costs increases the likelihood that a cover-up is exposed and, as a second-order effect, makes the government less likely to conceal illegitimate violence. This latter effect creates positive belief spillovers that enhance third-party support via equilibrium beliefs. Therefore, governments benefit from transparency institutions when the second-order effect dominates the first. In particular, moderately strong transparency institutions can leave the government the least well-off because they do not induce the government to truthfully disclose but still help NGOs expose cover-ups.

RELATED LITERATURE

Research on underreporting bias in conflict data focuses on nonstrategic sources: inconsistent media coverage (Hendrix and Salehyan 2015; Weidmann 2015), aggregation bias from combining sources (Cook and Weidmann 2019), or geographic biases associated with cellphone coverage (Weidmann 2016). In contrast, we explore the interdependent strategic forces that determine underreporting bias in government- and NGO-provided data, considering the motivations of each actor when reporting unpopular violence. Thus, our paper relates to Drakos and Gofas (2006), who study how terrorists anticipate media coverage, creating underreporting bias. Whereas their work is largely empirical and focuses on the decision of terrorist groups to attack, this paper is largely theoretical and focuses on the decision of governments to acknowledge illegitimate violence. Our analysis demonstrates how incentives to report and investigate illegitimate violence affect underreporting bias even when the underlying frequency of illegitimate violence is exogenous.

To do this, we construct a formal model similar to those in the literature on auditing and risk disclosure (Avenhaus, Von Stengel, and Zamir 2002; Dobler 2008). Auditing games appear in the study of arms control (Arena and Wolford 2012; Baliga and Sjöström 2008), covert affairs (Spaniel and Poznansky 2018), and

https://doi.org/10.1017/S0003055422001162 Published online by Cambridge University Press

cyberwarfare with attribution problems (Baliga, Bueno de Mesquita, and Wolitzky 2020). Besides our focus on the production and analysis of conflict data, the model below departs from this literature in two noticeable ways. First, whereas previous theoretical work has one uninformed audience, in our model the government (our inspectee) sends a report to two different uninformed audiences with different preferences over outcomes, the NGO (our inspector) and a third-party observer. Ex post, both types of governments benefit from third-party support, but only those perpetrating legitimate violence benefit from NGO investigations. Nonetheless, we show that governments benefit ex ante from stronger NGOs if the resulting investigations are thorough enough to commit governments to the truth.

Second, in our model the government's actions are reports or disclosures. As such, they only affect government payoffs indirectly through endogenous third-party support instead of having a direct effect or altering the structure of strategic interaction. Thus, lying costs are fully endogenous in our model and arise via the observer's distaste for cover-ups and equilibrium beliefs. In contrast, previous work treats lying costs as a black box by appealing to long-term reputational costs (e.g., Crescenzi et al. 2011; Smith 2021). This distinction is important because we show that when lying costs are endogenous, it becomes more difficult to observe them in data generated from the model's equilibrium. The difficulty arises because uninformed third-parties internalize the possibility of cover-ups when the government conceals illegitimate violence on the equilibrium path—even truthful governments are affected by lying costs.

More broadly, our analysis contributes to the literature examining how media shapes regime accountability—see Graber (2003) and Baum and Potter (2008) for reviews. Briefly, a robust civil society provides information to citizens that allows them to hold leaders accountable for unpopular decisions. Even autocrats may adopt partial press freedoms to incentivize local or bureaucratic officials (Egorov, Guriev, and Sonin 2009; Lorentzen 2014) or balance coup threats (Boleslavsky, Shadmehr, and Sonin 2021; Hollyer, Rosendorff, and Vreeland 2019). In the domain of national security, Colaresi (2012) and Bell and Martinez Machain (2018) argue that increasing transparency institutions is a win-win for democratic governments: strong transparency institutions simultaneously satisfy temporary secrecy demands and long-term accountability demands through retrospective oversight by the media.[2] Implicit in these accounts is an assumption that "the media serve primarily as a linkage mechanism rather than as an independent, strategic actor in the policymaking process" (Baum and Potter 2008, 50). In contrast, we treat NGOs as strategic actors who allocate investigative effort according to budgetary pressures and expectations about government behavior. Doing so helps to elucidates the relationship between transparency institutions and actual information disclosed by governments and also accounts for why governments adopt these institutions.

THEORETICAL APPROACH AND ILLUSTRATIVE CASES

Three premises constitute the foundation of our formal model. First, warring parties rely on the support of third-party noncombatants. In civil war, localized support can produce resources or information about the tactical strategies of the opponent (Condra and Shapiro 2012; Kalyvas 2006; Shaver and Shapiro 2021). In international conflict, support from locals in theater provides similar benefits, but leaders in both autocracies and democracies need the support of their selectorate at home to ensure political survival (Bueno de Mesquita et al. 1999; Prorok 2016; Weeks 2012). Over the course of a conflict, a government may, intentionally or otherwise, perpetrate violence that they expect their potential supporters to find abhorrent, should they learn of it. Because violent conflict is messy, potential supporters do not have complete information about the nature of violence. Consequently, governments may attempt to conceal illegitimate violence to ensure continued support in the immediate term.

Second, although governments would like to conceal their use of illegitimate violence, they also wish to avoid getting caught in a cover-up, which could further undermine support. Societies with a minimally free press present the potential that watchdog NGOs or newspapers may also investigate the conflict and, in doing so, they may uncover evidence that contradicts the government's official narrative about its wartime behavior. Watchdog, transparency, or human rights organizations view their primary purpose as gathering and disseminating information, and they derive monetary benefit via donors from doing so. For example, Human Rights Watch and Global Witness rely on donations from the public to continue fact finding and publishing reports on human rights violations both in and outside the context of violent conflict. These benefits might be even larger when their reports contradict information provided by the government. Therefore, when the government has concealed illegitimate violence, it cannot ensure that the violence will not be exposed by this kind of watchdog NGO. Such exposure may cause the government to lose supporters who care about the government's honesty.

Third, NGOs and governments provide different types of information to observers. Governments have more information on the state of violence, given their participation in the event and the chain of command that regularly allows for the transfer of information through the ranks. However, even if the government would like to release verifiable information about the nature of a conflict event, constraints on military intelligence may prevent them from doing so. Nongovernmental organizations do not know the nature of violence at the outset

[2] Colaresi (2012) uses "institutions of oversight" to encompass FOI laws and press protections, whereas we use transparency institutions. They also include national-security legislative oversight powers, which are outside the scope of our analysis.

https://doi.org/10.1017/S0003055422001162 Published online by Cambridge University Press

but need to exert costly effort to investigate, providing verifiable information.[3] This creates an incentive for governments to conceal illegitimate violence that affects their support in the short term, so long as the potential cost of lying is not perceived to be too great in the long run. Of course, NGO investigations may not always be successful. Their likelihood of success depends on the amount of devoted resources, which is strategically allocated according to expectations about what an investigation could find. Although NGOs may be more or less driven by this expectation, they are unlikely to fabricate illegitimate violence because their survival and funding is contingent on their reputation as credible actors.

This theoretical framework captures primary empirical features of cases as disparate as the Naxalite conflict in India and civilian deaths from US drone strikes abroad. These cases motivate our main assumptions and illustrate how the model works on the ground. We do not use the cases to test the model or in a process-tracing exercise. Instead, we rely on them to identify prominent contextual features that we believe a model of strategic reporting should include. Our model is, in that sense, motivated by the cases. The model further provides a precise and internally consistent account of how reporting incentives of governments and NGOs interact, producing a set of intuitive as well as counter-intuitive implications for conflict scholars.

Naxalite Conflict, India

The Naxalite conflict is a Maoist insurgency in India that originated in 1967. Initially, Naxalites were a small, ideological group that split from the main Marxist party in India. In 2004, the conflict escalated from a low-level skirmish with tens of fatalities annually to a full-scale insurgency with thousands of casualties per year. The states most affected were parts of Jharkand, Bihar, Andra Pradesh, Orissa, and most of Chhattissgarh. The escalation was in part due to the emergence of village-level militias, called the Salwa Judum, fighting to support the government. The warring parties—the Naxalites, the Salwa Judum, and the Indian military forces—allegedly engaged in firefights, rape, targeted killings, and destruction of villages. For civilians, it is difficult to differentiate between the Salwa Judum and the Indian military. Moreover, Naxalites are not regularly dressed in recognizable uniforms and belong to no easily identifiable tribe or caste. Consequently, conflict events that occur in the rural regions of Chhattisgarh, for instance, are difficult to attribute although much of the territory outside the urban centers in Chhattisgarh is recognized as Naxal-held territory during the time of interest.

The Indian government considers the Naxalites one of the most dangerous threats to internal security. Consequently, it began maintaining a list of conflict events and associated fatalities in an online platform called the South Asia Terror Portal (SATP), which is widely used by academics and policy makers. The platform initially provided the location (state) of the event, the date, and some contextual details that might have included the event's perpetrator and civilian casualties. These events often include smaller skirmishes after rebels attack police outposts or when government forces patrol villages. The resulting government narrative suggests Naxalites are the primary instigators in months and areas with the largest number of casualties and fatalities.

In 2007, multiple NGOs started investigating the veracity of the SATP between 2005 and 2007 and gathering their own data about the conflict with a focus on Chhattisgarh, the state hardest hit by the insurgency. Through fact-finding missions in suspected areas of violence, the NGOs compiled lists of conflicts, detailing their locations and the names of individual casualties. The NGO data differ significantly from the SATP data, although the latter has greater temporal coverage than the former. Furthermore, in the years since the NGOs' reporting, the SATP has become less detailed, providing only aggregate numbers of deaths on an annual basis.[4] Recall that investigating NGOs rely on external funding resources to operate, some of which is contingent on the truthfulness or revelatory nature of their findings. Because of the large number of NGOs in India, there is significant competition for these resources. Thus NGOs cannot afford to fabricate data outright or they risk damaging their reputation and losing their funding.

The US Drone Strikes and Targeted-Killings Program

After the September 11 attacks, the Bush administration began a secretive targeted-killing program in regions without ongoing hostilities, including Pakistan, Somalia, and Yemen (at the time). Relying primarily on drone strikes to carry out assassinations of alleged terrorists, President Barack Obama continued this program throughout his administration. Information about civilian casualties was originally limited. Unlike the lists of civilian casualties in active military theater, the administration produced no such list for the targeted-killing program. It was not until Obama's second term that the program's existence was even acknowledged. Although the use of assassinations outside of active war zones appears to have recently gained greater acceptance among governments, the noncombatant casualties associated with these tactics remain a point of moral outrage. According to the Pew Research Center (2015), 60% of US citizens support drone strikes targeting extremists,

[3] As described below, the baseline model allows NGOs to provide hard information that verifies either type of violence. In an extension, we consider the possibility that NGOs provide hard information that only verifies illegitimate violence whereas legitimate violence is unverifiable.

[4] It provides the number of events initiated by Naxalites versus those initiated by the government. Earlier SATP versions suggest that the government has much more detailed information about civilian deaths.

https://doi.org/10.1017/S0003055422001162 Published online by Cambridge University Press

but 80% are concerned about the attacks endangering the lives of innocent civilians.

The US government acknowledges that it has verifiable information about the number of fatalities associated with these targeted strikes: "government post-strike reviews involve the collection and analysis of multiple sources of intelligence before, during and after a strike, including video observations, human sources and assets, signals intelligence, geospatial intelligence, accounts for local officials on the ground, and open source reporting" (Director of National Intelligence 2016, 2). It cannot release such information due to military and intelligence constraints. It was only in 2015 that the Obama administration passed an executive order requiring an annual report of the number of noncombatants killed during the program. Two reports were issued; one covers 2009–2015 and another 2016 (Director of National Intelligence 2015; 2016).[5] The reports are vague, providing a single numerical estimate of aggregate noncomabatant fatalities across strikes in all three countries, similar to the current SATP list. Unlike the Naxalite case, the Obama administration refused to acknowledge the existence of the drone program and in doing so withheld information about noncombatant casualties until these reports were issued. Although security hawks may view this as a lie of omission, security doves may view it more nefariously.

Several years before the Obama administration passed the executive order, the Bureau of Investigative Journalism (BIJ), the Long Wars Project, and the New America Foundation began gathering data on the annual number of civilian casualties associated with targeted drone strikes in Pakistan, Somalia, and Yemen. Drawing on media reports and contacts on the ground, these NGOs have each developed a list of strikes, the total number of fatalities, and the number of noncombatant fatalities in each of these countries from 2004 onward. Although these sources differ from each other in their accounting, they differ further still from the two government reports: every independent investigation of drone strikes has found more noncombatant deaths than admitted by the administration (Shane 2015). The NGOs and a variety of media outlets believe that it was this investigative reporting that led Obama to issue the executive order requiring reporting on the targeted-killings program. Although the costs of not revealing this program in the face of NGO reports may be difficult to observe, Obama continues to face criticism from the left, even postpresidency, for failing to justify noncombatant casualties resulting from the drone strikes (e.g., Friedersdorf 2016; Williams 2017).

FORMAL FRAMEWORK

The model consists of three actors: a government, a watchdog NGO, and a third-party observer who is a potential supporter of the government, labeled G, N, and O, respectively. The government is involved in an ongoing violent conflict. At the beginning of the interaction, the current state or type of violence is $v \in \{0, 1\}$. The state $v = 1$ denotes an occurrence of illegitimate violence—for example, violations of human rights, violence against noncombatants, or destruction of critical infrastructure. In contrast, the state $v = 0$ means no illegitimate violence occurred. Initially, the state of violence is known only to the government, and the probability that violence is illegitimate is $\Pr(v = 1) = q$. The parameter q captures at least three sources of illegitimate violence including how often the government chooses to violate laws or norms of war, agency problems between leaders and on-the-ground troops, and mere bad luck. These last two sources are outside the government's purview, so $q > 0$.

After observing state v, the government chooses whether to acknowledge that illegitimate violence occurred (denoted $m = 1$) or not (denoted $m = 0$). The message $m = 1$ corresponds to the government disclosing its use of illegitimate violence to reflect the deaths of civilians, for instance. We interpret $m = 0$ as the business-as-usual message where the government does not update its list of government-perpetrated noncombatant killings or its list of drone strikes with civilian casualties. The government's disclosure decision may not occur immediately after the state of violence is revealed, but this disclosure phase occurs before NGOs can investigate.

Both the third-party and the NGO observe the government's message m. Subsequently, the observer chooses an initial level of support for the government $s_1 \in \mathbb{R}$. The NGO then chooses a level of effort $e \in [0, 1]$ for investigating the state of violence. With probability e, enough information is uncovered to publish a report revealing the type of violence ($r = 1$). With probability $1{-}e$, the investigation fails to uncover enough information to publish ($r = 0$). If a report is published ($r = 1$), then the type of violence v is revealed to the observer. If the report is not released ($r = 0$), then the type of violence remains unknown.[6] In other words, NGOs find and disseminate evidence that verifies either type of violence $v \in \{0, 1\}$.[7] The likelihood that they find such evidence depends on their chosen effort e. The observer then chooses a second level of support $s_2 \in \mathbb{R}$.

Consider how this setup captures the cases. In the Naxalite conflict, government troops regularly encounter rebels when patrolling villages. During these encounters, noncombatants may be targeted or killed by government forces. The third-party observer is the local, noncombatant population that determines the degree to which to support government forces—for example, by providing tactical information that could

[5] Trump ended the practice by executive order.

[6] We assume that the observer sees the NGO report if it is published, but this assumption can be relaxed, e.g., the report is read with a fixed probability. In this version, the equilibrium characterization would not substantively change, but the government would be less likely to disclose illegitimate violence in equilibrium.

[7] Because we focus on verifiable information, there is no possibility that the NGO lies in the model, reflecting the NGO's incentive to maintain legitimacy to secure funding. In Appendix H we consider a version of the model in which only illegitimate violence is verifiable.

https://doi.org/10.1017/S0003055422001162 Published online by Cambridge University Press

be used to defeat the rebels. If noncombatants are killed in an encounter, either accidentally or because they were incorrectly labeled as Naxalites, then the locals may consider the violence to be illegitimate. Initially, only the Indian government knows whether its forces engaged in illegitimate violence. Because war is messy and information is incomplete, locals outside of those immediately affected by the violence do not know its legitimacy. The government decides whether to report noncombatant causalities (i.e., disclose the legitimacy of violence) when updating its list of rebel encounters. The local population then decides whether to lend support to the government—for example, by providing more or less useful tips to the government about the insurgents' tactical operations. At this point, NGOs may decide to investigate whether the government is being truthful in its description of the conflict. After their reports are published, potential supporters may change their level of support based on the findings.

In the US's targeted-killing program, drone strikes may or may not entail noncombatant causalities or the destruction of critical civilian infrastructure. In contrast to the Naxalite case, the observer represents a constituency that decides the degree to which to support the Obama administration at the voting booth or in the court of public opinion. If the administration's goal is to create broad support, then the observer could be a representative US citizen. In contrast, if the goal is to motivate the progressive base, then it could be a representative member of the Democratic Party. The model accommodates either interpretation, although the preferences of the specific observer—which we describe below—would change across interpretations. Strikes that only destroy munitions stockpiles or terrorist cells are likely to be considered legitimate violence, but those that kill noncombatants or destroy hospitals, for example, are less likely to be viewed as legitimate. When these events occur, US citizens have limited information about them. The government, in contrast, has explicitly stated that it knows the civilian cost of each of its targeted attacks (Director of National Intelligence 2016). If citizens gain information about the degree of noncombatant casualties associated with drone strikes through NGO reports, such as those published by the BIJ, then public opinion of the administration or Obama's legacy within the Democratic Party may change.

For payoffs, the government wants to maximize support from the observer—support offered before the NGO publishes and support offered after any potential dishonesty is revealed. Its payoff is

$$u_G(s_1, s_2) = g(s_1) + \delta g(s_2).$$

Above, the function $g : \mathbb{R} \to \mathbb{R}$ maps support into some benefit. The function g is strictly increasing and continuously differentiable with a nonvanishing derivative ($g'(s) > 0$ for all s). These benefits naturally depend on the nature of the conflict and the interpretation of the third-party. In the Naxalite conflict, support comes in the form of tactical information reported to the government as tips from the locals that can be used to defeat the insurgency (Kalyvas 2006; Lyall, Shiraito, and Imai 2015; Shaver and Shapiro 2021). In the drone-strike case, support refers to Obama's poll numbers that can be used as political capital required for reelection or his progressive legacy upon leaving office. The parameter $\delta > 0$ captures the relative importance of timing. If $\delta < 1$, then the government prioritizes immediate support. If $\delta > 1$, then the government prioritizes final support.

The observer has an ideal level of support $\hat{s}$ that depends on the state of violence, the government's message, and a baseline popularity level:

$$\hat{s} = \underbrace{\beta}_{\text{baseline}} - \underbrace{\gamma v}_{\text{dislike of illegitimate violence}} - \underbrace{\kappa \boldsymbol{I}[m = 0, v = 1]}_{\text{dislike of cover ups}}$$

Above, $\beta \in \mathbb{R}$ is the baseline support for the government. The parameter $\gamma > 0$ denotes the observer's distaste for the government using illegitimate violence, and $\kappa > 0$ is the observer's distaste for the government after it hides illegitimate violence.[8] With ideal support level $\hat{s}$, the observer's payoffs are

$$u_O(s_1, s_2) = -(s_1 - \hat{s})^2 - (s_2 - \hat{s})^2,$$

which is the quadratic loss between the chosen support and ideal support.[9]

Finally, the watchdog NGO wants to discover enough information to publish a report subject to some cost of investigating. Its payoff is

$$u_N(e, r; m, v) = \underbrace{(\lambda + (1-\lambda)\boldsymbol{I}[m = 0, v = 1])r}_{\text{benefit of revealing the state v}} - \underbrace{\frac{\rho}{2}e^2}_{\text{effort cost}}.$$

Above, we normalize the NGO's benefit of revealing the state of violence to one but divide this benefit into two components. The first term $\lambda \in (0, 1]$ captures the proportion of benefit from releasing reports regardless of whether there is a cover-up. The second term $1-\lambda$ captures the proportion of benefit that arises from catching the government in a cover-up. If λ is close to zero, then a substantial proportion of NGO publication benefits depend on exposing government cover-ups. If λ is close to 1, then NGO benefits depend on releasing reports regardless of their salaciousness.

The term $\frac{\rho}{2}e^2$ is the cost of exerting effort, and the parameter ρ captures NGO efficiency. For a fixed probability of success, less efficient NGOs (larger ρ) pay higher investigative costs than do more efficient ones. Efficiency likely varies across NGOs, depending

[8] As mentioned above, illegitimate violence might occur via mistakes even though the government's optimal level of illegitimate violence is zero. In this case, we expect γ to be smaller in magnitude than when the government explicitly commits illegitimate violence. When γ is small, the government is more truthful in equilibrium—see Proposition 1 and Implication 2.

[9] For a continuum of observers with potentially heterogeneous preferences, it is possible to interpret the parameters (β, γ, κ) as population averages when the vector of parameters is drawn identically and independently from a distribution that satisfies mild regularity conditions.

on funding and transparency institutions. Better funded NGOs will be more efficient, as they are not likely to face binding budget constraints and thus large opportunity costs, so ρ should be smaller. Likewise, NGOs operating in countries with transparency institutions such as FOI laws and press protections will face lower investigative costs because it is easier to gather information (Colaresi 2012).

Our two cases help to motivate these payoffs. Many watchdog NGOs rely on charitable donations for funding. The size and frequency of these donations relate to their ability to publish visible reports and to their provision of information that differs from the prevailing state narrative. Some NGOs benefit from a *surprise dividend*—that is, additional resources following their revelation of government cover-ups. Not all NGOs are similarly reliant on this surprise dividend and thus may have different preferences, representing cases where λ is closer to one. We expect λ to be small when there is competition among NGOs for attention and donations, as in India, which has one of the largest numbers of NGOs per capita. At least one of the NGOs investigating the Naxalite conflict was particularly resource scarce and relied heavily on a surprise dividend, attempting to be the first to reveal dramatic information contradicting the government's narrative.[10]

Notice that the government potentially trades off initial and final support, where δ captures the relative importance of final support to initial support. In the Naxalite conflict, local support in the Chhattissgarh region might be instrumental for the government's military success. Thus, we might suspect that δ is close to or smaller than one as the government might want to end the conflict as fast as possible. In the US case, δ might be correlated with the time until the next election. If the presidential election is far off, δ would be greater than 1, but if an election is immediate, δ is close zero. In addition, δ could capture the degree to which Obama prioritizes his postpresidential legacy. If this is sufficiently valued, δ would be greater than one.

Strategies and beliefs are straightforward. For the government, a strategy is a function $\sigma_G : \{0,1\} \to [0,1]$, where $\sigma_G(v)$ is the probability that the government admits that it used illegitimate violence after violence state v. For the observer, a strategy is a function $\sigma_O : \{0,1\} \times \{0,1,\varnothing\} \to \mathbb{R}$, where $\sigma_O(m,v)$ is the support O gives the government after message m when the state of violence is unknown ($v = \varnothing$), revealed to be legitimate ($r = 1$ and $v = 0$), or revealed to be illegitimate ($r = 1$ and $v = 1$). Finally, a strategy for the NGO is a function $\sigma_N : \{0,1\} \to [0,1]$, where $\sigma_N(m)$ is the amount of effort N chooses after message m. In addition, μ_m is the belief that conflict involved illegitimate violence after message m—that is, $\mu_m = Pr(v = 1|m)$.

We focus on perfect Bayesian equilibria where beliefs satisfy a version of the D1 criterion, referred to as equilibrium hereafter. Specifically, an equilibrium is an assessment (σ, μ) where (a) $\sigma = (\sigma_G, \sigma_O, \sigma_N)$ is a sequentially rational strategy profile given beliefs $\mu = (\mu_0, \mu_1)$ and (b) beliefs μ are consistent with the strategies and updated via Bayes rule whenever possible. In addition, for any message m not sent with positive probability in equilibrium, the belief μ_m satisfies a version of D1 modified to account for endogenous verification of the sender's type.[11] In the analysis, the refinement removes an equilibrium in which the government always admits to illegitimate violence regardless of v. Given the rarity of governments admitting fault in military combat, this is a substantively appealing criteria.

Before proceeding, it is important to remember that when illegitimate violence occurs, the business-as-usual message represents the government concealing the true state of violence. This concealment may take two forms, however. The government may omit the presence of illegitimate violence or may explicitly claim violence was legitimate. If the interpretation is the former, then we expect the lying costs κ to be comparatively smaller in magnitude than if the interpretation is the latter. This approach is justified for three reasons. First, our substantive implications focus on the frequency with which the government admits illegitimate violence, so the exact nature of the concealment is not a first-order concern. This is similar to empirical work that uses counts or indicators of illegitimate violence. Second, the type of concealment is difficult to classify in our cases. For example, in the Naxalite case, when the government does not report a clash with the rebels that resulted in noncombatant fatalities, this could be interpreted as concealment via omission. Another interpretation would be that, because the government reports no clashes, there could not have been illegitimate violence, which represents concealment via a lie. Third, in a version of the model with three messages—representing acknowledge illegitimate violence, omit discussing violence, and say explicitly no illegitimate violence occurred—the government has peculiar incentives after legitimate violence in nonseperating equilibria. Specifically, it might want to send unexpected messages suggesting that it concealed illegitimate violence, in which case the NGO would have greater incentives to investigate (as it expects a cover-up), which increases the probability that legitimate violence will be exposed and then increases final support for the government. This incentive seems disconnected from our cases where we do not observe governments, after claiming legitimate violence, trying to convince NGOs that there was indeed illegitimate violence to encourage investigations.

ANALYSIS

NGO's effort. To see how NGOs investigate, note that after the government sends message $m = 0$, the probability of a cover-up is μ_0. Then the NGO selects an effort level after message m such that

[10] Author personal correspondence and experience.

[11] We assume that in any subgame after the NGO releases a report ($r = 1$) the observer has correct beliefs (i.e., knows the state) even if the subgame is off the equilibrium path. See the proof of Lemma 1 for details.

$$\max_{e\in[0,\,1]} (\lambda+(1-\lambda)\mathbf{I}[m=0]\mu_0)e-\frac{\rho}{2}e^2$$

and its equilibrium effort therefore takes the form

$$\sigma_N(m)=\frac{\lambda+(1-\lambda)\mathbf{I}[m=0]\mu_0}{\rho}. \qquad (1)$$

If $\lambda<1$, then a proportion of NGO publication benefits depends on exposing government cover-ups. In this case, Equation 1 says that as the NGO expects the government to more frequently conceal illegitimate violence (i.e., μ_0 increases), it allocates more investigative effort because it is more likely to expose a cover-up, which entails $1-\lambda$ of additional benefit. Likewise, because countries with press protections and FOI laws will have smaller investigative costs ρ, Equation 1 says that NGOs in these countries will exert more effort than those in countries without such transparency institutions, all else equal.

Observer's support. When the observer chooses support (either initial s_1 or final s_2) it may not know whether the government concealed illegitimate violence. Also, note that the observer does not have private information and its actions do not influence the NGO's equilibrium incentives in Equation 1. Thus, when the observer does not know the value of v but sees message m, its equilibrium support satisfies

$$\max_{s_t\in\mathbb{R}} -\mu_m(s_t-\beta-\gamma-\kappa(1-m))^2-(1-\mu_m)(s_t-\beta)^2,$$

for $t=1,2$. In contrast, when the observer knows v—for example, when the NGO successfully investigated the government—then the observer can choose its level of support to match $\hat{s}$. Overall, this discussion implies that in equilibrium, third-party support takes the form:

$$\sigma_O(m,v)=\begin{cases}\beta-\gamma\mu_m-\kappa(1-m)\mu_m & \text{if } v=\varnothing\\ \beta & \text{if } v=0\\ \beta-\gamma-\kappa(1-m) & \text{if } v=1\end{cases}. \qquad (2)$$

Equation 2 illustrates why it is difficult to use variation in observed support over time to identify the distaste of cover-ups and illegitimate violence. Suppose violence is illegitimate and the government conceals it by sending message $m=0$. In equilibrium, initial support is $s_1=\beta-(\gamma+\kappa)\mu_0$. If the NGO does not release a report, then final support is also uninformed, $s_2=s_1$, but if the NGO releases a report, then $s_2=\beta-\gamma-\kappa$. After an investigation reveals a cover-up, the change in support is therefore $s_2-s_1=(\gamma+\kappa)(\mu_0-1)$, which is muted by equilibrium beliefs μ_0. When the observer anticipates cover-ups and illegitimate violence, μ_0 is large, so a smaller shift in support occurs than suggested by γ and κ.

Government's message. The government sends message m to maximize its expected benefits given the type of violence v and assessment (σ,μ). The first result says that the government is truthful in equilibrium after legitimate violence ($v=0$).

Lemma 1. *If violence was legitimate, then the government sends the business-as-usual message—that is, $\sigma_G(0)=0$ in every equilibrium (σ,μ). After the business-as-usual message, equilibrium beliefs are*

$$\mu_0=\frac{(1-\sigma_G(1))q}{(1-\sigma_G(1))q+(1-q)},$$

which is strictly decreasing in $\sigma_G(1)$ and weakly increasing in the prior q.

Only after illegitimate violence ($v=1$) does the government have incentives to lie. On the one hand, the government can send a truthful message ($m=1$), thereby avoiding a lie but decreasing support. On the other hand, the government can lie ($m=0$) to increase initial support in hopes that the NGO does not reveal the lie, in which case it enjoys uninformed support in both the immediate and long term.

The expected benefits of lying are thus endogenous to equilibrium behavior. After illegitimate violence, if the government lies by sending the business-as-usual message $m=0$, then its payoff is

$$U_G^{\sigma,\mu}(m=0;v=1)=(1+\delta(1-\sigma_N(0)))\underbrace{g(\sigma_O(0,\ \varnothing))}_{\text{uninformed support}} \qquad (3)$$
$$+\,\delta\sigma_N(0)\underbrace{g(\sigma_O(0,\ 1))}_{\text{informed support}}.$$

In Equation 3, the government potentially receives two different levels of support if it sends the business-as-usual message after illegitimate violence. The observer's first level of support (made before the NGO report) will be uninformed, $\sigma_O(0,\varnothing)=\beta-(\gamma+\kappa)\mu_0$. The second will be informed $\sigma_O(0,1)=\beta-\gamma-\kappa$ with probability $\sigma_N(0)=\frac{\lambda+(1-\lambda)\mu_0}{\rho}$ and will be uninformed with complimentary probability.

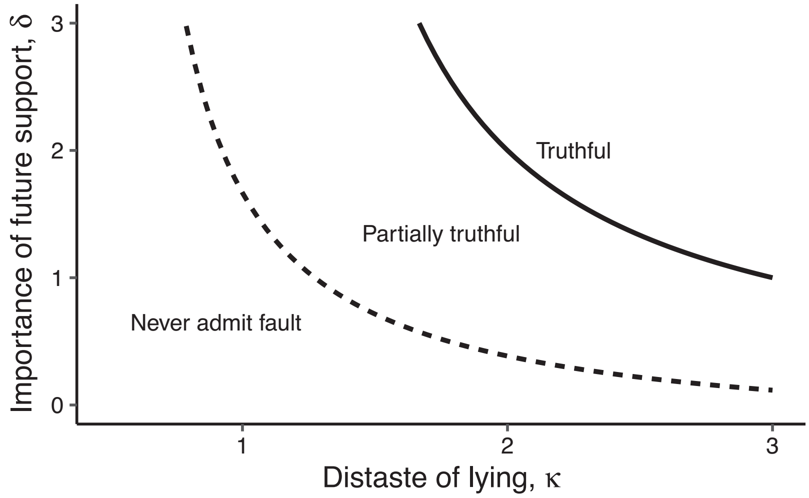
In Equation 3 both the NGO's effort and the observer's uninformed support depend on equilibrium beliefs μ_0. By Lemma 1, $\mu_0\in[0,q]$ is strictly decreasing in $\sigma_G(1)$—that is, the truthfulness of the government. If the government is expected to lie—$\sigma_G(1)$ close to zero—then the NGO exerts more effort to investigate and the observer reduces uninformed support. These forces decrease the government's benefit from lying. If the government is expected to tell the truth—$\sigma_G(1)$ close to one—then the NGO exerts less effort to investigate and the observer increases uninformed support. These forces increase the government's benefit from lying. The following result details how the government balances these trade-offs in equilibrium.

Proposition 1. *The government's behavior is unique in equilibrium:*

1. *The government is truthful—$\sigma_G(v)=v$—in equilibrium if and only if*

$$g(\beta-\gamma-\kappa)\le g(\beta)-\rho\frac{(1+\delta)[g(\beta)-g(\beta-\gamma)]}{\delta\lambda}. \qquad (4)$$

FIGURE 1. Government's Equilibrium Behavior from Proposition 1



Note: Example generated assuming $g(s) = s$, $\gamma = 1$, $\lambda = 0.5$, $\rho = 1$, and $q = 0.2$.

2. *The government never admits fault—$\sigma_G(v) = 0$—in equilibrium if and only if*

$$g(\beta - \gamma - \kappa) \geq g(\beta - (\gamma + \kappa)q) - \rho \frac{(1+\delta)[g(\beta - (\gamma+\kappa)q) - g(\beta - \gamma)]}{\delta(q + (1-q)\lambda)}. \quad (5)$$

3. *The government admits fault after illegitimate violence with probability strictly between zero and one —$\sigma_G(1) \in (0, 1)$—if and only if both inequalities in Equations 4 and 5 are not satisfied.*

Figure 1 illustrates the inequalities in Proposition 1 as functions of the cost of lying and the importance of future support. The main implication for conflict scholars is that as the observer's distaste of lying increases or the government prioritizes long-term rather than immediate support, then the government becomes more truthful in equilibrium. In Appendix H, we illustrate how the equilibrium characterization changes when only illegitimate violence is verifiable rather than both types of violence being verifiable as in the baseline model. The substantive features of the equilibrium characterization do not change, but the government is weakly less truthful in equilibrium.

IMPLICATIONS

Underreporting bias

Scholars often rely on government data because of its temporal span and ease of access. Yet domestic and international NGOs may criticize this data as incomplete or biased in favor of the government. For example, Human Rights Watch provides alternative accounts of the government's use of violence in the Naxalite conflict (Human Rights Watch 2008). In the drone-strike case as well, there are multiple lists recording the extent of illegitimate violence in the conflict. How can researchers or policy makers know which ones to prioritize and when?

To answer these questions, define the probability that actor $i = N,G$ reports illegitimate violence given strategy profile σ:

$$PIV_i(\sigma) = \begin{cases} q\sigma_G(1) + (1-q)\sigma_G(0) & \text{if } i = G \\ q[\sigma_G(1)\sigma_N(1) + (1-\sigma_G(1))\sigma_N(0)] & \text{if } i = N \end{cases}.$$

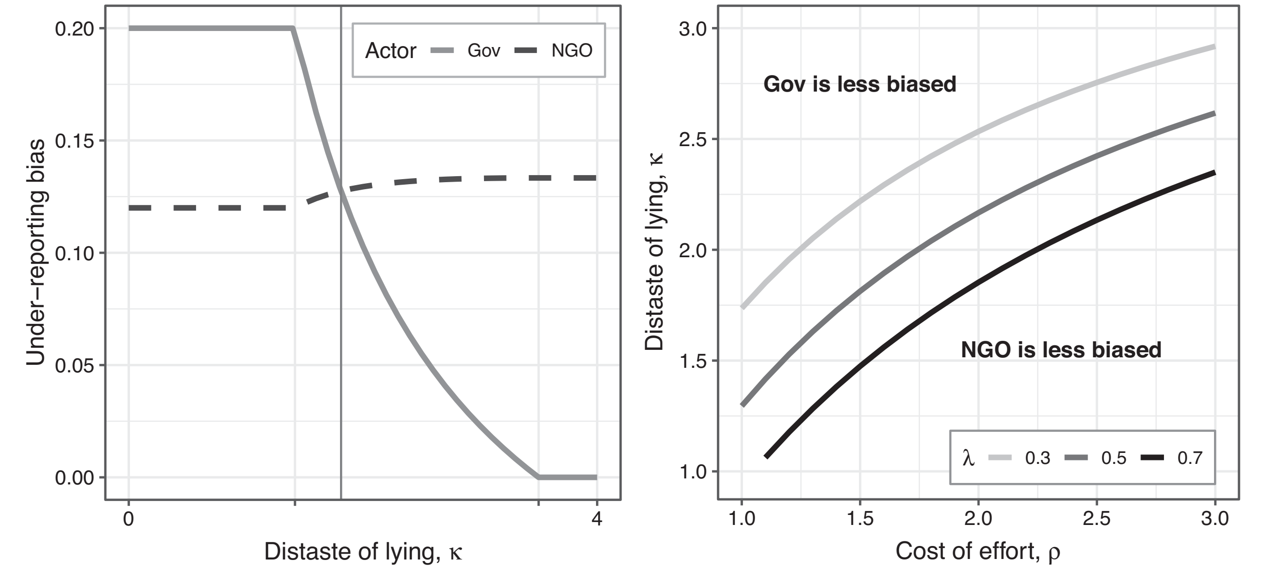
Then i's underreporting bias is $B_i(\sigma) = q - PIV_i(\sigma)$. In words, underreporting bias is the difference between the baseline frequency of illegitimate violence, q, and i's frequency of reporting illegitimate violence PIV_i. In equilibrium, both actors have a tendency to underreport.[12] The government's source of underreporting bias is its incentive to conceal illegitimate violence. The NGO's source of underreporting bias is that it needs to exert costly effort to uncover the truth. Which source does the most damage? To answer this question, we first introduce the following assumption.

Assumption 1. *The benefits of support are sufficiently responsive: there exists $s \in \mathbb{R}$ such that $g(s) < g(\beta) - \rho \frac{(1+\delta)[g(\beta) - g(\beta-\gamma)]}{\delta\lambda}$.*

Comparing Assumption 1 to Equation 4, the assumption says that we can find a distaste of cover-ups, κ, that

[12] Scholars often use conflict event lists to count the number of incidences of illegitimate violence in a given region and period without observing the total number of events. We view these counts as aggregating several draws of outcomes from the equilibrium (σ, μ) that is determined by parameters that are fixed throughout a given region and period. Furthermore, when scholars do not observe the underlying events, it is also difficult to aggregate the lists to improve biases (Cook and Weidmann 2019).

FIGURE 2. Comparison of Government and NGO Underreporting Bias



Note: Left panel graphs the actors' equilibrium level of underreporting bias B_i as a function of κ. Right panel graphs κ^* as a function of ρ and λ. Graphs generated assuming $g(s) = s$, $\gamma = 1$, and $q = 0.2$. In the left panel, we fix $\rho = 1.5$ and $\lambda = 0.5$, implying that $\kappa^* \approx 1.83$.

is large enough to ensure that the government is truthful in equilibrium. The assumption holds if g is concave, for example. The next result describes two cutpoints on the distaste of lying that demarcate the three equilibria.

Lemma 2. *Under Assumption 1, there exist cutpoints* $\bar{\kappa}$, $\underline{\kappa} \in (0, \infty)$ *such that* $\underline{\kappa} < \bar{\kappa}$ *and the following hold in every equilibrium* (σ, μ):

1. *if* $\kappa \geq \bar{\kappa}$, *then the government is always truthful and has underreporting bias* $B_G(\sigma) = 0$;
2. *if* $\kappa \leq \underline{\kappa}$, *then the government never admits fault and has underreporting bias* $B_G(\sigma) = q$; *and*
3. *if* $\kappa \in (\underline{\kappa}, \bar{\kappa})$, *then the government admits its use of illegitimate violence with probability strictly between zero and one and has underreporting bias* $B_G(\sigma) \in (0, q)$.

In the left panel of Figure 2, we graph the underreporting bias for each actor as a function of κ. If the distaste of lying is large enough ($\kappa \geq \bar{\kappa}$), then the government is always truthful. This corresponds to the government having a bias of zero and the NGO having a bias of $q\left(1-\frac{\lambda}{\rho}\right) > 0$. In contrast, if the government's cost of lying is small ($\kappa \leq \underline{\kappa}$), then the government never admits fault. In this case, its bias is q and the NGO's bias is $q\left(1-\frac{\lambda+(1-\lambda)q}{\rho}\right) < q$. In the intermediate range $\kappa \in (\underline{\kappa}, \bar{\kappa})$, the government admits fault after illegitimate violence with probability strictly between zero and one. This probability is strictly increasing in κ, so the government's bias decreases to zero as κ gets larger. As the government becomes more truthful, however, the NGO is less likely to catch the government in a cover-up, so it invests less effort, thereby increasing its bias. As the distaste of cover-ups κ moves from $\underline{\kappa}$ to $\bar{\kappa}$, the government's bias becomes smaller than the NGO's bias at the point κ^*.

Implication 1. *Under Assumption 1, there exists cutpoint* $\kappa^* > 0$ *such that the NGO's underreporting bias is smaller than the government's if and only if* $\kappa < \kappa^*$. *Furthermore,* $\frac{\partial \kappa^*}{\partial \rho} > 0$ *if g is concave and* $\frac{\rho(1-q)\delta\lambda}{q(\rho+\delta(\rho+1-2\lambda))} \geq 1$.

Notice that the cutpoint κ^* can increase as the NGO's cost of effort, ρ, increases, which is illustrated in Figure 2's right panel. In words, the NGO's underreporting bias is more likely to be smaller than the government's (in the set inclusion sense) as the NGO becomes less effective at investigating conflict events.[13] Implication 1 states a sufficient condition for this relationship, which is more likely to hold when q is small and δ is sufficiently large. This captures situations where the government's use of illegitimate violence is not rampant and the government cares about its long-term prospects.

One could imagine the opposite result: greater investigative costs disincentivize NGO effort, thereby making NGO data less reliable relative to government data. This story misses the strategic interplay between the NGO and the government, however. When the costs of investigating increase, two effects emerge in equilibrium. In the direct effect, the NGO invests less effort, leading to greater underreporting bias in NGO data. In the indirect effect, the government anticipates the direct effect and also becomes less truthful, leading to greater underreporting bias in government data. Thus, both data sources become more biased after an

[13] This result also holds when only illegitimate violence is verifiable—see Appendix H.

increase in investigative costs, but the indirect effect dominates under the sufficient condition in Implication 1. In other words, as investigative costs increase, both government and NGO reports will exhibit more underreporting bias, but the effect will be larger for the government. For similar reasons, κ^* can be decreasing in λ—that is, the proportion of NGO publication benefits that depends on releasing information regardless of cover-ups, as illustrated in Figure 2's right panel.

Overall, the analysis suggests two important considerations for conflict researchers. First, it establishes conditions under which NGO data should exhibit less underreporting bias relative to government data: (a) when NGO investigations are relatively inefficient (large ρ) and (b) when NGOs rely heavily on surprise dividends (small λ). As mentioned above, the first condition likely holds when NGOs do not have long-term, robust funding or when NGOs operate in countries without transparency institutions. The second likely holds with substantial competition among NGOs for influence and attention when, for example, there are many NGOs per capita.

Second, underreporting bias in NGO and government data should be positively correlated across cases. When NGOs are substantially underreporting illegitimate violence, there are few incentives for governments to truthfully disclose illegitimate violence, as the likelihood of being exposed in a cover-up is small. Inversely, when NGOs correctly report illegitimate violence, then the government has stronger incentives to tell the truth to avoid cover-ups. Thus, the model suggests that combining government and NGO data will have limited benefits when addressing underreporting bias. When one source consistently misses violence incidents, it is likely the other source will as well.

Illegitimate Violence and Support

Recent work in the counterinsurgency literature estimates the effect of noncombatant casualties on local support for the side responsible. Broadly, state-caused collateral damage can depress support for the government, but the effect is attenuated when examining insurgent-caused damage and support for insurgent groups (Condra and Shapiro 2012; Lyall, Blair, and Imai 2013; Shaver and Shapiro 2021). To measure violence against noncombatants, researchers use self-reported exposure in surveys (e.g., Lyall, Blair, and Imai 2013) or NGO-reported conflict events (e.g., Condra and Shapiro 2012; Shaver and Shapiro 2021). In addition, Lyall, Shiraito, and Imai (2015) measure exposure to violence using data from the International Security Assistant Force, a coalition of NATO governments charged with securing Afghanistan against the Taliban insurgency (842). With these data, they find "no consistent association between indirect exposure to violence and individual attitudes" (844–5).[14] Inspired by these studies and our cases, we compare the observed effect of illegitimate violence on equilibrium support to the true effect when governments strategically report illegitimate violence.

We begin by assuming that researchers observe government messages m and final support s_2 from several draws from one equilibrium.[15] For example, they observe whether the government reports causing collateral damages or not and the resulting level of civilian support. With such data, researchers can compare expected support after the government reports illegitimate violence (noncombatant causalities in this context) to expected support after the government reports no illegitimate violence, all else equal.[16] Definition 1 states this comparison formally.

Definition 1. *Given a strategy profile σ, the observed effect of illegitimate violence using government data is* $\mathrm{E}[s_2|m \neq 1, \sigma] - \mathrm{E}[s_2|m = 1, \sigma] \equiv \Gamma(\sigma)$.

The observed effect underestimates the distaste of illegitimate violence when $\Gamma(\sigma) < \gamma$ and correctly estimates the distaste when $\Gamma(\sigma) = \gamma$.

Table 1 computes the observed effect of illegitimate violence in the truthful and partially truthful equilibria. In the never-admit-fault equilibrium, the government never sends message $m = 1$, so Γ is undefined. The rows enumerate all possible combinations of messages and violence states. Given (m, v), the column $\mathrm{E}[s_2|m, v, \sigma]$ refers to the expected level of support following message m and violence state v. The NA values correspond to pairs (m, v) that never appear on the equilibrium path. In the truthful equilibrium, government disclosures completely reveal its type so observed support includes no distaste of lying. In the partially truthful equilibrium, if the government admits illegitimate violence, then it is truthful and the observer correctly anticipates illegitimate violence. In contrast, if the government sends the business-as-usual message, then it is potentially lying. In this case, unobserved support is biased downward when violence was legitimate but biased upward when violence was illegitimate. Expected support after each message then follows from the law of total expectation.

Thus, the observed effect of illegitimate violence on support correctly estimates γ only when the government is truthful. In the partially truthful equilibrium, however, the observed effect is smaller than the true value because the observer tempers their support after the business-as-usual message because the government may be concealing illegitimate violence. It becomes particularly important to know under what conditions the government will be truthful and the

[14] The study also uses NGO-reported and self-reported exposure to violence. Using the latter, victimization by coalition security forces is associated with a reduction in support to the counterinsurgency in some treatments (Lyall, Shiraito, and Imai 2015, 845).

[15] Our analysis would not change if we used average support, i.e., $\alpha s_1 + (1-\alpha)s_2$ for $\alpha \in [0, 1]$, but focusing on either initial or final support makes the exposition easier.

[16] Civilian support is measured through frequency of informant "tips" in Shaver and Shapiro (2021), attitudes about counterinsurgency informant programs in Lyall, Shiraito, and Imai (2015), and attitudes about coalition forces in Lyall, Blair, and Imai (2013).

https://doi.org/10.1017/S0003055422001162 Published online by Cambridge University Press

TABLE 1. Observed Effect of Illegitimate Violence on Equilibrium Support

	m	v	$\mathbb{E}[s_2 \mid m, v, \sigma]$	$\mathbb{E}[s_2 \mid m, \sigma]$	$\Gamma(\sigma)$
Truthful equilibrium	0	0	β	β	γ
	0	1	NA		
	1	0	NA	$\beta–\gamma$	
	1	1	$\beta–\gamma$		
Partially truthful equilibrium	0	0	$\beta–(\gamma+\kappa)(1–\sigma_N(0))\mu_0$	$\beta–(\gamma+\kappa)\mu_0$	$\gamma–(\gamma+\kappa)\mu_0$
	0	1	$\beta–(\gamma+\kappa)(\sigma_N(0)+(1–\sigma_N(0))\mu_0)$		
	1	0	NA	$\beta–\gamma$	
	1	1	$\beta–\gamma$		

Note: Rows denote all possible message-violence pairs in the truthful (*top*) and partially truthful (*bottom*) equilibria, and NA denotes message-violence pairs that do not emerge in equilibrium. Columns denote the values used to compute the observed effect of illegitimate violence on support Γ. The value Γ is not defined in the never-admit-fault equilibrium.

difference between the true effect, γ, and its observed counterpart, Γ. Recall that by Lemma 2, when g is concave, there exists a $\bar{\kappa} \in$ R such that a truthful equilibrium exists if and only if $\kappa \geq \bar{\kappa}$. So when $\bar{\kappa}$ becomes larger, government disclosures are more likely (in the set inclusion sense) to provide incorrect estimates of γ in the case of the partially truthful equilibrium or infeasible estimates in the case of the never-admit-fault equilibrium.

Implication 2. *Assume $g(s) = s$. Then $\bar{\kappa} = \gamma\left(\frac{(1+\delta)\rho}{\delta\lambda}–1\right)$, and the truthful equilibrium becomes less likely in the set inclusion sense as the distaste for illegitimate violence γ increase—that is, $\frac{\partial\bar{\kappa}}{\partial\gamma} > 0$. Moreover, in the partially truthful equilibrium (σ, μ), the difference between the observed and the true distaste for illegitimate violence $\Delta = \gamma–\Gamma(\sigma)$ is increasing in γ—that is, $\frac{\partial\Delta}{\partial\gamma} > 0$.*

A dilemma thus arises when estimating the effects of illegitimate violence on popular support using government reports. If the distaste of illegitimate violence, γ, is small, then the government is truthful. This means that the observed level of support will not be biased due to strategic reasons (though it might not be easily detectable in smaller samples). If this distaste is large, then the government is unlikely to be truthful, and the observed effect will be biased toward zero. The magnitude of this attenuation increases in the size of the true effect. Likewise, the bias emerges even though government disclosures and popular support are observed without measurement error. The result indicates that the estimates associated with the government-provided data in Lyall, Shiraito, and Imai (2015) could be interpreted as a lower bound on the degree to which civilians punish governments for exposure to violence.

Finally, the result also indicates that the truthful equilibrium is more likely to occur (in the set inclusion sense) when NGOs have small investigative costs, ρ, and the government cares about long-term support—δ is large. Thus, the observed effect Γ should correctly estimate the parameter γ in environments with well-funded NGOs, strong transparency institutions, and governments that prioritize long-term support.

Implication 2 focuses on the relationship between illegitimate violence and observed equilibrium support. Our model also includes a parameter β describing the government's baseline popularity absent illegitimate violence and cover-ups. We can therefore explore the relationship between popularity and the government's propensity to disclose illegitimate violence, which means government transparency can be assessed via government baseline popularity, a potentially observable quantity.

Implication 3. *Assume g is strictly concave. As the government's baseline popularity, β, increases, the following hold:*

1. *The truthful equilibrium becomes less likely in the set inclusion sense—that is, $\frac{\partial\bar{\kappa}}{\partial\beta} > 0$.*
2. *The never-admit-fault equilibrium becomes more likely in the set inclusion sense—that is, $\frac{\partial\underline{\kappa}}{\partial\beta} > 0$, if and only if*

$$\frac{g'(\beta–\gamma–\underline{\kappa})–g'(\beta–(\gamma+\underline{\kappa})q)}{g'(\beta–\gamma)–g'(\beta–(\gamma+\underline{\kappa})q)} > \frac{\rho(1+\delta)}{\delta(q+(1–q)\lambda)}. \tag{6}$$

In other words, the government generally becomes less truthful as its baseline popularity increases *if the decreasing marginal returns to support are sufficiently strong.* Specifically, if g is strictly concave, then higher levels of baseline support imply a smaller set of parameters sustaining the truthful equilibrium (Implication 3.1). In addition, the left-hand side of Equation 6 is a measure of the strength of decreasing marginal returns.

The intuition for this is straightforward. With strong decreasing marginal returns to support, the loss of support that follows an exposed cover-up is more detrimental to a government with low baseline popularity. Thus, the government can more easily afford the costs of lying when it enjoys broad baseline support. Therefore, the truthful equilibrium becomes more difficult to sustain and the never-admit-fault equilibrium becomes easier to sustain as the baseline support increases. Thus, the result suggests that conflict researchers should have greater concerns about underreporting bias from government data when the government enjoys a high baseline popularity from the observer, all else equal.

https://doi.org/10.1017/S0003055422001162 Published online by Cambridge University Press

How and When NGOs Benefit Governments

Scholars have sought to explain why governments create transparency institutions at all given the benefits of controlling information about its behavior (Grigorescu 2003). Several study the variation in FOI laws (Berliner 2014), but transparency institutions include broader legal and regulatory frameworks that facilitate civil society's access to information (as in Egorov, Guriev, and Sonin 2009; Lorentzen 2014). Understanding the drivers and effects of these institutions is especially important in this domain of national security, as governments are especially reticent to disclose information (Colaresi 2012).[17] In the Naxalite case, for example, the government imposed significant costs on NGOs investigating the conflict via the Chhattisgarh Special Public Security Act of 2005. Those convicted of contacting suspected Naxalite rebels faced six years in prison. This policy and others that determine press protections affect investigative costs, ρ, and the equilibrium strategies capture the transparency behavior of the government. Thus, we use the model to study the effects of transparency institutions (via smaller ρ) on government truthfulness and the conditions under which the government would have incentives to manipulate NGO investigative costs through changes to transparency institutions.

Lemma 3. *In the partially truthful equilibrium, the government becomes less likely to disclose illegitimate violence as NGO investigative costs, ρ, increase—that is, $\frac{\partial \sigma_G(1)}{\partial \rho} < 0$ and $\frac{\partial \mu_0}{\partial \rho} > 0$. In the other equilibria, the government's strategy and thus beliefs are constant in ρ.*

Because transparency institutions reduce investigative costs, they encourage government disclosures in the partially truthful equilibrium. This is illustrated in Figure 3's left panel. Notice that transparency institutions are not necessary for government truthfulness, however. Even in their absence, as long as $\rho < \infty$, NGO investigations still expose cover-ups in equilibrium, which means the government could truthfully disclose illegitimate violence when the importance of long-term support, δ, and the distaste of cover-ups, κ, are large.

Recall that uninformed support after message $m = 0$ depends on the probability that the government lied, μ_0. By increasing the equilibrium probability that governments disclose illegitimate violence, efficient NGOs create positive belief spillover effects to governments after legitimate violence via enhanced uninformed support. The next implication states when this effect can increase the government's ex ante expected utility.

Implication 4. *If g is strictly concave, then the following hold:*

1. *In the partially truthful equilibrium, the government's ex ante expected utility is strictly decreasing in NGO investigative costs, ρ, if $\lambda \geq \frac{\rho(1+\delta)-2q\delta}{\delta(1-2q)}$. This inequality always holds if $q \leq \frac{1}{2}$.*
2. *In the never-admit-fault equilibrium, the government's ex ante expected utility is strictly increasing in ρ.*
3. *In the truthful equilibrium, the government's ex ante expected utility is constant in ρ.*

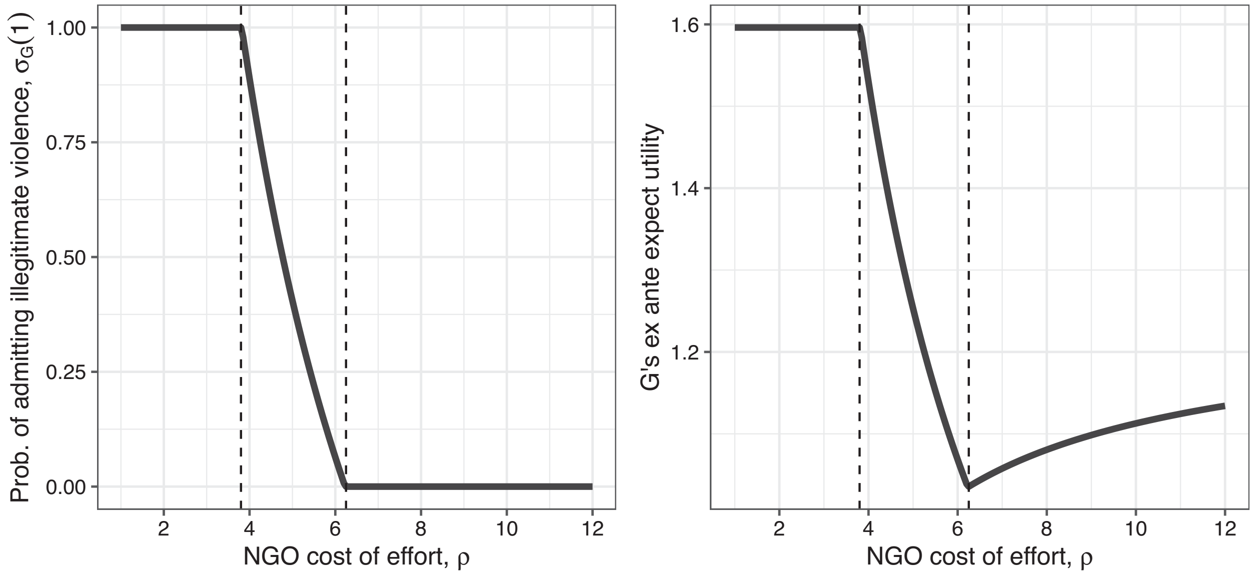
In other words, when g is strictly concave, governments can benefit ex ante from *more* efficient NGOs—that is, smaller ρ produced from transparency institutions—only in the partially truthful equilibrium. Implication 4 states two sufficient conditions for this to happen: the probability of illegitimate violence is small or the NGO is sufficiently motivated to investigate events regardless of the surprise dividend. To see the intuition, notice that, in the partially truthful equilibrium, the government wants to commit to telling the truth ex ante. After illegitimate violence, the government is mixing and thus indifferent between lying and telling the truth, the latter entails a payoff of $(1+\delta)g(\beta - \gamma)$, which is independent of ρ. After legitimate violence, the government sends message $m = 0$ and would like the message to be believed with certainty, and Lemma 3 shows that the message becomes more believable with more efficient NGOs. Thus, increasing the efficiency of NGOs via transparency institutions can weakly increase the expected utility of both types of government as the business-as-usual message $m = 0$ becomes more believable.

In contrast, in the never-admit-fault equilibrium, more efficient NGOs decrease the government's expected payoffs. Here, the government expects three levels of final support: uninformed $s_2 = \beta - (\gamma + \kappa)q$ with probability $1 - \sigma_N(0)$, informed after legitimate violence $s_2 = \beta$ with probability $\sigma_N(0)(1-q)$, and informed after illegitimate violence $s_2 = \beta - \gamma - \kappa$ with probability $\sigma_N(0)q$. As the NGO faces higher investigative costs, there is greater probability that final support will be uninformed, reducing uncertainty from an ex ante perspective. When the government is risk averse (g is strictly concave), this increases the government's expected utility. In the truthful equilibrium, government disclosures remove uncertainty about the type of violence, so the NGO's report and thus the NGO's cost of effort does not affect its payoffs.

Overall, whether the government benefits from transparency institutions and more efficient NGOs depends on exactly how truthful the government becomes after their adoption. If the government does not become very truthful, then the NGO will better expose cover-ups, meaning the government is worse off. If the government becomes very truthful, then uninformed support increases, benefiting the government and minimizing the chances of being caught in a cover-up. These competing effects produce the nonmonotonic relationship between investigative costs ρ and the government's expected payoffs in Figure 3's right panel. Here, the government is worse off with moderately efficient NGOs, $\rho \approx 6$, and would prefer either more or less efficient investigating NGOs.

[17] Absent the capacity to manipulate transparency institutions, officials could use other means of manipulating NGO costs such as attacks against journalists (Carey and Gohdes 2021; Davenport 2009).

FIGURE 3. Effects of NGO Efficiency on the Government's Strategy and Payoffs



Note: Left panel graphs the government's equilibrium probability of truthfully reporting illegitimate violence, $\sigma_G(1)$, as a function of the NGO's cost of effort, ρ. Right panel graphs the government's ex ante expected utility as a function of ρ. Dashed vertical lines demarcate the three types of equilibrium behavior: truthful (small ρ), partially truthful (moderate ρ), and never admit fault (large ρ). Graphs generated assuming $g(s) = \log s$, $\beta = 2$, $\kappa = 1$, $\gamma = 0.95$, $\delta = 2$, and $q = 0.25$.

CONNECTIONS TO OUR CASES

Several implications are born out in our cases. First, recall that NGO and government data exist for the Naxalite conflict, specifically in Chhattisgarh during 2005–07 and that substantial differences arise between lists. Implication 1 suggests that one list is likely to be a more accurate depiction of the illegitimate violence in conflict than the other. In the Naxalite case, this is because at least one of the NGOs was especially interested in catching the government in a lie given the competition for attention and resources among NGOs.[18] This places the Naxalite context along the light-gray line in Figure 2, suggesting that the NGO list is likely to exhibit less underreporting bias than the SATP list of encounters. Regarding the number of civilians killed, we would expect the NGO list to more accurately represent the nature of the conflict.

We can also observe the relationship between baseline popularity and truth telling. India's long-running National Congress Party was in power and widely popular between 2004 and 2008, although it was in a coalition government. Marginal returns to additional support were minimal. Consequently, the resulting loss of support from an exposed cover-up would impose a smaller cost than it would have for a less popular governing party, which can explain the differences between the government and NGO accounts.

Implication 3 helps explain the emergence of the Chhattisgarh Special Public Security Act in 2005, which drastically increased the cost of NGO investigations, ρ, by introducing harsh penalties for affiliating or communicating with suspected Naxalites. Figure 3's left panel illustrates that such a change in ρ can move behavior into a never-admit-fault equilibrium from a partially truthful equilibrium. Figure 3's right panel illustrates that if the initial ρ was not too small and the subsequent increase in ρ was large, then the government is strictly better off by having passed the 2005 Act.

In the US targeted-killings program, we can think of this case as exhibiting two different periods: before the release of any data by the administration and after the release of the first report in 2016. In the first period, only NGO data are available and the government's list is effectively zero, with the administration operating in the never-admit-fault equilibrium. In the later period, it is likely operating in the partially truthful equilibrium, releasing reports that convey only a portion of noncombatant deaths resulting from the drone strikes. This difference in equilibrium could be explained by a change in Obama's time preferences over support, δ. In 2011, δ is arguably small, as Obama was running for reelection and immediate political support was crucial. In the second period, δ increases drastically as Obama becomes more concerned with his legacy than immediate electoral support. Consider Figure 1 to see that an increase in δ could move the government away from the never-admit-fault equilibrium into the partially truthful one.

The effect of baseline popularity is more difficult to interpret in this case. Obama faced lower than average approval ratings throughout most of his presidency; his

[18] Author first-hand experience.

https://doi.org/10.1017/S0003055422001162 Published online by Cambridge University Press

favorability surpassed 50% only during his fourth and seventh years in office. During the second peak in popularity, the administration acknowledges the targeted-killings program and begins reporting associated civilian casualties. This pattern is inconsistent with the comparative statics in Implication 3, which suggests that the government's truthfulness should decrease after an increase in baseline support (assuming g is sufficiently concave). Nonetheless, at the end of his second term, Obama may have prioritized his progressive legacy, in which case truthfulness would have increased due to an increase in δ.

Implication 3 helps us understand the effect of the BIJ, which began systematic data collection in 2010, publishing its first list of casualties in 2011. It was a well-funded actor that selectively chose reporting projects. As such, it constitutes a highly efficient NGO (i.e., it has a small ρ). As illustrated in the left panel of Figure 3, when the costs of NGO investigations decrease, the government's likelihood of disclosing illegitimate violence weakly increases. Furthermore, it is possible that such a change could shift the behavior from the never-admit-fault equilibrium to the partially truthful equilibrium. The overall effect of this shift on the government's expected utility is ambiguous. If such a change sufficiently commits the government to the truth, then it is better off after the BIJ begins its investigations.

CONCLUSION

We explore the mechanisms that lead governments to strategically disclose illegitimate violence and establish implications for the production and analysis of conflict data. We find that both government and NGO reports of illegitimate violence are likely to suffer from underreporting bias, and these biases should be positively correlated across cases. When NGOs face higher investigative costs, they invest less effort in reporting and are therefore less likely to expose cover-ups. At the same time, however, governments will have larger incentives to conceal illegitimate violence.

In addition, we illustrate a dilemma that arises when estimating the effects of collateral damage on civilian support using government reports: if this effect is small and unimportant, then government disclosures are likely to be truthful and the effect can be correctly estimated using standard research designs. If this effect is large and substantial, however, then government disclosures will understate the amount of illegitimate violence, leading to attenuation bias even when reports and support are observed without error. Finally, our analysis suggests that governments will have nonmonotonic preferences over the strength of transparency institutions, where moderately strong institutions leave the government the worst off.

Future research might evaluate how the model applies to domains outside the production and analysis of conflict data. For these applications, we highlight three core assumptions. First, governments release information that is not immediately verified, an assumption best characterizing the national security context where governments cannot reveal hard information without risking a security threat. This might also be true of particular kinds of financial information, the revelation of which might cause significant market volatility. Second, the watchdog NGO or media must be able to produce hard, verifiable information (e.g., pictures of mass graves or videos of noncombatant casualties) that is released to the observer. Our model is therefore not applicable to the study of partisan news organizations if they produce false or unverifiable reports or selectively choose what to report. Third, the government has sufficient certainty about the true state of the world. If, alternatively, the government sees sufficiently noisy signals of the state of the world, then it may hedge against the potential cost of a cover-up by admitting wrongdoing even after seeing a signal that suggests appropriate behavior. These incentives are particularly strong when the distaste of cover-ups and the prior probability of wrongdoing are large.

We note that, although some of our modeling parameters are unlikely to change over the course of a conflict, other parameters—for example, baseline popularity—may vary significantly during a conflict. Variation in these parameters might suggest that some periods may exhibit more or less underreporting bias than others within the same conflict and same dataset. We do not study how forward-looking governments disclose violence today to influence the evolution of their popularity throughout a conflict, although these dynamics could be explored in future research. Likewise, the occurrence of illegitimate violence is exogenous in our model in part because conflict is messy and violence against noncombatants often occurs unintentionally. Nonetheless, future research could endogenize the government's use of illegitimate violence to study how this changes the conditions under which government transparency arises.

Finally, understanding the incentives of warring parties to report the true nature of conflict events has implications for postconflict reconciliation and transitional justice. Recent work has demonstrated that transitional justice initiatives are often ineffective at promoting postconflict peace and reconciliation (Loyle 2018; Loyle and Davenport 2016). This may be because even genuine attempts to implement transitional justice institutions (e.g., tribunals, reintegration policies) are often built on a shared record of violence among warring parties. This shared record is often a compilation of violent conflict events provided by NGOs and the government. As such, it may fail to map accurately onto individual experiences. Understanding the biases that exist in this record may help explain why so many people feel left out of the transitional justice process and why such a record may serve as an ineffective tool for promoting transitional justice.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://doi.org/10.1017/S0003055422001162.

DATA AVAILABILITY STATEMENT

Replication code files for this study are available at the American Political Science Review Dataverse: https://doi.org/10.7910/DVN/PKI60Z.

ACKNOWLEDGMENTS

First Draft: April 2021. Thanks to Emiel Awad, Casey Crisman-Cox, Jacqueline DeMeritt, Alex Hirsch, Phil Hoffman, Kathy Ingram, Federica Izzo, Jin Yeub Kim, Cyanne Loyle, and Brad Smith for comments and discussions. This paper has benefited from audience feedback at MPSA 2021, APSA 2021, and ISA 2022.

CONFLICT OF INTEREST

The authors declare no ethical issues or conflicts of interest in this research.

ETHICAL STANDARDS

The authors affirm this research did not involve human subjects.

REFERENCES

Arena, Philip, and Scott Wolford. 2012. “Arms, Intelligence, and War.” *International Studies Quarterly* 56 (2): 351–65.

Avenhaus, Rudolf, Bernhard Von Stengel, and Shmuel Zamir. 2002. “Inspection Games.” *Handbook of Game Theory with Economic Applications* 3:1947–87.

Baliga, Sandeep, Ethan Bueno de Mesquita, and Alexander Wolitzky. 2020. “Deterrence with Imperfect Attribution.” *American Political Science Review* 114 (4): 1155–78.

Baliga, Sandeep, and Tomas Sjöström. 2008. “Strategic Ambiguity and Arms Proliferation.” *Journal of Political Economy* 116 (6): 1023–57.

Baum, Matthew, and Philip Potter. 2008. “The Relationships between Mass Media, Public Opinion, and Foreign Policy: Toward a Theoretical Synthesis.” *Annual Review of Political Science* 11:39–65.

Bell, Sam R., and Carla Martinez Machain. 2018. “Democracy, Transparency, and Secrecy in Crisis.” *Foreign Policy Analysis* 14 (4): 592–602.

Benmelech, Efraim, Claude Berrebi, and Esteban F. Klor. 2015. “Counter-Suicide-Terrorism: Evidence from House Demolitions.” *Journal of Politics* 77 (1): 27–43.

Berliner, Daniel. 2014. “The Political Origins of Transparency.” *Journal of Politics* 76 (2): 479–91.

Boleslavsky, Raphael, Mehdi Shadmehr, and Konstantin Sonin. 2021. “Media Freedom in the Shadow of a Coup.” *Journal of the European Economic Association* 19 (3): 1782–815.

Bueno de Mesquita, Bruce, James D. Morrow, Randolph M. Siverson, and Alastair Smith. 1999. “An Institutional Explanation of the Democratic Peace.” *American Political Science Review* 93 (4): 791–807.

Carey, Sabine, and Anita Gohdes. 2021. “Understanding Journalist Killings.” *Journal of Politics* 83 (4): 1216–28.

Colaresi, Michael. 2012. “A Boom with Review: How Retrospective Oversight Increases the Foreign Policy Ability of Democracies.” *American Journal of Political Science* 56 (3): 671–89.

Condra, Luke N., and Jacob N. Shapiro. 2012. “Who Takes the Blame? The Strategic Effects of Collateral Damage.” *American Journal of Political Science* 56 (1): 167–87.

Cook, Scott J., and Nils B. Weidmann. 2019. “Lost in Aggregation: Improving Event Analysis with Report-Level Data.” *American Journal of Political Science* 63 (1): 250–64.

Crescenzi, Mark, Kelly Kadera, Sara McLaughlin Mitchell, and Clayton Thyne. 2011. “A Supply Side Theory of Mediation.” *International Studies Quarterly* 55 (4): 1069–94.

Davenport, Christian. 2009. *Media Bias, Perspective, and State Repression: The Black Panther Party*. Cambridge: Cambridge University Press.

Director of National Intelligence. 2015. “Summary of 2009-2015 Information Regarding United States Counterterrorism Strikes outside Areas of Active Hostilities.” Technical Report. Washington, DC: U.S. Government.

Director of National Intelligence. 2016. “Summary of 2016 Information Regarding United States Counterterrorism Strikes outside Areas of Active Hostilities.” Technical Report. Washington, DC: U.S. Government.

Dobler, Michael. 2008. “Incentives for Risk Reporting: A Discretionary Disclosure and Cheap Talk Approach.” *International Journal of Accounting* 43 (2): 184–206.

Drakos, Konstantinos, and Andreas Gofas. 2006. “The Devil You Know but Are Afraid to Face.” *Journal of Conflict Resolution* 50 (5): 714–35.

Egorov, Georgy, Sergei Guriev, and Konstantin Sonin. 2009. “Why Resource-Poor Dictators Allow Freer Media: A Theory and Evidence from Panel Data.” *American Political Science Review* 103 (4): 645–68.

Friedersdorf, Conor. 2016. “Obama’s Weak Defense of His Record on Drone Killings.” *The Atlantic*, December 23. https://www.theatlantic.com/politics/archive/2016/12/president-obamas-weak-defense-of-his-record-on-drone-strikes/511454/.

Gibilisco, Michael, and Jessica Steinberg. 2022. “Replication Data for: Strategic Reporting: A Formal Model of Biases in Conflict Data.” Havard Dataverse. Dataset. https://doi.org/10.7910/DVN/PKI60Z.

Graber, Doris. 2003. “The Media and Democracy: Beyond Myths and Stereoptypes.” *Annual Review of Political Science* 6:139–60.

Grigorescu, Alexandru. 2003. “International Organizations and Government Transparency: Linking the International and Domestic Realms.” *International Studies Quarterly* 47 (4): 643–67.

Hendrix, Cullen S., and Idean Salehyan. 2015. “No News Is Good News: Mark and Recapture for Event Data When Reporting Probabilities Are Less Than One.” *International Interactions* 41 (2): 392–406.

Hollyer, James R., B. Peter Rosendorff, and James Raymond Vreeland. 2019. “Why Do Autocrats Disclose?” *Journal of Conflict Resolution* 63 (6): 1488–516.

Human Rights Watch. 2008. “‘Being Neutral is Our Biggest Crime’: Government, Vigilante, and Naxalite Abuses in India’s Chhattisgarh State.” Technical Report, Human Rights Watch. July 14. https://www.hrw.org/report/2008/07/14/being-neutral-our-biggest-crime/government-vigilante-and-naxalite-abuses-indias.

Kalyvas, Stathis N. 2006. *The Logic of Violence in Civil War*. Cambridge: Cambridge University Press.

Lorentzen, Peter. 2014. “China’s Strategic Censorship.” *American Journal of Political Science* 58 (2): 402–14.

Loyle, Cyanne E. 2018. “Transitional Justice and Political Order in Rwanda.” *Ethnic and Racial Studies* 41 (4): 663–80.

Loyle, Cyanne E., and Christian Davenport. 2016. “Transitional Injustice: Subverting Justice in Transition and Postconflict Societies.” *Journal of Human Rights* 15 (1): 126–49.

Lyall, Jason, Graeme Blair, and Kosuke Imai. 2013. “Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan.” *American Political Science Review* 107 (4): 679–705.

Lyall, Jason, Yuki Shiraito, and Kosuke Imai. 2015. “Coethnic Bias and Wartime Informing.” *Journal of Politics* 77 (3): 833–48.

Pew Research Center. 2015. “Public Continues to Back U.S. Drone Attacks.” May 28. https://www.pewresearch.org/politics/2015/05/28/public-continues-to-back-u-s-drone-attacks/.

Prorok, Alyssa K. 2016. “Leader Incentives and Civil War Outcomes.” *American Journal of Political Science* 60 (1): 70–84.

Shane, Scott. 2015. “Drone Strikes Reveal Uncomfortable Truth: U.S. Is Often Unsure about Who Will Die.” *New York Times*, April 24. nytimes.com/2015/04/24/world/asia/drone-strikes-reveal-uncomfortable-truth-us-is-often-unsure-about-who-will-die.html.

Shaver, Andrew, and Jacob N. Shapiro. 2021. "The Effect of Civilian Casualties on Wartime Informing: Evidence from the Iraq War." *Journal of Conflict Resolution* 65 (7–8): 1337–77.

Smith, Bradley C. 2021. "Military Coalitions and the Politics of Information." *Journal of Politics* 83 (4): 1369–82.

Spaniel, William, and Michael Poznansky. 2018. "Credible Commitment in Covert Affairs." *American Journal of Political Science* 62 (3): 668–81.

Weeks, Jessica L. 2012. "Strongmen and Straw Men: Authoritarian Regimes and the Initiation of International Conflict." *American Political Science Review* 106 (2): 326–47.

Weidmann, Nils B. 2015. "On the Accuracy of Media-Based Conflict Event Data." *Journal of Conflict Resolution* 59 (6): 1129–49.

Weidmann, Nils B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *American Journal of Political Science* 60 (1): 206–18.

Williams, Jennifer. 2017. "From Torture to Drone Strikes: The Disturbing Legal Legacy Obama Is Leaving for Trump." *Vox*. November 14. https://www.vox.com/policy-and-politics/2016/11/14/13577464/obama-farewell-speech-torture-drones-nsa-surveillance-trump.

https://doi.org/10.1017/S0003055422001162 Published online by Cambridge University Press

A APPENDIX (online only)

B Proof of Lemma 1

First, note that if $\sigma_G(0) = 0$ then Bayes rule implies that the probability of illegitimate violence after message $m = 0$ is

$$\mu_0 = \frac{\Pr(m=0|v=1)\Pr(v=1)}{\Pr(m=0)} = \frac{(1-\sigma_G(1))q}{(1-\sigma_G(1))q + (1-q)}.$$

Thus, it suffices to show that $\sigma_G(0) = 0$ in every equilibrium. To do this, we need two intermediate claims.

Claim 1. *In every equilibrium* (σ, μ)*, if* $\sigma_G(0) > 0$*, then* $\mu_0 > 0$ *and* $\sigma_G(1) = 1$*.*

To see this, first note that G's expected utility from sending $m = 1$ after legitimate violence in equilibrium (σ, μ) is:

$$\begin{aligned} U_G^{\sigma,\mu}(m=1; v=0) &= g(\sigma_O(1,\varnothing)) + \delta\left[\sigma_N(1)g(\beta) + (1-\sigma_N(1))g(\sigma_O(1,\varnothing))\right] \\ &\leq g(\sigma_O(1,\varnothing)) + \delta\left[\sigma_N(0)g(\beta) + (1-\sigma_N(0))g(\sigma_O(1,\varnothing))\right]. \end{aligned}$$

The inequality follows because $\beta \geq \sigma_O(1,\varnothing)$ by O's equilibrium condition in Equation 2, and $\sigma_N(0) \geq \sigma_N(1)$ by the NGO's equilibrium condition in Equation 1. Second, note that

G's expected utility from sending $m = 0$ after legitimate violence is:

$$U_G^{\sigma,\mu}(m = 0; v = 0) = g(\sigma_O(0, \varnothing)) + \delta \left[\sigma_N(0)g(\beta) + (1 - \sigma_N(0))g(\sigma_O(0, \varnothing))\right].$$

Because $\sigma_G(0) > 0$ implies $U_G^{\sigma,\mu}(m = 1; v = 0) \geq U_G^{\sigma,\mu}(m = 0; v = 0)$ in equilibrium, the above two inequalities imply $g(\sigma_1(1, \varnothing)) \geq g(\sigma_O(0, \varnothing))$, i.e., $\sigma_1(1, \varnothing) \geq \sigma_O(0, \varnothing)$ as g is strictly increasing. By O's equilibrium condition in Equation 2, this is only possible if

$$-\gamma\mu_1 \geq -(\gamma + \kappa)\mu_0.$$

The above inequality implies that, if $\mu_0 = 0$, then $\mu_1 = 0$. But $\mu_m = 0$ for both messages m is not possible in equilibrium when $q > 0$.

Turning our attention to the government's decision when $v = 1$, if it sends message m its payoff is:

$$\begin{aligned}
U_G^{\sigma,\mu}(m = 1; v = 1) &= g(\sigma_O(1, \varnothing)) + \delta \left[\sigma_N(1)g(\beta - \gamma) + (1 - \sigma_N(1))g(\sigma_O(1, \varnothing))\right] \\
&\geq g(\sigma_O(0, \varnothing)) + \delta \left[\sigma_N(1)g(\beta - \gamma) + (1 - \sigma_N(1)))g(\sigma_O(0, \varnothing))\right] \\
&> g(\sigma_O(0, \varnothing)) + \delta \left[\sigma_N(1)g(\beta - \gamma - \kappa) + (1 - \sigma_N(1)))g(\sigma_O(0, \varnothing))\right] \\
&\geq g(\sigma_O(0, \varnothing)) + \delta \left[\sigma_N(0)g(\beta - \gamma - \kappa) + (1 - \sigma_N(0)))g(\sigma_O(0, \varnothing))\right] \\
&= U_G^{\sigma,\mu}(m = 0; v = 1).
\end{aligned}$$

The first inequality follows because $\sigma_1(1, \varnothing) \geq \sigma_O(0, \varnothing)$, as proved above. The second (strict) inequality follows because $\delta, \sigma_N(m), \kappa > 0$ and g is strictly increasing. The third inequality follows because $\sigma_N(0) \geq \sigma_N(1)$ by N's equilibrium condition in Equation 1. So we have shown $U_G^{\sigma,\mu}(m = 1; v = 1) > U_G^{\sigma,\mu}(m = 0; v = 1)$, which implies $\sigma_G(1) = 1$.

Claim 2. *In every equilibrium (σ, μ), if $\sigma_G(0) > 0$, then $\sigma_G(0) = 1$.*

Proof. If not, then $\sigma_G(0) \in (0, 1)$ some equilibrium (σ, μ). Because $\sigma_G(0) > 0$, Claim 1 implies $\sigma_G(1) = 1$. So governments with legitimate violence $v = 0$ are sending message $m = 0$ with positive probability and the government with illegitimate violence is always sending $m = 1$. So $\mu_0 = 0$, which contradicts Claim 1. □

To prove the Lemma, consider some equilibrium (σ, μ) such that $\sigma_G(0) > 0$. By Claims 1 and 2, $\sigma_G(m) = 1$ for all m, so $\mu_1 = q$. It therefore suffices to argue that when the government is always admitting fault ($\sigma_G(m) = 1$ for all m), the only off-path belief, μ_0, satisfying D1 is $\mu_0 = 0$, which contradicts Claim 1 and establishes the Lemma.

To do this, define

$$\begin{aligned}
EU_G(e, s; v, m' = 0) &= g(s) + \delta \left[eg(\beta - (\gamma + \kappa)v) + (1 - e)g(s)\right] \\
&= (1 + \delta(1 - e))g(s) + \delta eg(\beta - (\gamma + \kappa)v)
\end{aligned}$$

which is the government's utility from sending message $m' = 0$ with violence quality v given it expects effort e and support s when the observer does not know v.[19] Then define

$$WD(v, m' = 0) = \left\{ (e, s) \in \left[\frac{\lambda}{\rho}, \frac{1}{\rho}\right] \times [\beta - \gamma - \kappa, \beta] : EU_G(e, s; v, m' = 0) \geq U_G^{\sigma,\mu}(m = 1; v) \right\}.$$

Above, $U_G^{\sigma,\mu}(m = 1; v)$ is the expected utility of sending message $m = 1$ in equilibrium (σ, μ) such that $\sigma_G(m) = 1$:

$$\begin{aligned} U_G^{\sigma,\mu}(m = 1; v) &= g(\beta - \gamma q) + \delta\left[\sigma_N(1) g(\beta - \gamma v) + (1 - \sigma_N(1)) g(\beta - \gamma q)\right] \\ &= (1 + \delta(1 - \frac{\lambda}{\rho})) g(\beta - \gamma q) + \delta \frac{\lambda}{\rho} g(\beta - \gamma v) \end{aligned}$$

The interval $\left[\frac{\lambda}{\rho}, \frac{1}{\rho}\right]$ is the set of effort levels that can be supported after sending message $m' = 0$ given any beliefs $\mu_0' \in [0, 1]$ when N best responds according to Equation 1.[20] Likewise, the interval $[\beta - \gamma - \kappa, \beta]$ is the set of support that can be generated after message $m' = 0$ given any beliefs $\mu_0' \in [0, 1]$ by Equation 2. Thus, $WD(v, m' = 0)$ is the set of potential best responses that make governments with type v weakly want to deviate to message $m' = 0$ over the equilibrium strategy of always admitting fault. In a similar vein, define

$$SD(v, m' = 0) = \left\{ (e, s) \in \left[\frac{\lambda}{\rho}, \frac{1}{\rho}\right] \times [\beta - \gamma - \kappa, \beta] : EU_G(e, s; v, m' = 0) > U_G^{\sigma,\mu}(m = 1; v) \right\}.$$

So $SD(v, m' = 0)$ is the set of potential best responses that make governments with type v strictly want to deviate to message $m' = 0$ over the equilibrium strategy of always admitting fault.

To show that D1 implies $\mu_0 = 0$, we prove that $WD(1, 0) \subsetneq SD(0, 0)$ (Fudenberg and Tirole 1991, Definition 11.6). That is, there exist rational responses (e, s) that attract governments of type $v = 0$ to deviate to sending message $m' = 0$ but that do not attract governments of type $v = 1$ to deviate.

To see that $WD(1, 0) \subseteq SD(0, 0)$, we show that $(e, s) \in WD(1, 0)$ implies (a) $s > \beta - \gamma q$ and (b) $(e, s) \in SD(0, 0)$. Note that $(e, s) \in WD(1, 0)$ is equivalent to $EU_G(e, s; v = 1, m' = 0) \geq U_G^{\sigma,\mu}(m = 1; v = 1)$. That is:

$$(1 + \delta(1 - e)) g(s) + \delta e g(\beta - \gamma - \kappa) \geq (1 + \delta(1 - \frac{\lambda}{\rho})) g(\beta - \gamma q) + \delta \frac{\lambda}{\rho} g(\beta - \gamma)$$

[19] In $EU_G(e, s; v, m' = 0)$, we are implicitly assuming that, after a successful report revealing the type of violence v, which occurs with probability e, the observer chooses its ideal level of second period support, $s_2 = \beta - (\gamma + \kappa) v$.

[20] Notice we do not consider mixed best responses as Equations 1 and 2 guarantee that the observer and the NGO have unique best responses to every belief $\mu_0' \in [0, 1]$. See Fudenberg and Tirole (1991, 452).

To see that this implies $s > \beta - \gamma q$, suppose not. Then

$$\begin{aligned}
U_G^{\sigma,\mu}(m = 1; v = 1) &= (1 + \delta(1 - \frac{\lambda}{\rho}))g(\beta - \gamma q) + \delta\frac{\lambda}{\rho}g(\beta - \gamma) \\
&\geq (1 + \delta(1 - \frac{\lambda}{\rho}))g(s) + \delta\frac{\lambda}{\rho}g(\beta - \gamma) \\
&> (1 + \delta(1 - \frac{\lambda}{\rho}))g(s) + \delta\frac{\lambda}{\rho}g(\beta - \gamma - \kappa) \\
&\geq (1 + \delta(1 - e))g(s) + \delta e g(\beta - \gamma - \kappa) \\
&= EU_G(e, s; v = 1, m' = 0).
\end{aligned}$$

where the last inequality follows because $e \in \left[\frac{\lambda}{\rho}, \frac{1}{\rho}\right]$ and $s \in [\beta - \gamma - \kappa, \beta]$. Thus, $s \leq \beta - \gamma q$ implies $EU_G(e, s; v = 1, m' = 0) < U_G^{\sigma,\mu}(m = 1; v = 1)$, contradicting $(e, s) \in WD(1, 0)$.

To see that $(e, s) \in WD(1, 0)$ implies $(e, s) \in SD(0, 0)$,

$$\begin{aligned}
U_G^{\sigma,\mu}(m = 1; v = 0) &= (1 + \delta(1 - \frac{\lambda}{\rho}))g(\beta - \gamma q) + \delta\frac{\lambda}{\rho}g(\beta) \\
&< (1 + \delta(1 - \frac{\lambda}{\rho}))g(s) + \delta\frac{\lambda}{\rho}g(\beta) \\
&\leq (1 + \delta(1 - e))g(s) + \delta e g(\beta) \\
&= EU_G(e, s; v = 0, m' = 0).
\end{aligned}$$

Finally, to see that $WD(1, 0) \subsetneq SD(0, 0)$, consider $(e^*, s^*[\epsilon]) = \left(\frac{1}{\rho}, \beta - \gamma q + \epsilon\right)$ where $\epsilon \in (0, \gamma q)$ is small. Using the expected utility calculations above it is straightforward to show that $(e^*, s^*[\epsilon]) \in SD(0, 0)$. We show that $(e^*, s^*[\epsilon]) \notin WD(1, 0)$ for ϵ small enough. To do this, notice that $EU_G(e, s; v = 1, m' = 0)$ is continuous in s and s^* is continuous in ϵ. So $EU_G(e^*, s^*[\epsilon]; v = 1, m' = 0)$ is continuous in ϵ, and it suffices to show that $EU_G(e^*, s^*[0]; v = 1, m' = 0) < U_G^{\sigma,\mu}(m = 1; v = 1)$. This condition holds because

$$\begin{aligned}
U_G^{\sigma,\mu}(m = 1; v = 1) &= (1 + \delta(1 - \frac{\lambda}{\rho}))g(\beta - \gamma q) + \delta\frac{\lambda}{\rho}g(\beta - \gamma) \\
&\geq (1 + \delta(1 - \frac{1}{\rho}))g(\beta - \gamma q) + \delta\frac{1}{\rho}g(\beta - \gamma) \\
&= (1 + \delta(1 - e^*))g(s^*[0]) + \delta e^* g(\beta - \gamma) \\
&> (1 + \delta(1 - e^*))g(s^*[0]) + \delta e^* g(\beta - \gamma - \kappa) \\
&= EU_G(e^*, s^*[0]; v = 1, m' = 0).
\end{aligned}$$

Above, the first inequality follows because $0 < \frac{\lambda}{\rho} \leq \frac{1}{\rho}$ and $-\gamma q > -\gamma$.

C Proof of Proposition 1

Claim 3. *An equilibrium (σ, μ) in which the government is truthful ($\sigma_G(v) = v$) exists if and only if the inequality in Equation 4 holds.*

Proof. If (σ, μ) is a truthful equilibrium, then $\mu_m = m$. After an incidence of illegitimate violence, $v = 1$, if G admits the truth its payoff is

$$U_G^{\sigma,\mu}(m = 1; v = 1) = (1 + \delta)g(\beta - \gamma).$$

If G with type $v = 1$ lies and sends message $m = 0$, its payoff is

$$\begin{aligned} U_G^{\sigma,\mu}(m = 0; v = 1) &= g(\beta) + \delta[\sigma_N(0)g(\beta - \gamma - \kappa) + (1 - \sigma_N(0))g(\beta)] \\ &= (1 + \delta(1 - \sigma_N(0)))g(\beta) + \delta\sigma_N(0)g(\beta - \gamma - \kappa) \\ &= \left(1 + \delta\left(1 - \frac{\lambda}{\rho}\right)\right)g(\beta) + \frac{\delta\lambda}{\rho}g(\beta - \gamma - \kappa) \end{aligned}$$

Above, the second equality follows because $\sigma_O(0, \varnothing) = \beta - (\gamma + \kappa)\mu_0 = 0$ and $\sigma_O(0, 1) = \beta - \gamma - \kappa$. The third follows from the NGO's equilibrium conditions in Equation 1 with $\mu_m = m$. To rule out profitable deviations, we need $U_G^{\sigma,\mu}(m = 1; v = 1) \geq U_G^{\sigma,\mu}(m = 0; v = 1)$, which is equivalent to:

$$g(\beta - \gamma - \kappa) \leq g(\beta) - \rho\frac{(1 + \delta)[g(\beta) - g(\beta - \gamma)]}{\delta\lambda}.$$

Thus, being truthful is incentive compatible for the government after $v = 1$ if and only if Equation 4 holds. To conclude the proof, note that $U_G^{\sigma,\mu}(m = 0; v = 0) = (1 + \delta)g(\beta)$, which is G's largest equilibrium payoff when s_1 and s_2 satisfy 2. So after legitimate violence ($v = 0$), G will never have a profitable deviation from a truthful equilibrium. □

Claim 4. *An equilibrium (σ, μ) in which the government never admits fault ($\sigma_G(v) = 0$) exists if and only if*

$$g(\beta - \gamma - \kappa) \geq g(\beta - (\gamma + \kappa)q) - \rho\frac{(1 + \delta)[g(\beta - (\gamma + \kappa)q) - g(\beta - \gamma)]}{\delta(q + (1 - q)\lambda)}.$$

Proof. We first show that a never-admit-fault equilibrium cannot exist if the inequality does not hold and then argue that never admitting fault is an equilibrium with off-path belief $\mu_1 = 1$ if the inequality holds.

Step 1. Suppose (σ, μ) is a never admit fault equilibrium. Then $\mu_0 = q$, which implies $\sigma_O(0, \varnothing) = \beta - (\gamma + \kappa)q$ by Equation 2. With $v = 1$, the government's payoff from not admitting illegitimate violence is

$$\begin{aligned} U_G^{\sigma,\mu}(m=0; v=1) &= (1+\delta(1-\sigma_N(0)))g(\beta-(\gamma+\kappa)q) + \delta\sigma_N(0)g(\beta-\gamma-\kappa) \\ &= \left(1+\delta\left(1-\frac{\lambda+(1-\lambda)q}{\rho}\right)\right)g(\beta-(\gamma+\kappa)q) + \delta\frac{\lambda+(1-\lambda)q}{\rho}g(\beta-\gamma-\kappa), \end{aligned}$$

where the second equality follows from the NGO's optimal effort level after $m=0$ with beliefs $\mu_0=0$ in Equation 1. The government's payoff from deviating and admitting illegitimate violence is

$$\begin{aligned} U_G^{\sigma,\mu}(m=1; v=1) &= (1+\delta(1-\sigma_N(1)))g(\sigma_O(1,\varnothing)) + \delta\sigma_N(1)g(\beta-\gamma) \\ &= \left(1+\delta\left(1-\frac{\lambda}{\rho}\right)\right)g(\beta-\gamma\mu_1) + \delta\frac{\lambda}{\rho}g(\beta-\gamma) \\ &\geq (1+\delta)\,g(\beta-\gamma) \end{aligned}$$

where the inequality follows because $\sigma_O(1,\varnothing)=\beta-\gamma\mu_1$ is strictly decreasing in $\mu_1\leq 1$. Notice that G has a profitable deviation if

$$(1+\delta)\,g(\beta-\gamma) > U_G^{\sigma,\mu}(m=0; v=1).$$

This condition is equivalent to

$$g(\beta-\gamma-\kappa) < g(\beta-(\gamma+\kappa)q) - \rho\frac{(1+\delta)[g(\beta-(\gamma+\kappa)q)-g(\beta-\gamma)]}{\delta(q+(1-q)\lambda)}.$$

Step 2. Suppose Equation 5 holds. Construct the assessment (σ,μ) as follows: $\sigma_G(v)=0$ and $\mu_1=1$. In addition, $\mu_0=q$ is defined as in Lemma 1, and $\sigma_N(m)$ and $\sigma_O(m)$ follow Equations 1 and 2, respectively. By previous analysis, N and O are best responding to σ_G, and μ_0 is derived via Bayes rule. In addition, the expected utility calculations in Step 1 prove that that G does not have a profitable deviation when $v=1$, $\mu_1=1$, and Equation 5 holds. To see that G does not have a profitable deviation when $v=0$, first note that Equation 5 implies $g(\beta-(\gamma+\kappa)q) > g(\beta-\gamma)$. If not, then we would have $g(\beta-\gamma)\geq g(\beta-(\gamma+\kappa)q)$ and therefore

$$\begin{aligned} g(\beta-\gamma-\kappa) &\geq g(\beta-(\gamma+\kappa)q) - \rho\frac{(1+\delta)[g(\beta-(\gamma+\kappa)q)-g(\beta-\gamma)]}{\delta(q+(1-q)\lambda)} \\ &\geq g(\beta-(\gamma+\kappa)q) > g(\beta-\gamma-\kappa), \end{aligned}$$

a contradiction. Therefore, we can establish that

$$\begin{aligned}
U_G^{\sigma,\mu}(m=0;v=0) &= (1+\delta(1-\sigma_N(0)))g(\beta-(\gamma+\kappa)q)+\delta\sigma_N(0)g(\beta)\\
&\geq (1+\delta(1-\sigma_N(1)))g(\beta-(\gamma+\kappa)q)+\delta\sigma_N(1)g(\beta)\\
&> (1+\delta(1-\sigma_N(1)))g(\beta-\gamma)+\delta\sigma_N(1)g(\beta)\\
&= (1+\delta(1-\sigma_N(1)))g(\sigma_O(1,\varnothing))+\delta\sigma_N(1)g(\sigma_O(1,0))\\
&= U_G^{\sigma,\mu}(m=1;v=0)
\end{aligned}$$

where the weak inequality follows because $g(\beta) > g(\beta-(\gamma+\kappa)q)$, and $\sigma_N(0) \leq \sigma_N(1)$ by Equation 1, and the strict inequality follows from $g(\beta-(\gamma+\kappa)q) > g(\beta-\gamma)$. □

Claim 5. *An equilibrium (σ,μ) in which the government is admits fault after illegitimate with probability strictly between zero and one ($\sigma_G(1) \in (0,1)$) exists if and only if both inequalities in Equations 4 and 5 are not satisfied.*

Proof. In a partially truthful equilibrium (σ,μ) where $\sigma_G(1) > 0$ and $\sigma_G(0) = 0$, $\mu_1 = 1$. Thus, if $v=1$ and G acknowledges illegitimate violence, then its payoff is

$$U_G^{\sigma,\mu}(m=1,v=1) = (1+\delta)g(\beta-\gamma).$$

If G with $v=1$ does not disclose illegitimate violence, its payoff is

$$\begin{aligned}
U_G^{\sigma,\mu}(m=0,v=1) &= (1+\delta(1-\sigma_N(0)))g(\sigma_O(0,\varnothing))+\delta\sigma_N(0)g(\sigma_O(0,1))\\
&= \left(1+\delta\left(1-\frac{\lambda+(1-\lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}\right)\right)g(\beta-(\gamma+\kappa)\tilde{\mu}_0[\sigma_G(1)])\\
&\quad+\ \delta\left(\frac{\lambda+(1-\lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}\right)g(\beta-\gamma-\kappa),
\end{aligned}$$

where $\tilde{\mu}_0[\sigma_G(1)]$ denotes the posterior belief in Lemma 1

$$\tilde{\mu}_0[\sigma_G(1)] = \frac{(1-\sigma_G(1))q}{(1-\sigma_G(1))q+(1-q)}.$$

Notice $\sigma_N(0) = \frac{\lambda+(1-\lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}$ is strictly increasing in $\tilde{\mu}_0$, i.e., the NGO invests more effort if it believes the government lied after sending message $m=0$. In addition, $\sigma_O(0,\varnothing) = \beta-(\gamma+\kappa)\tilde{\mu}_0[\sigma_G(1)]$ is strictly decreasing in $\tilde{\mu}_0$, i.e., the uninformed observer provides less support after message $m=0$ when it believes the government is lying. Because $\sigma_O(0,\varnothing) \geq \sigma_O(0,1) = \beta-\gamma-\kappa$, $U_G^{\sigma,\mu}(m=0,v=1)$ is strictly decreasing in $\tilde{\mu}_0$. Because $\tilde{\mu}_0$ is strictly decreasing in $\sigma_N(0)$, $U_G^{\sigma,\mu}(m=0,v=1)$ is strictly increasing in $\sigma_G(1)$.

Define the function $F:[0,1]\to\mathbb{R}$ as

$$F(x) = U_G^{\sigma,\mu}(m=0,v=1)\big|_{\sigma_G(1)=x} - U_G^{\sigma,\mu}(m=1,v=1).$$

In a partially truthful equilibrium (σ,μ) we must have $F(\sigma_G(1)) = 0$. Furthermore, if $x \in (0,1)$ and $F(x) = 0$, then we can construct a partially truthful equilibrium as follows:

1. $\sigma_G(1) = x$ and $\sigma_G(0) = 0$;
2. $\mu_0 = \tilde{\mu}_0[x]$, $\mu_1 = 1$;
3. σ_N and σ_O follow Equations 1 and 2, respectively.

In this assessment, the government with type $v = 0$ does not have a profitable deviation to send message $m = 1$: $\sigma_N(0) \geq \sigma_N(1)$, and $F(x) = 0$ implies $g(\beta - (\gamma + \kappa)\tilde{\mu}_0[x]) > g(\beta - \gamma)$. The government of type $v = 1$ is indifferent between admitting and covering up illegitimate violence by construction.

Notice that F is continuous and strictly increasing in x by the discussion above. It suffices to show that (a) $F(1) > 0$ is equivalent to the negation of Equation 4 and (b) $F(0) < 0$ is equivalent to the negation of Equation 5. To see the former, note that $\tilde{\mu}_0[1] = 0$. Thus, $F(1) > 0$ is equivalent to

$$\left(1 + \delta\left(1 - \frac{\lambda}{\rho}\right)\right) g(\beta) + \frac{\delta\lambda}{\rho} g(\beta - \gamma - \kappa) - (1 + \delta) g(\beta - \gamma) > 0.$$

Rewriting in terms of $g(\beta - \gamma - \kappa)$ shows that

$$g(\beta - \gamma - \kappa) > g(\beta) - \rho\frac{(1 + \delta)[g(\beta) - g(\beta - \gamma)]}{\delta\lambda},$$

which is the negation of Equation 4. To see the latter, note that $\tilde{\mu}_0[0] = q$, which means $F(0) < 0$ is equivalent to

$$\left(1 + \delta\left(1 - \frac{\lambda + (1 - \lambda)q}{\rho}\right)\right) g(\beta - (\gamma + \kappa)q) + \delta\left(\frac{\lambda + (1 - \lambda)q}{\rho}\right) g(\beta - \gamma - \kappa) - (1 + \delta) g(\beta - \gamma) < 0.$$

Rewriting in terms of $g(\beta - \gamma - \kappa)$ shows that

$$g(\beta - \gamma - \kappa) < g(\beta - (\gamma + \kappa)q) - \rho\frac{(1 + \delta)[g(\beta - (\gamma + \kappa)q) - g(\beta - \gamma)]}{\delta(q + (1 - q)\lambda)},$$

which is the negation of Equation 5. □

Claim 6. *The inequalities in Equations 4 and 5 are mutually exclusive.*

Proof. We need to show

$$g(\beta) - \rho\frac{(1 + \delta)[g(\beta) - g(\beta - \gamma)]}{\delta\lambda} < g(\beta - (\gamma + \kappa)q) - \rho\frac{(1 + \delta)[g(\beta - (\gamma + \kappa)q) - g(\beta - \gamma)]}{\delta(q + (1 - q)\lambda)}.$$

Notice that the right-hand-side is decreasing in $g(\beta)$ because $\frac{\rho(1+\delta)}{\delta\lambda} > 1$. Rewriting the above inequality in terms of $g(\beta)$ means that the inequality holds if and only if $g(\beta)$ is strictly greater than

$$g(\beta-\gamma)\underbrace{\frac{q(1+\delta)(1-\lambda)\rho}{(q(1-\lambda)+\lambda)((1+\delta)\rho-\delta\lambda)}}_{\equiv w_1}+g(\beta-(\gamma+\kappa)q)\underbrace{\frac{\lambda(\rho(1+\delta)-\delta(q(1-\lambda)+\lambda))}{(q(1-\lambda)+\lambda)((1+\delta)\rho-\delta\lambda)}}_{\equiv w_2}$$

Because $g(\beta)>g(\beta-\gamma)$ and $g(\beta)>g(\beta-(\gamma+\kappa)q)$, it suffices to show that $w_1\geq 0$, $w_2\geq 0$, and $w_1+w_2\leq 1$.

To see that $w_k\geq 0$ ($k=1,2$) note that their denominator is positive: $(q(1-\lambda)+\lambda)>0$ (because $\lambda\in(0,1]$ and $q\in(0,1)$) and $((1+\delta)\rho-\delta\lambda)>0$ (because $\delta>0$, $\rho\geq 1\geq\lambda$). As $\lambda\in(0,1]$, the numerator of w_1 is positive. As $\lambda\in(0,1]$, the numerator of w_2 is positive because $\rho\geq 1$ and $q(1-\lambda)+\lambda\in(0,1]$. Therefore w_k is positive. In addition, adding w_1+w_2 shows that $w_1+w_2=1$. □

D Proof of Lemma 2

Throughout the proof, we maintain Assumption 1. To see (1), by Proposition 1 the government is always truthful in equilibrium if and only if Equation 4 holds. Notice the right-hand side of Equation 4 is constant in κ. Because g is strictly increasing and thus $g(\beta-\gamma-\kappa)$ is strictly decreasing in κ, Assumption 1 implies that as $\kappa\to\infty$ the left-hand becomes strictly smaller than the right hand side. Finally, note that

$$\begin{aligned}\lim_{\kappa\to 0}g(\beta-\gamma-\kappa)&=g(\beta-\gamma)\\&>g(\beta-\gamma)\frac{\rho(1+\delta)}{\delta\lambda}+g(\beta)\left[1-\frac{\rho(1+\delta)}{\delta\lambda}\right]\\&=g(\beta)-\rho\frac{(1+\delta)[g(\beta)-g(\beta-\gamma)]}{\delta\lambda}\end{aligned}$$

where the first equality follows because g is continuous and the first inequality follows because $g(\beta-\gamma)<g(\beta)$ and $\frac{\rho(1+\delta)}{\delta\lambda}>1$. Because $g(\beta-\gamma-\kappa)$ is continuous as a function of κ, the intermediate value theorem then implies there exists $\bar{\kappa}>0$ such that

$$g(\beta-\gamma-\bar{\kappa})=g(\beta)-\rho\frac{(1+\delta)[g(\beta)-g(\beta-\gamma)]}{\delta\lambda}.$$

Because $g(\beta-\gamma-\kappa)$ is strictly decreasing in κ, $\bar{\kappa}$ is unique and $g(\beta-\gamma-\kappa)\leq g(\beta-\gamma-\bar{\kappa})$ if and only if $\kappa\geq\bar{\kappa}$.

To see (2), by Proposition 1 the government is never admitting fault in equilibrium if and only if Equation 5 holds. We can rewrite this condition as $D(\kappa)\geq 0$ where

$$\begin{aligned}D(\kappa)&=g(\beta-\gamma-\kappa)-g(\beta-(\gamma+\kappa)q)+\rho\frac{(1+\delta)[g(\beta-(\gamma+\kappa)q)-g(\beta-\gamma)]}{\delta(q+(1-q)\lambda)}\\&=g(\beta-\gamma-\kappa)+(c-1)\cdot g(\beta-(\gamma+\kappa)q)-c\cdot g(\beta-\gamma)\end{aligned}$$

and

$$c \equiv \frac{\rho(1+\delta)}{\delta(q+(1-q)\lambda)} > 1,$$

We first argue that $D(0) > 0$. To see this, note that

$$\begin{aligned} D(0) &= (1-c)\cdot g(\beta-\gamma) + (c-1) \\ &= (c-1)\left[g(\beta-\gamma q) - g(\beta-\gamma)\right] \end{aligned}$$

which is greater than zero because $c > 1$ and $g(\beta-\gamma) < g(\beta-\gamma q)$. Second, we argue that there exists $\kappa > 0$ such that $D(\kappa) < 0$. To see this, because g is strictly increasing, we can bound $D(\kappa)$ from above

$$\begin{aligned} D(\kappa) &\leq g(\beta-(\gamma+\kappa)q) + (c-1)\cdot g(\beta-(\gamma+\kappa)q) - c\cdot g(\beta-\gamma) \\ &= c\left[g(\beta-(\gamma+\kappa)q) - g(\beta-\gamma)\right]. \end{aligned}$$

The term $[g(\beta-(\gamma+\kappa)q) - g(\beta-\gamma)]$ is negative for $\kappa > \gamma\frac{1-q}{q}$. Because D is continuous, the intermediate value theorem implies there exists $\underline{\kappa} > 0$ such that $D(\underline{\kappa}) = 0$. Because D is strictly decreasing, $\underline{\kappa}$ is unique and $\kappa \leq \underline{\kappa}$ if and only if $D(\kappa) \geq 0$.

Notice we have proved $\kappa \geq \bar{\kappa}$ is equivalent to Equation 4 and $\kappa \leq \underline{\kappa}$ is equivalent to Equation 5. Thus $\underline{\kappa} < \bar{\kappa}$ because we have already proved that the two Equations contain mutually exclusive inequalities—see Claim 6. So by Proposition 1, the government admits fault after illegitimate violence with probability strictly between zero and one ($\sigma_G(1) \in (0,1)$) if and only if $\kappa \in (\underline{\kappa}, \bar{\kappa})$.

E Proof of Implication 1

Recall that, in equilibrium, G is always truthful if legitimate violence is used, i.e., $\sigma_G(0) = 0$. G may lie after illegitimate violence however. Using Lemma 2, we can write G's equilibrium probability of admitting to illegitimate violence as a function of κ:

$$S_G(\kappa) = \begin{cases} \{0\} & \text{if } \kappa < \underline{\kappa} \\ \{x \in \mathbb{R} : F(x,\kappa) = 0\} & \text{if } \kappa \in (\underline{\kappa}, \bar{\kappa}) \\ \{1\} & \text{if } \kappa > \bar{\kappa}, \end{cases}$$

where F is defined in the proof of Proposition 1 (see Claim 5):

$$\begin{aligned} F(x,\kappa) = &\left(1+\delta\left(1-\frac{\lambda+(1-\lambda)\tilde{\mu}_0[x]}{\rho}\right)\right) g(\beta-(\gamma+\kappa)\tilde{\mu}_0[x]) \\ &+ \delta\left(\frac{\lambda+(1-\lambda)\tilde{\mu}_0[x]}{\rho}\right) g(\beta-\gamma-\kappa) - (1+\delta)g(\beta-\gamma). \end{aligned}$$

Because g is C^1, F is C^1 as its partial derivatives exist and are continuous. Furthermore, F is strictly increasing in x and $\frac{\partial F}{\partial x}(x,\kappa) > 0$. Specifically,

$$\begin{aligned}\frac{\partial F}{\partial x} = \frac{\tilde{\mu}_0'[x]}{\rho}\,&[\delta(\lambda-1)(g(\beta-(\gamma+\kappa)\tilde{\mu}_0[x])-g(\beta-\gamma-\kappa))\\&-(\gamma+\kappa)(\rho+\delta(\rho-\lambda)-\delta(1-\lambda)\tilde{\mu}_0[x])g'(\beta-(\gamma+\kappa)\tilde{\mu}_0[x])]\,,\end{aligned}$$

where $\tilde{\mu}_0'[x] = \frac{q(q-1)}{(1-qx)^2} < 0$ These properties are sufficient conditions in the implicit function theorem, which we make use of here.

Claim 7. *S_G is a continuous, weakly increasing function of κ. If $\kappa \in (\underline{\kappa},\bar{\kappa})$, then S_G is continuously differentiable at κ and $\frac{\partial S_G}{\partial \kappa} > 0$.*

Proof. First, by Lemma 2, $\kappa \in (\underline{\kappa},\bar{\kappa})$ is equivalent to neither inequality in Equations 4 nor 5 holding. So $\kappa \in (\underline{\kappa},\bar{\kappa})$ implies that the government is mixing after illegitimate violence and the equation $F(x,\kappa) = 0$ characterizes the mixing probability. Thus, $S_G(\kappa) \neq \emptyset$. In Claim 5, we proved $F(x,\kappa)$ is strictly increasing in x, so $\kappa \in (\underline{\kappa},\bar{\kappa})$ implies $|S_G(\kappa)| = 1$. So S_G is a function.

To see that S_G is continuous, note that F satisfies the sufficient conditions of the implicit function theorem. As such, S_G is C^1 and therefore continuous at every $\kappa \in (\underline{\kappa},\bar{\kappa})$. We need to verify that S_G is continuous at $\underline{\kappa}$ and $\bar{\kappa}$. Note that $\lim_{\kappa\to\underline{\kappa}^-} S_G(\kappa) = 0$ and $\lim_{\kappa\to\bar{\kappa}^+} S_G(\kappa) = 1$. So we need to verify (a) $\lim_{\kappa\to\underline{\kappa}^+} S_G(\kappa) = 0$ and (b) $\lim_{\kappa\to\bar{\kappa}^-} S_G(\kappa) = 1$. To do this, we show (a′) $F(0,\underline{\kappa}) = 0$ and (b′) $F(1,\bar{\kappa}) = 0$, respectively.

To see (a′), note that $\tilde{\mu}_0[0] = q$, so we can write $F(0,\underline{\kappa})$ as

$$\left(1+\delta\left(1-\frac{\lambda+(1-\lambda)q}{\rho}\right)\right)g(\beta-(\gamma+\underline{\kappa})q)+\underbrace{\delta\left(\frac{\lambda+(1-\lambda)q}{\rho}\right)g(\beta-\gamma-\underline{\kappa})}_{\equiv W}-(1+\delta)g(\beta-\gamma).$$

Focusing on W, recall $D(\underline{\kappa}) = 0$ means $g(\beta-\gamma-\underline{\kappa}) = g(\beta-(\gamma+\underline{\kappa})q)-\rho\frac{(1+\delta)[g(\beta-(\gamma+\underline{\kappa})q)-g(\beta-\gamma)]}{\delta(q+(1-q)\lambda)}$. So we can write

$$\begin{aligned}W &= \delta\left(\frac{\lambda+(1-\lambda)q}{\rho}\right)\left[g(\beta-(\gamma+\underline{\kappa})q)-\rho\frac{(1+\delta)[g(\beta-(\gamma+\underline{\kappa})q)-g(\beta-\gamma)]}{\delta(q+(1-q)\lambda)}\right]\\&= \delta\left(\frac{\lambda+(1-\lambda)q}{\rho}\right)g(\beta-(\gamma+\underline{\kappa})q)-(1+\delta)g(\beta-(\gamma-\underline{\kappa})q)+(1+\delta)g(\beta-\gamma).\end{aligned}$$

Subsisting W into the original expression proves that $F(0,\underline{\kappa}) = 0$.

To see (b′), note that $\tilde{\mu}_0[1] = 0$, so we can write $F(1,\bar{\kappa})$ as

$$\left(1+\delta\left(1-\frac{\lambda}{\rho}\right)\right)g(\beta)+\frac{\delta\lambda}{\rho}g(\beta-\gamma-\bar{\kappa})-(1+\delta)g(\beta-\gamma).$$

Substituting $g(\beta-\gamma-\bar{\kappa}) = g(\beta)-\rho\frac{(1+\delta)[g(\beta)-g(\beta-\gamma)]}{\delta\lambda}$ proves the result.

Finally, to see that S_G is continuous differentiable and weakly decreasing, consider some $\kappa \in (\underline{\kappa}, \bar{\kappa})$. By the implicit function theorem, $\frac{\partial S_G}{\partial \kappa}$ exists and is continuous. Furthermore,

$$\frac{\partial S_G}{\partial \kappa} = -\frac{\frac{\partial F}{\partial \kappa}}{\frac{\partial F}{\partial x}}.$$

As described above, denominator is positive. To sign the numerator, differentiate $F(x, \kappa)$ with respect to κ:

$$\begin{aligned}\frac{\partial F}{\partial \kappa}(x, \kappa) = &- \overbrace{\left(1 + \delta\left(1 - \frac{\lambda + (1-\lambda)\tilde{\mu}_0[x]}{\rho}\right)\right)}^{>0} \overbrace{\tilde{\mu}_0[x]}^{>0} \overbrace{g'(\beta - (\gamma + \kappa)\tilde{\mu}_0[x])}^{>0} \\ &- \underbrace{\delta\left(\frac{\lambda + (1-\lambda)\tilde{\mu}_0[x]}{\rho}\right)}_{>0} \underbrace{g'(\beta - \gamma - \kappa)}_{>0}\end{aligned}$$

Above, $\tilde{\mu}_0[x] > 0$ because $\kappa \in (\underline{\kappa}, \bar{\kappa})$ implies that a solution x to $F(x, \kappa) = 0$ must be $x \in (0, 1)$. In addition, $g'(s) > 0$ for all s because g is strictly increasing with a non-vanishing derivative. As such $\frac{\partial F}{\partial \kappa} < 0$, implying that $\frac{\partial S_G}{\partial \kappa} > 0$. □

In equilibrium after the government admits to illegitimate violence ($m = 1$), the NGO knows the government is truthful (Lemma 1) and invests effort $\frac{\lambda}{\rho}$ (Equation 1). After the government sends the business as usual message, the NGO's effort can be written as a function of κ via Equation 1 and the previous claims:

$$S_N(\kappa) = \frac{\lambda + (1-\lambda)\tilde{\mu}_0[S_G(\kappa)]}{\rho}.$$

Claim 8. *There exists $\kappa^* \in (\underline{\kappa}, \bar{\kappa})$ such that $B_N(\sigma) < B_G(\sigma)$ if and only if $\kappa < \kappa^*$.*

Proof. In equilibrium, $\sigma_G(0) = 0$, and we can write G's bias as a function of κ:

$$B_G(\kappa) = \begin{cases} q & \kappa \leq \underline{\kappa} \\ q(1 - S_G(\kappa)) & \kappa \in (\underline{\kappa}, \bar{\kappa}) \\ 0 & \kappa \geq \bar{\kappa} \end{cases}$$

Notice B_G is weakly decreasing, continuous, and ranges from q to 0. We can write N's bias as

$$B_N(\kappa) = \begin{cases} q\left(1 - \frac{\lambda + (1-\lambda)q}{\rho}\right) & \kappa \leq \underline{\kappa} \\ q\left(1 - S_G(\kappa)\frac{\lambda}{\rho} - (1 - S_G(\kappa))S_N(\kappa)\right) & \kappa \in (\underline{\kappa}, \bar{\kappa}) \\ q\left(1 - \frac{\lambda}{\rho}\right) & \kappa \geq \bar{\kappa}, \end{cases}$$

which is weakly increasing, continuous. $B_G(\underline{\kappa}) - B_N(\underline{\kappa}) = q\frac{\lambda + (1-\lambda)q}{\rho} > 0$, and $B_G(\bar{\kappa}) - B_N(\bar{\kappa}) = -q(1 - \frac{\lambda}{\rho}) < 0$. Because $B_G(\kappa) - B_N(\kappa)$ is continuous there exists $\kappa^* \in (\underline{\kappa}, \bar{\kappa})$ such that $B_G(\kappa^*) = B_N(\kappa^*)$. Because $B_G(\kappa) - B_N(\kappa)$ is strictly decreasing on the interval $(\underline{\kappa}, \bar{\kappa})$, κ^* is unique and $\kappa < \kappa^*$ if and only if $B_G(\kappa) > B_N(\kappa)$. □

Claim 9. *If $\kappa \in (\underline{\kappa}, \bar{\kappa})$, then $\frac{\partial S_G}{\partial \rho} < 0$.*

Proof. To see this, first note that:

$$\frac{\partial F}{\partial \rho}(x, \kappa) = \frac{\delta}{\rho^2}\left[g(\beta - (\gamma + \kappa)\tilde{\mu}_0[x]) - g(\beta - \gamma - \kappa)\right](\lambda + (1 - \lambda)\tilde{\mu}_0[x]) > 0.$$

Second, $\kappa \in (\underline{\kappa}, \bar{\kappa})$ implies that the solution x^* such that $F(x^*, \kappa) = 0$ will be interior, i.e., $x^* < 1$. If $x^* < 1$, then $\tilde{\mu}_0[x^*] > 0$. Thus, $\frac{\partial F}{\partial \rho}(x^*, \kappa) > 0$ at any solution x^* such that $F(x^*, \kappa) = 0$. We then invoke the implicit function theorem:

$$\frac{\partial S_G}{\partial \rho} = -\frac{\frac{\partial F}{\partial \rho}}{\frac{\partial F}{\partial x}} < 0,$$

where the inequality follows because $\frac{\partial F}{\partial x} > 0$. Furthermore, using the definitions of $\frac{\partial F}{\partial \rho}$ and $\frac{\partial F}{\partial x}$, $\frac{\partial S_G}{\partial \rho}$ takes the form:

$$\frac{\partial S_G}{\partial \rho} = -\frac{\delta\Delta^-(\lambda + (1 - \lambda)\tilde{\mu}_0[x])}{\rho\tilde{\mu}_0'[x](\delta\Delta^-(\lambda - 1) - (\gamma + \kappa)(\rho + \delta(\rho - \lambda) - \delta(1 - \lambda)\tilde{\mu}_0[x])g'(\beta - (\gamma + \kappa)\tilde{\mu}_0[x])}$$

where $\Delta^- \equiv g(\beta - (\gamma + \kappa)\tilde{\mu}_0[x]) - g(\beta - \gamma - \kappa)$. □

Claim 10. *If g is concave and*

$$\frac{\rho(1 - q)\delta\lambda}{q(\rho + \delta(\rho + 1 - 2\lambda))} \geq 1,$$

then $\frac{\partial \kappa^*}{\partial \rho} > 0$.

Proof. By construction, at $\kappa^* \in (\underline{\kappa}, \bar{\kappa})$ $B_G(\sigma) = B_N(\sigma)$ in equilibrium (σ, μ). This is equivalent to

$$\begin{aligned}
B_G(\sigma) = \mathrm{B}_N(\sigma) &\iff q(1 - \sigma_G(1)) = q(1 - [\sigma_G(1)\sigma_N(1) + (1 - \sigma_G(1))\sigma_N(0)]) \\
&\iff \sigma_G(1) = \sigma_G(1)\sigma_N(1) + (1 - \sigma_G(1))\sigma_N(0) \\
&\iff \sigma_G(1) = \frac{\sigma_N(0)}{1 + \sigma_N(0) - \sigma_N(1)} \\
&\iff S_G(\kappa^*) = \frac{S_N(\kappa^*)}{1 + S_N(\kappa^*) - \frac{\lambda}{\rho}} \\
&\iff S_G(\kappa^*) - \frac{\lambda + (1 - \lambda)\tilde{\mu}_0[S_G(\kappa^*)]}{\rho + (1 - \lambda)\tilde{\mu}_0[S_G(\kappa^*)]} = 0. \qquad (\star)
\end{aligned}$$

Equation ($\star$) above implicitly defines κ^* as a function of ρ. Differentiating the left-hand side with respect to S_G gives us

$$1 - \frac{(1-\lambda)(\rho-\lambda)}{(\tilde{\mu}_0[S_G(\kappa^*)](1-\lambda)+\rho)^2}\tilde{\mu}_0'[S_G(\kappa^*)] > 0,$$

where the inequality follows because the fraction above is nonnegative and $\tilde{\mu}_0'[x] < 0$. Because $\frac{\partial S_G}{\partial \kappa} > 0$, the derivative of the left-hand side of Equation ($\star$) with respect to κ is positive by the chain rule. Thus, it suffices to show that the derivative of the left-hand side of Equation ($\star$) with respect to ρ is negative, in which case the implicit function theorem implies that $\frac{\kappa^*}{\partial \rho} > 0$.

For this last step, differentiating the left-hand side of Equation ($\star$) with respect to ρ gives us

$$\underbrace{\frac{\lambda + (1-\lambda)\tilde{\mu}_0[S_G(\kappa^*)]}{(\rho + (1-\lambda)\tilde{\mu}_0[S_G(\kappa^*)])^2}}_{\text{direct effect}} + \underbrace{\frac{\partial S_G}{\partial \rho}\left(1 - \frac{(\rho-\lambda)(1-\lambda)\tilde{\mu}_0'[S_G(\kappa^*)]}{(\rho + (1-\lambda)\tilde{\mu}_0[S_G(\kappa^*)])^2}\right)}_{\text{indirect effect}}.$$

Notice this expression is strictly negative if

$$\frac{\partial S_G}{\partial \rho} < -\frac{1}{\rho^2}. \tag{E.1}$$

Furthermore, the expression for $\frac{\partial S_G}{\partial \rho}$ in Claim 9 is strictly increasing as a function of $g'(\beta - (\gamma + \kappa)\tilde{\mu}_0[S_G(\kappa^*)])$. Because g is concave, $g'(\beta - (\gamma+\kappa)\tilde{\mu}_0[S_G(\kappa^*)]) \le \frac{g(\beta-(\gamma+\kappa)\tilde{\mu}_0[S_G(\kappa^*)]) - g(\beta-\gamma-\kappa)}{(\gamma+\kappa)(1-\tilde{\mu}_0[S_G(\kappa^*)])}$. Thus, a sufficient condition for the inequality in Equation E.1 is

$$\frac{\delta(\tilde{\mu}_0[S_G(\kappa^*)] + (1-\tilde{\mu}_0[S_G(\kappa^*)])\lambda)(1-\tilde{\mu}_0[S_G(\kappa^*)])}{\tilde{\mu}_0'[S_G(\kappa^*)]\rho(\rho + \delta(\rho + 1 - 2\lambda(1-\tilde{\mu}_0[S_G(\kappa^*)]) - 2\tilde{\mu}_0[S_G(\kappa^*)]))} < -\frac{1}{\rho^2}.$$

Rearranging gives us

$$\frac{\delta(\tilde{\mu}_0[S_G(\kappa^*)] + (1-\tilde{\mu}_0[S_G(\kappa^*)])\lambda)(1-\tilde{\mu}_0[S_G(\kappa^*)])}{\tilde{\mu}_0'[S_G(\kappa^*)](\rho + \delta(\rho + 1 - 2\lambda(1-\tilde{\mu}_0[S_G(\kappa^*)]) - 2\tilde{\mu}_0[S_G(\kappa^*)]))} > \frac{1}{\rho}. \tag{E.2}$$

The right-hand side of the above inequality is strictly decreasing as a function of $S_G(\kappa^*)$. Thus, a sufficient condition of the inequality in Equation E.2 is

$$\frac{\delta(\tilde{\mu}_0[1] + (1-\tilde{\mu}_0[1])\lambda)(1-\tilde{\mu}_0[1])}{\tilde{\mu}_0'[1](\rho + \delta(\rho + 1 - 2\lambda(1-\tilde{\mu}_0[1]) - 2\tilde{\mu}_0[1]))} = \frac{(1-q)\delta\lambda}{q(\rho + \delta(\rho + 1 - 2\lambda))} \ge \frac{1}{\rho}.$$

Rearranging this inequality gives the sufficient condition in the Implication for κ^* to increase in ρ. □

F Proof of Implication 2

For the first result, if $g(s) = s$, then g is concave. So the proof of Lemma 2 establishes that $\bar{\kappa}$ solves

$$\beta - \gamma - \bar{\kappa} = \beta - \rho\frac{(1+\delta)[\beta - (\beta - \gamma)]}{\delta\lambda}.$$

Rearranging gives us, $\bar{\kappa} = \gamma\left(\frac{(1+\delta)\rho}{\delta\lambda} - 1\right)$, which is increasing in γ as $\rho \geq 1$, $\delta > 0$, and $\lambda \in (0, 1]$.

For the second result, note that

$$\frac{\partial\Delta}{\partial\gamma} = \mu_0 + (\gamma + \kappa)\frac{\partial\mu_0}{\partial\gamma}.$$

Here μ_0 is the direct effect. As γ increases, all else equal, unobserved support after message $m = 0$, i.e., $\sigma_O(0, \varnothing)$, decreases because in the mixed strategy equilibrium the observer anticipates government coverups. $(\gamma + \kappa)\frac{\partial\mu_0}{\partial\gamma}$ is an indirect effect. As γ changes, equilibrium behavior and hence beliefs change. Recall that in the mixed strategy equilibrium, $\mu_0 = \tilde{\mu}_0[\sigma_G(1)]$, i.e., beliefs are a function of government behavior. So can use the chain rule to rewrite the above Equation as

$$\frac{\partial\Delta}{\partial\gamma} = \underbrace{\tilde{\mu}_0[\sigma_G(0)]}_{>0} + \underbrace{(\gamma + \kappa)}_{>0}\underbrace{\tilde{\mu}_0'[\sigma_G(1)]}_{<0}\frac{\partial\sigma_G(1)}{\partial\gamma}.$$

So we only need to find $\frac{\partial\sigma_G(0)}{\partial\gamma}$. Recall that in the mixed strategy equilibrium the government's strategy is implicitly defined by $F(\sigma_G(1)) = 0$, where F is increasing in $\sigma_G(0)$. Assuming $g(s) = s$ and differentiating F with respect to γ gives

$$\begin{aligned}\frac{\partial F}{\partial\gamma} &= (1+\delta) - \delta\frac{\lambda + (1-\lambda)\tilde{\mu}_0[\sigma_G(0)]}{\rho} - \tilde{\mu}_0[\sigma_G(1)]\left(1 + \delta\left(1 - \frac{\lambda + (1-\lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}\right)\right)\\ &= \frac{(1 - \tilde{\mu}_0[\sigma_G(1)])\left(\rho + \delta(\rho - \lambda) - \delta(1-\lambda)\tilde{\mu}_0[\sigma_G(1)]\right)}{\rho} > 0.\end{aligned}$$

So the implicit function theorem implies $\frac{\partial\sigma_G(1)}{\partial\gamma} = -\frac{\partial F}{\partial\gamma}(\frac{\partial F}{\partial\sigma_G(1)})^{-1} < 0$. Using the equation above, we have $\frac{\partial\Delta}{\partial\gamma} > 0$.

G Proof of Implication 3

Claim 11. *Assume g is strictly concave. As the population's bias (β) increases, the truthful equilibrium becomes less likely in the set inclusion sense.*

Proof. When g is strictly concave (and strictly increasing), Assumption 1 holds. Under Assumption 1, Lemma 2 demonstrates that there exists $\bar{\kappa} > 0$ such that the truthful equilibrium exists if and only if $\kappa \geq \bar{\kappa}$. In addition, the cutpoint $\bar{\kappa}$ is implicitly defined by the equation:

$$\underbrace{g(\beta-\gamma-\bar{\kappa}) - g(\beta) + \rho\frac{(1+\delta)[g(\beta)-g(\beta-\gamma)]}{\delta\lambda}}_{\equiv C(\kappa)} = 0. \tag{G.1}$$

We show that $\bar{\kappa}$ is increasing in β. First, $\frac{\partial C}{\partial \kappa} < 0$ because $g'(s) > 0$ for all support levels s. Second,

$$\frac{\partial C}{\partial \beta} = g'(\beta-\gamma-\bar{\kappa}) - g'(\beta) + \frac{\rho(1+\delta)}{\delta\lambda}[g'(\beta-\gamma) - g'(\beta)]$$

Because g is strictly concave, $\tilde{s} > s$ implies $g'(\tilde{s}) < g'(s)$. So $g'(\beta) < g'(\beta-\gamma)$ and $g'(\beta) < g'(\beta-\gamma-\kappa)$. Thus, $\frac{\partial C}{\partial \beta} > 0$, and the Implicit Function Theorem implies $\frac{\partial \bar{\kappa}}{\partial \beta} > 0$. □

Claim 12. *As the population's bias (β) increases, the never-admit-fault equilibrium becomes more (less) likely in the set inclusion sense if and only if*

$$\frac{g'(\beta-\gamma-\underline{\kappa}) - g'(\beta-(\gamma+\underline{\kappa})q)}{g'(\beta-\gamma) - g'(\beta-(\gamma+\underline{\kappa})q)} > (<) \frac{\rho(1+\delta)}{\delta(q+(1-q)\lambda)}$$

Proof. Under Assumption 1, Lemma 2 demonstrates their exists $\underline{\kappa} > 0$ such that never-admit-fault equilibrium exists if and only if $\kappa \leq \underline{\kappa}$. In addition, the cutpoint $\underline{\kappa}$ is implicitly defined by the equation $D(\underline{\kappa}) = 0$, where

$$D(\underline{\kappa}) = g(\beta-\gamma-\underline{\kappa}) + (c-1)\cdot g(\beta-(\gamma+\underline{\kappa})q) - c\cdot g(\beta-\gamma)$$

and $c = \frac{\rho(1+\delta)}{\delta(q+(1-q)\lambda)} > 1$. First, $\frac{\partial D}{\partial \kappa} < 0$ as $g'(s) > 0$ and $c > 1$. Second,

$$\begin{aligned}\left.\frac{\partial D}{\partial \beta}\right|_{\kappa=\underline{\kappa}} &= g'(\beta-\gamma-\underline{\kappa}) + (c-1)g'(\beta-(\gamma+\underline{\kappa})q) - cg'(\beta-\gamma)\\ &= g'(\beta-\gamma-\underline{\kappa}) - g'(\beta-(\gamma+\underline{\kappa})q) + c[g'(\beta-(\gamma+\underline{\kappa})q) - g'(\beta-\gamma)]\end{aligned}$$

Because $\frac{\partial D}{\partial \kappa} < 0$, the sign of $\left.\frac{\partial D}{\partial \beta}\right|_{\kappa=\underline{\kappa}}$ will determine the sign of $\frac{\partial \underline{\kappa}}{\partial \beta}$ by the Implicit Function Theorem. First, notice that strict concavity implies, $g'(\beta-\gamma-\underline{\kappa}) > g'(\beta-(\gamma+\underline{\kappa})q)$. Second, notice that $\underline{\kappa} < \frac{\gamma(1-q)}{q}$. If not, then $(\gamma+\underline{\kappa})q \geq \gamma$ and $g(\beta-\gamma) \geq g(\beta-(\gamma+\underline{\kappa})q)$—but this would mean $D(\underline{\kappa}) < 0$, a contradiction. Because $\underline{\kappa} < \frac{\gamma(1-q)}{q}$, $g(\beta-(\gamma+\underline{\kappa})q) > g(\beta-\gamma)$ and $g'(\beta-(\gamma+\underline{\kappa})q) < g'(\beta-\gamma)$ as g is strictly increasing and strictly concave. Rewriting the above expression in terms of c gives us $\left.\frac{\partial D}{\partial \beta}\right|_{\kappa=\underline{\kappa}} > 0$ if and only if

$$\frac{g'(\beta-\gamma-\underline{\kappa}) - g'(\beta-(\gamma+\underline{\kappa})q)}{g'(\beta-\gamma) - g'(\beta-(\gamma+\underline{\kappa})q)} > c = \frac{\rho(1+\delta)}{\delta(q+(1-q)\lambda)}.$$ □

H Proof of Lemma 3 & Implication 4

H.1 Proof of Lemma 3

In the partially truthful equilibrium (σ, μ), $\frac{\partial \sigma_G(1)}{\partial \rho} < 0$ and $\frac{\partial \mu_0}{\partial \rho} > 0$ follow from Claim 9 and the beliefs in μ_0 in Lemma 1. In the truthful or never-admit-fault equilibrium, the government is using a pure strategy which is independent of ρ.

H.2 Never-admit-fault equilibrium

In the never-admit-fault equilibrium, the government sends message $m = 0$ regardless of its type, implying $\mu_0 = q$. On the equilibrium path of play, the observer gives uninformed support $\sigma_O(0; \varnothing) = \beta - (\gamma + \kappa)q$ and the NGO invests effort $\sigma_N(0) = \frac{\lambda + (1-\lambda)q}{\rho}$. Taken together, G's ex ante expected utility is

$$\overbrace{g(\beta - (\gamma + \kappa)q)}^{\text{initial support}} + \delta \overbrace{\left[\sigma_N(0) \underbrace{(qg(\beta - \gamma - \kappa) + (1-q)g(\beta))}_{v \text{ revealed}} + \underbrace{(1 - \sigma_N(0))g(\beta - (\gamma + \kappa)q)}_{v \text{ not revealed}}\right]}^{\text{final support}}$$

Notice G's expected benefits from its final level of support is a convex combination of $(qg(\beta - \gamma - \kappa) + (1-q)g(\beta))$ and $g(\beta - (\gamma + \kappa)q)$ with weights $\sigma_N(0) = \frac{\lambda + (1-\lambda)q}{\rho}$ and $1 - \sigma_N(0) = 1 - \frac{\lambda + (1-\lambda)q}{\rho}$, respectively. As ρ increases, more weight is put on the latter term. This strictly increases G ex anted expected utility if and only if

$$g(\beta - (\gamma + \kappa)q) > qg(\beta - \gamma - \kappa) + (1-q)g(\beta).$$

Note the above inequality always holds if g is strictly concave.

H.3 Partially truthful equilibrium

In the partially truthful equilibrium, governments with type $v = 1$ are indifferent between admitting to illegitimate violence and not. If they admit to illegitimate violence and send $m = 1$, then $\mu_1 = 1$ and G's ex post expected utility is therefore $(1 + \delta)g(\beta - \gamma)$, which is constant in ρ. This means can we just focus on the expected utility of governments with type $v = 0$. For governments with type $v = 0$, their ex post expected utility is

$$g(\beta - (\gamma + \kappa)\tilde{\mu}_0[\sigma_G(1)]) + \delta[\sigma_N(0)g(\beta) + (1 - \sigma_N(0))g(\beta - (\gamma + \kappa)\tilde{\mu}_0[\sigma_G(1)])]$$

which is equal to

$$(1 + \delta(1 - \sigma_N(0))) \underbrace{g(\beta - (\gamma + \kappa)\tilde{\mu}_0[\sigma_G(1)])}_{\text{uninformed support}} + \delta \sigma_N(0) \underbrace{g(\beta)}_{\substack{\text{informed} \\ \text{support}}} .$$

Substituting $\sigma_N(0) = \frac{\lambda+(1-\lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}$ gives us

$$\left(1+\delta\left(1-\frac{\lambda+(1-\lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}\right)\right)g(\beta-(\gamma+\kappa)\tilde{\mu}_0[\sigma_G(1)])+\delta\left(\frac{\lambda+(1-\lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}\right)g(\beta) \tag{H.1}$$

Note that Equation H.1 is G's expected utility in the partially truthful equilibrium given $v = 0$. We want to know how increasing ρ affects this expression. Notice that, in the partially truthful equilibrium, $\sigma_G(1)$ is a C^1 function. So we can differentiate the above expression with respect to ρ. Doing so, shows that the derivative with respect to ρ takes the form:

$$E_1 + (E_2 + E_3)\tilde{\mu}_0'[\sigma_G(1)]\frac{\partial\sigma_G(1)}{\partial\rho} \tag{H.2}$$

Set $\Delta^+ = g(\beta) - g(\beta-(\gamma+\kappa)\tilde{\mu}_0[\sigma_G(1)]) > 0$ as the difference between informed and uninformed support. We can detail the effects above as follows:

1. E_1 is the direct effect of ρ on the utility in Equation H.1:

$$E_1 \equiv -\frac{\delta(\tilde{\mu}_0[\sigma_G(1)] - \lambda(1-\tilde{\mu}_0[\sigma_G(1)]))\Delta^+}{\rho^2} < 0.$$

   The effect is negative.

2. $\tilde{\mu}_0'[\sigma_G(1)]\frac{\partial\sigma_G(1)}{\partial\rho} > 0$ is how ρ affects beliefs.

3. The effects E_2 and E_3 are the indirect effects about how the change in beliefs affect the expected payoffs in Equation H.1.

   - Effect E_2 is an effort effect: $E_2 \equiv \frac{\delta(1-\lambda)\Delta^+}{\rho} > 0$.
   - Effect E_3 is a support effect:

$$-(\gamma+\kappa)\left(1+\delta\left(1-\frac{\lambda+(1-\lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}\right)\right)g'(\beta-(\gamma+\kappa)\tilde{\mu}_0[\sigma_G(1)]) < 0.$$

     It is negative.

Note that a sufficient condition for Equation H.2 to be negative is $E_2 \leq -E_3$. Because g is concave,

$$g'(\beta-(\gamma+\kappa)\tilde{\mu}_0[\sigma_G(1)]) \geq \frac{\Delta^+}{(\gamma+\kappa)\tilde{\mu}_0[\sigma_G(1)]}.$$

Thus, a sufficient condition for $E_2 \leq -E_3$ is

$$\frac{\delta(1-\lambda)}{\rho} \leq \left(1+\delta\left(1-\frac{\lambda+(1-\lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}\right)\right)\frac{1}{\tilde{\mu}_0[\sigma_G(1)]}$$

which can be rewritten as

$$0 \leq \frac{\rho(1+\delta) - 2\delta\tilde{\mu}_0[\sigma_G(1)](1-\lambda) - \delta\lambda}{\rho\tilde{\mu}_0[\sigma_G(1)]}$$

Notice $\rho \geq 1$ and $\tilde{\mu}_0[\sigma_G(1)] \in (0,q)$ in the partially truthful equilibrium. So a (necessary and) sufficient condition for the above inequality is

$$2\delta\tilde{\mu}_0[\sigma_G(1)](1-\lambda) + \delta\lambda \leq \rho(1+\delta)$$

Notice if $q \leq \frac{1}{2}$, then the left-hand side is bounded above by δ, which is strictly less than the right-hand side. In addition, the inequality holds strictly with $\lambda = 1$, and the left-hand-side is strictly increasing in $\tilde{\mu}_0[\sigma_G]$, which is bounded above by q. Solving for λ, $\lambda \geq \frac{\rho(1+\delta)-2q\delta}{\delta(1-2q)}$ is therefore a sufficient condition for $E_2 \leq -E_3$.

I Extension: Only Illegitimate Violence Is Verifiable

It could be the case that NGOs can only verify illegitimate violence. That is, it might be easier to identify when civilians are killed than when no civilians are killed. To capture this possibility, we amend the baseline model as follows. After the government sends message m and the observer chooses initial support s_1, the NGO investigates with effort $e \in [0,1]$. The investigations produces signal r in the following manner:

- If $v = 1$, then $r = 1$ with probability e, and $r = 0$ with probability $1 - e$.
- If $v = 0$, then $r = 0$ with probability 1.

The observer sees r and e and then chooses final support level s_2.[21] The key here is that $r = 1$ implies $v = 1$, but $r = 0$ does not imply $v = 0$. The payoffs for the government and the observer are the same as above, but we modify the payoffs of the NGO as follows:

$$u_N(e, r; m) = \lambda e + (1-\lambda)r(1-m) - \frac{\rho}{2}(e)^2$$

Comparing this payoff to the baseline model, $(1-\lambda)r(1-m)$ corresponds to N's payoff for exposing the a government coverup, which happens after the NGO verifies that illegitimate violence occurred ($r = 1$) but the government did not acknowledge it ($m = 0$). The term $\frac{\rho}{2}(e)^2$ captures the NGO's investigative costs. Finally, λe is the benefit of the NGO from issuing a quality report.[22] As we show below, this formulation of the NGO's payoffs leads to an identical equilibrium effort condition as in the baseline model. What is changing, however, is what the observer learns after seeing signal $r = 0$.

[21] Even if the observer did not see the amount of effort chosen, our results would not change. NGO payoffs are independent of second-period support s_2. In equilibrium, the NGO is using a pure strategy and the observer would therefore anticipate the equilibrium effort choice.

[22] In the baseline model, the NGO releases a report if and only if it uncovers verifiable information about the state, whether legitimate of illegitimate violence occurs, which occurred with probability e. Here, it is not possible to verify that legitimate violence occurred.

Namely, the degree to which signal $r = 0$ is informative depends on the NGO's equilibrium effort level. To see this, suppose after seeing message m, the posterior belief that $v = 1$ is ν. Then suppose the NGO invests effort e_m which produces signal r. If $r = 1$, then O knows violence was illegitimate. If $r = 0$, then the posterior belief that $v = 1$ is:

$$\begin{aligned}
\Pr(v = 1|e_m, r = 0, m) &= \frac{\Pr(r = 0|v = 1, e_m, m)\Pr(v = 1|e_m, m)}{\Pr(r = 0|e_m, m)} \\
&= \frac{\Pr(r = 0|v = 1, e_m, m)\nu}{\Pr(r = 0|e_m, m)} \\
&= \frac{(1 - e_m)\nu}{\Pr(r = 0|e_m, m)} \\
&= \tfrac{(1-e_m)\nu}{\Pr(r=0|e_m,m,v=1)\Pr(v=1|e_m,m)+\Pr(r=0|e_m,m,v=0)\Pr(v=0|e_m,m)} \\
&= \frac{(1 - e_m)\nu}{(1 - e_m)\nu + (1 - \nu)} \\
&= \frac{(1 - e_m)\nu}{1 - e_m\nu}
\end{aligned}$$

Notice when $e_m = 1$ this posterior belief is 0. That is, when the NGO exerts full effort $r = 0$ reveals that illegitimate violence could not have happened or else r would have been 1. When $e_m = 0$, this posterior belief is ν, that is no new information is acquired with zero effort.

Strategies for the government and the NGO are identical to those defined above. For the observer, a strategy is a function $\sigma_O : \{0, 1\}\times[0, 1] \to \mathbb{R}$ where $\sigma_O(m, \nu)$ is the support O gives the government after message m given it believes $v = 1$ with probability $\nu \in [0, 1]$. Finally, μ_m is the belief that $v = 1$ after message m but before the NGO report, and μ_m^0 is the belief that $v = 1$ after message m and report $r = 0$. An equilibrium is an assessment (σ, μ) where $\sigma = (\sigma_G, \sigma_O, \sigma_N)$ is a sequentially rational strategy profile given beliefs $\mu = (\mu_m, \mu_m^0)_{m\in\{0,1\}}$ and beliefs are consistent with strategies and updated via Bayes rule whenever possible. As in the baseline model, we are implicitly assuming that the observer will have correct beliefs after seeing $r = 1$ in any subgame, i.e., $\Pr(v = 1|r = 1) = 1$.

I.1 Analysis

We first state conditions on the NGO's effort and the observer's support that must be true in any equilibrium. These conditions and their derivation mirror those in Equations 1 and 2 from baseline analysis. After message $m = 0$, when the NGO chooses it's effort level, the belief of a coverup is μ_0. This coverup is revealed after signal $r = 1$, which occurs with probability e. So its equilibrium effort takes the form

$$\sigma_N(m) = \frac{\lambda + (1 - \lambda)\mathbf{I}[m = 0]\mu_0}{\rho}.$$

When the observer chooses support, let the belief that $v = 1$ be ν. Then its equilibrium support satisfies

$$\sigma_O(m, \nu) = \beta - \gamma\nu - \kappa(1 - m)\nu.$$

We now illustrate how Proposition 1 changes when the only illegitimate violence is verifiable: Although the government is weakly more likely to lie in equilibrium, the characterization of equilibria is largely the same. Specifically, if $g(\beta-\gamma-\kappa)$ is sufficiently small, the government is always truthful. When $g(\beta-\gamma-\kappa)$ is sufficiently large, then the government never admits fault. When $g(\beta-\gamma-\kappa)$ is moderate, then the government is partially truthful.

Claim 13. *An equilibrium (σ, μ) in which the government is truthful ($\sigma_G(v) = v$) exists if and only if*

$$g(\beta-\gamma-\kappa) \leq g(\beta) - \rho\frac{(1+\delta)[g(\beta)-g(\beta-\gamma)]}{\delta\lambda}, \tag{I.1}$$

which is the same condition as in the baseline model (Proposition 1, Equation 4) where legitimate violence is verifiable.

Proof. If (σ,μ) is a truthful equilibrium, then $\mu_m = \mu_m^0 = m$. After an incidence of illegitimate violence ($v = 1$) if G admits the truth, then its payoff is $U_G^{\sigma,\mu}(m = 1; v = 1) = (1+\delta)g(\beta-\gamma)$. If G lies and sends message $m = 0$,then its payoff is

$$U_G^{\sigma,\mu}(m=0;v=1) = g(\sigma_O(0,\mu_0)) + \delta\left[\sigma_N(0)g(\sigma_O(0,1)) + (1-\sigma_N(0))g(\sigma_O(0,\mu_0^0))\right].$$

In the above equation, initial support is $\sigma_O(0,\mu_0)$. With probability $\sigma_N(0)$, $r = 1$, $v = 1$ is revealed, and final support is $\sigma_O(0,1)$. With probability $1-\sigma_N(0)$, $r = 0$ in which case final support is $\sigma_O(0,\mu_0^0)$. Using the NGO's equilibrium condition, $\sigma_N(0) = \frac{\lambda}{\rho}$ as $\mu_0 = 0$. Using O's equilibrium condition, $\sigma_O(0,\mu_0) = \beta$ and $\sigma_O(0,1) = \beta-\gamma-\kappa$. Finally, in equilibrium

$$\mu_0^0 = \Pr(v = 1|e = \frac{\lambda}{\rho}, r = 0, m = 0) = \frac{(1-\frac{\lambda}{\rho})\mu_0}{1-\frac{\lambda}{\rho}\mu_0} = 0$$

where the second equality comes from the derivation of $\Pr(v = 1|e, r = 0, m)$ above and the third follows from $\mu_0 = 0$. This implies $\sigma_O(0,\mu_0^0) = \beta$. Making this substitutions gives us

$$U_G^{\sigma,\mu}(m=0;v=1) = g(\beta) + \delta\left[\frac{\lambda}{\rho}g(\beta-\gamma-\kappa) + \left(1-\frac{\lambda}{\rho}\right)g(\beta)\right].$$

To rule out profitable deviations, we need $U_G^{\sigma,\mu}(m = 1; v = 1) \geq U_G^{\sigma,\mu}(m = 0; v = 1)$ which is equivalent to

$$g(\beta-\gamma-\kappa) \leq g(\beta) - \rho\frac{(1+\delta)[g(\beta)-g(\beta-\gamma)]}{\delta\lambda}.$$

□

Claim 14. *An equilibrium (σ, μ) in which the government never admits faults exists if and only if*

$$g(\beta - \gamma - \kappa) \geq g(\beta - (\gamma + \kappa)b) - \rho \frac{g(\beta - (\gamma + \kappa)q) + \delta g(\beta - (\gamma + \kappa)b) - (1 + \delta)g(\beta - \gamma)}{\delta(q + (1 - q)\lambda)} \tag{I.2}$$

where $b = \mu_0^0 = \frac{q(\lambda+(1-\lambda)q-\rho)}{q(\lambda+(1-\lambda)q)-\rho}$. Furthermore, the right-hand side of the inequality is strictly less than the corresponding expression in the baseline model (Proposition 1, Equation 5) where legitimate violence is verifiable.

Proof. We first show that a never-admit-fault equilibrium does not exist if Equation I.2 does not hold. We then argue that never admitting fault is an equilibrium with off-path beliefs $\mu_1 = \mu_1^0 = 1$ if Equation I.2 holds. We finally argue that Equation I.2 is less restrictive than the corresponding never-admit-fault condition in the baseline model (Proposition 1, Equation 5).

Step 1. Suppose (σ, μ) is a never admit fault equilibrium. With $v = 1$, the government's payoff from not admitting illegitimate violence is

$$U_G^{\sigma,\mu}(m = 0; v = 1) = g(\sigma_O(0, \mu_0)) + \delta \left[\sigma_N(0)g(\sigma_O(0, 1)) + (1 - \sigma_N(0))g(\sigma_O(0, \mu_0^0))\right].$$

In the equation above $\mu_0 = q$ as both types of the government pool on $m = 0$. In equilibrium we have

$$\mu_0^0 = \Pr(v = 1 | e = \sigma_N(0), r = 0, m = 0) = \frac{(1 - \sigma_N(0))\mu_0}{1 - \sigma_N(0)\mu_0} = \frac{(1 - \sigma_N(0))q}{1 - \sigma_N(0)q}.$$

Substitution gives us

$$\begin{aligned} U_G^{\sigma,\mu}(m = 0; v = 1) = g(\beta - (\gamma + \kappa)q) + \delta \Bigg[&\frac{\lambda + (1 - \lambda)q}{\rho} g(\beta - \gamma - \kappa) \\ &+ \left(1 - \frac{\lambda + (1 - \lambda)q}{\rho}\right) g\left(\beta - (\gamma + \kappa)\mu_0^0\right) \Bigg], \end{aligned}$$

where $\mu_0^0 = \frac{\left(1 - \frac{\lambda+(1-\lambda)q}{\rho}\right)q}{1 - \frac{\lambda+(1-\lambda)q}{\rho} q} = \frac{q(\lambda+(1-\lambda)q-\rho)}{q(\lambda+(1-\lambda)q)-\rho}$. If the government with $v = 1$ deviates and sends message $m = 1$, its payoff is

$$\begin{aligned} U_G^{\sigma,\mu}(m = 1; v = 1) &= g(\sigma_O(1, \mu_1)) + \delta \left[\sigma_N(1)g(\sigma_O(1, 1)) + (1 - \sigma_N(1))g(\sigma_O(1, \mu_1^0))\right] \\ &= g(\beta - \gamma\mu_1) + \delta \left[\sigma_N(1)g(\beta - \gamma) + (1 - \sigma_N(1))g(\beta - \gamma\mu_1^0)\right] \\ &\geq g(\beta - \gamma) + \delta \left[\sigma_N(1)g(\beta - \gamma) + (1 - \sigma_N(1))g(\beta - \gamma)\right] \\ &= (1 + \delta)g(\beta - \gamma). \end{aligned}$$

In the above expression, note that after G sends message m with probability $\sigma_N(1)$ the NGO produces a report with verifiable information that $v = 1$. Notice that G with type $v = 1$ has a profitable deviation if $(1+\delta)\, g(\beta-\gamma) > U_G^{\sigma,\mu}(m=0; v=1)$. This condition is equivalent to

$$g(\beta-\gamma-\kappa) > g(\beta-(\gamma+\kappa)\mu_0^0) - \rho\frac{g(\beta-(\gamma+\kappa)q) + \delta g(\beta-(\gamma+\kappa)\mu_0^0) - (1+\delta)g(\beta-\gamma)}{\delta(q+(1-q)\lambda)}.$$

Step 2. Assume the inequality in Equation I.2 holds. Consider an assessment (σ, μ) such that $\sigma_G(v) = 0$ and $\mu_1 = \mu_1^0 = 1$. In addition, $\mu_0 = q$, $\mu_0^0 = b = \frac{q(\lambda+(1-\lambda)q-\rho)}{q(\lambda+(1-\lambda)q)-\rho}$, and σ_N and σ_O are defined in the equilibrium conditions above. By previous analysis, N and O are best responding to σ_G, and the beliefs μ_0 and μ_0^0 are derived via Bayes rule. Furthermore, the expected utility calculations in Step 1 prove that that G does not have a profitable deviation when $v = 1$, $\mu_1 = \mu_1^0 = 1$, and Equation I.2 holds.

To see that G does not have a profitable deviation when $v = 0$, first note that Equation I.2 implies $g(\beta-(\gamma+\kappa)q) + \delta g(\beta-(\gamma+\kappa)\mu_0^0) > (1+\delta)g(\beta-\gamma)$. Thus, the payoff $U_G^{\sigma,\mu}(m=0; v=0) = g(\beta-(\gamma+\kappa)q) + \delta g(\beta-(\gamma+\kappa)\mu_0^0)$ is strictly larger than $U_G^{\sigma,\mu}(m=1; v=0) = (1+\delta)g(\beta-\gamma)$. Here, were G to send message $m = 1$ when $v = 0$, the posterior belief is $\mu_1 = \mu_1^0 = 1$, and $v = 0$ cannot be verified by the NGO. So both rounds of support after the deviation are $\beta - \gamma$.

Step 3. Consider the right-hand side of Equation I.2. This expression is strictly decreasing in the variable $y = g(\beta-(\gamma+\kappa)b)$ because $\rho > 0$ and $q + (1-q)\lambda \in (0, 1]$. Furthermore, $g(\beta-(\gamma+\kappa)q) < g(\beta-(\gamma+\kappa)b)$ as $b < q$ and g is strictly increasing. Substituting $g(\beta-(\gamma+\kappa)q)$ for $y = g(\beta-(\gamma+\kappa)b)$ then proves the result. □

Claim 15. *An equilibrium (σ, μ) in which the government admits fault after illegitimate violence with probability strictly between zero and one ($\sigma_G(1) \in (0, 1)$) and never admits fault after legitimate violence exists if and only if the inequalities in Equations I.1 and I.2 do not hold. Furthermore, the equilibrium probability of admitting illegitimate violence is strictly less than in the baseline model where legitimate violence is verifiable.*

Proof. In such an equilibrium $\mu_1 = \mu_1^0 = 1$ because only governments with $v = 1$ are admitting to illegitimate violence, and they do so with positive probability. Thus, if $v = 1$ and G acknowledges illegitimate violence, then its payoff is

$$U_G^{\sigma,\mu}(m=1, v=1) = (1+\delta)g(\beta-\gamma).$$

If G with $v = 1$ does not disclose illegitimate violence, its payoff is

$$\begin{aligned} U_G^{\sigma,\mu}(m = 0, v = 1) &= g(\sigma_O(0, \mu_1)) + \delta \left[\sigma_N(0)g(\sigma_O(0, 1)) + (1 - \sigma_N(0))g(\sigma_O(0, \mu_0^0))\right] \\ &= g(\beta - (\gamma + \kappa)\tilde{\mu}_0[\sigma_G(1)]) + \delta \left[\sigma_N(0)g(\beta - \gamma - \kappa) + \right. \\ &\qquad \left. (1 - \sigma_N(0))g(\beta - (\gamma + \kappa)\tilde{\mu}_0^0[\sigma_G(1)])\right] \\ &= g(\beta - (\gamma + \kappa)\tilde{\mu}_0[\sigma_G(1)]) + \delta \left[\frac{\lambda + (1 - \lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho} g(\beta - \gamma - \kappa) + \right. \\ &\qquad \left. \left(1 - \frac{\lambda + (1 - \lambda)\tilde{\mu}_0[\sigma_G(1)]}{\rho}\right) g(\beta - (\gamma + \kappa)\tilde{\mu}_0^0[\sigma_G(1)])\right] \end{aligned}$$

where $\tilde{\mu}_0[\sigma_G(1)]$ denotes the posterior belief in Lemma 1, And $\tilde{\mu}_0^0[\sigma_G(1)]$ is the posterior belief derived above:

$$\begin{aligned} \tilde{\mu}_0^0[\sigma_G(1)] &= \Pr(v = 1 | e = \sigma_N(0), r = 0, m = 0) \\ &= \frac{(1 - \sigma_N(0))\tilde{\mu}_0[\sigma_G(1)]}{1 - \sigma_N(0)\tilde{\mu}_0[\sigma_G(1)]} \\ &= \frac{\tilde{\mu}_0[\sigma_G(1)](\lambda + (1 - \lambda)\tilde{\mu}_0[\sigma_G(1)] - \rho)}{\tilde{\mu}_0[\sigma_G(1)](\lambda + (1 - \lambda)\tilde{\mu}_0[\sigma_G(1)]) - \rho} \end{aligned}$$

Define the function $G : [0, 1] \to \mathbb{R}$ as

$$G(x) = U_G^{\sigma,\mu}(m = 0, v = 1)|_{\sigma_G(1)=x} - U_G^{\sigma,\mu}(m = 1, v = 1).$$

In a partially truthful equilibrium (σ, μ) we must have $G(\sigma_G(1)) = 0$. Furthermore, if $x \in (0, 1)$ and $G(x) = 0$, then we can construct a partially truthful equilibrium as follows:

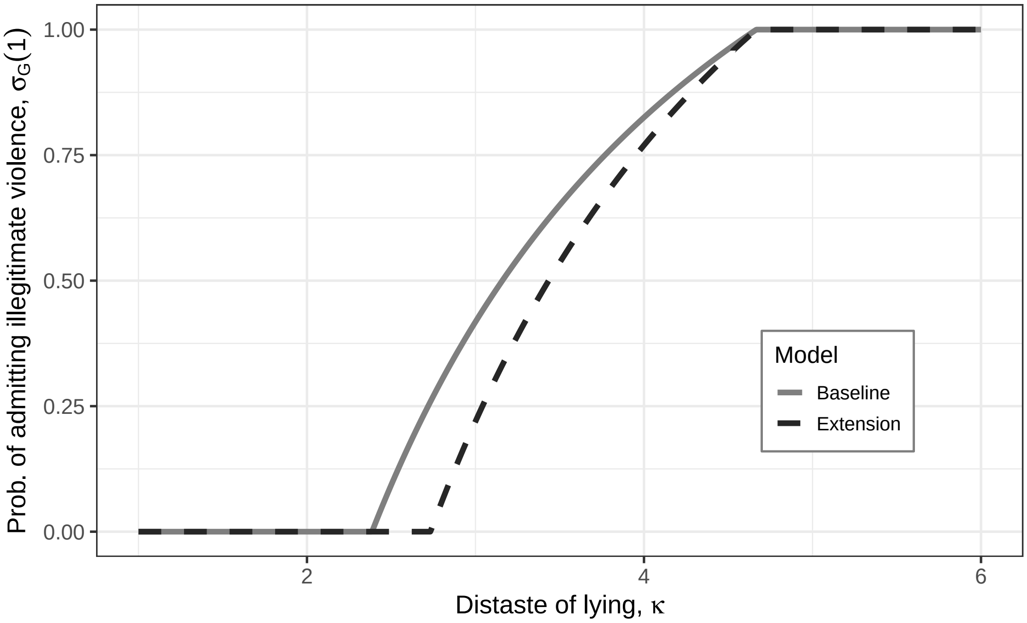
1. $\sigma_G(1) = x$ and $\sigma_G(0) = 0$;
2. $\mu_0 = \tilde{\mu}_0[x]$, $\mu_1 = \mu_1^0 = 1$, and $\mu_0^0 = \tilde{\mu}_0^0[x]$;
3. σ_N and σ_O follow the equilibrium conditions above.

Under this assessment, the government with type $v = 0$ does not have a profitable deviation to send message $m = 1$: $G(x) = 0$ implies $g(\beta - (\gamma + \kappa)\tilde{\mu}_0[x]) + \delta g(\beta - (\gamma + \kappa)\tilde{\mu}_0^0[x]) > g(\beta - \gamma)$, and $\sigma_N(0) \geq \sigma_N(1)$.

First, note that G is continuous. Second, it is also strictly increasing in x. To see this, note that we have already shown that $\tilde{\mu}_0$ is decreasing $\sigma_G(1)$. Furthermore, $\tilde{\mu}_0^0$ is increasing in $\tilde{\mu}_0$ so it is also decreasing in $\sigma_G(1)$. So uninformed support after message $m = 0$ (that is, $\sigma_O(0, \mu_0)$ and $\sigma_O(0, \mu_0^0)$), is increasing in the probability that the government admits illegitimate violence $\sigma_G(1)$. Furthermore, the NGO's equilibrium effort, $\sigma_N(0)$, and thus the probability of exposing a coverup, is decreasing in the truthfulness of the government, $\sigma_G(1)$. It suffices to show that (a) $G(1) > 0$ is equivalent to Equation I.1 and (b) $G(0) < 0$ is equivalent to Equation I.2. The algebra to show (a) and (b) follows along similar lines as in the proof of Proposition 1.

Finally, suppose Equations I.1 and I.2 do not hold. Then there exists equilibrium (σ, μ) in which the government admits fault after illegitimate with probability strictly between zero

Figure I.1: Comparison to the baseline model.



Notes: The solid blue line is the equilibrium probability that the government admits illegitimate violence, $\sigma_G(1)$, in the baseline model (Proposition 1). The dashed orange is the same probability in the extension where only illegitimate is verifiable. Graphs generated assuming $g(s) = s$, $\gamma = 2$, $q = 0.2$, $\delta = 4$, $\rho = 2$, and $\lambda = \frac{3}{4}$.

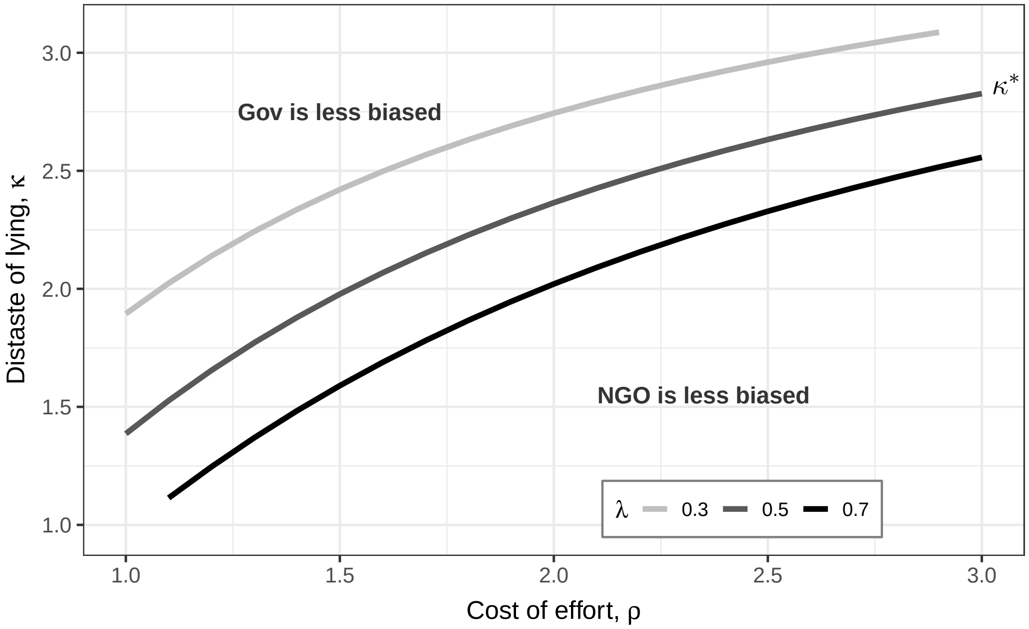
and one where $G(\sigma_G(1)) = 0$. Furthermore, by Claim 14 and Proposition 1 there exists an equilibrium in the baseline model (σ^b, μ^b) such that $\sigma^b_G(1) \in (0, 1)$. As we established previously, the probability of admitting illegitimate violence in the partially truthful equilibrium of the baseline model satisfies $F(\sigma^b_G(1)) = 0$. Notice both F and G are strictly increasing in their arguments. We now show that $G(x) - F(x) > 0$. Thus, if $F(x^b) = G(x) = 0$ for $x^b, x \in (0, 1)$, then $G(x^b) > 0$ and $x^b > x$. To see that $G(x) - F(x) > 0$, we can write the difference as:

$$G(x) - F(x) = \delta \left(1 - \frac{\lambda + (1-\lambda)\tilde{\mu}_0[x]}{\rho}\right) \left[g\left(\beta - (\gamma + \kappa)\tilde{\mu}^0_0[x]\right) - g(\beta - (\gamma + \kappa)\tilde{\mu}_0[x])\right]. \tag{I.3}$$

Because $\rho \geq 1$, $\frac{\lambda + (1-\lambda)\tilde{\mu}_0[x]}{\rho} < 1$. So we only need to show that $g\left(\beta - (\gamma + \kappa)\tilde{\mu}^0_0[x]\right) > g(\beta - (\gamma + \kappa)\tilde{\mu}_0[x])$. Because g is strictly increasing, this is equivalent to $\tilde{\mu}^0_0[x] < \tilde{\mu}_0[x]$, which holds by the definition of $\tilde{\mu}^0_0$. □

Figure I.1 illustrates how the government's equilibrium probability of admitting illegitimate violence changes across two versions of the model: in the baseline model, NGO reports can verify both legitimate and illegitimate violence, but in this extension, NGO report can only verify illegitimate violence. When $\sigma_G(1)$ is zero (small distaste of lying), the government is in the never-admit-fault equilibrium. When this probability is one (large distaste of lying),

Figure I.2: The Cutpoint κ^* When Legitimate Violence Is Not Verifiable.



Notes: Graph generated using the same example as Figure 2, where $g(s) = s$, $\gamma = 1$, and $q = 0.2$. In the original Figure, legitimate violence was verifiable, but here it is unverifiable.

the government is in the truthful equilibrium. As the graph demonstrates, when legitimate violence is unverifiable, the never-admit-fault equilibrium becomes more likely in the set inclusion sense (Claim 14). Furthermore, when the government is partially truthful in the extension, the government would be more truthful were legitimate violence to be verifiable (Claim 15). Finally, if the government is truthful in the baseline model, it would be truth were legitimate violence to not be verifiable and *vice versa* (Claim 13).

Even when legitimate violence is unverifiable, Implication 1 can still hold. Namely, when g is concave, we can find a $\kappa^* > 0$ such that the government has larger bias than the NGO if and only if $\kappa < \kappa^*$. The key to this is illustrated in Figure I.1: when κ is small, the government is never admiting fault so it's bias is larger than the NGO's. When κ is large, the government is truthful so it's bias is zero and smaller than the NGOs. Furthermore, the cutpoint can be increasing in the NGO's cost of effort. To illustrate this possibility, we graph κ^* as a function of ρ is Figure I.2. Notably, we use the same numerical example as the one generating Figure 2, which illustrated Implication 1 in the baseline model. With and without verifiable legitimate violence, the substantive takeaway is the same.