APPENDIX (online only)

A 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 $-\gamma q > -\gamma$.

B 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 $w_1+w_2=1$. □

C 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 $\kappa \in (\underline{\kappa}, \bar{\kappa})$.

D 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

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

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, $g'(\beta-(\gamma+\kappa)\tilde{\mu}_0[S_G(\kappa^*)]) \leq \frac{g(\beta-(\gamma+\kappa)\tilde{\mu}_0[S_G(\kappa^*)])-g(\beta-\gamma-\kappa)}{(\gamma+\kappa)(1-\tilde{\mu}_0[S_G(\kappa^*)])}$. Thus, a sufficient condition for the inequality in Equation D.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

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

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 D.2 is

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

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

E 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 $\frac{\partial\Delta}{\partial\gamma} > 0$.

F 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:

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

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 $\frac{\partial \kappa}{\partial \beta}$ 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

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

□

G Proof of Lemma 3 & Implication 4

G.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.

G.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

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

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.

G.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

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

Note that Equation G.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:

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

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 G.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 G.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 G.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 $E_2 \leq -E_3$.

H 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:

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

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., $\Pr(v = 1|r = 1) = 1$.

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*

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

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*

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

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 H.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 H.2 holds. We finally argue that Equation H.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 H.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 H.2 holds.

To see that G does not have a profitable deviation when $v=0$, first note that Equation H.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 H.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 H.1 and H.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

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

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

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

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:

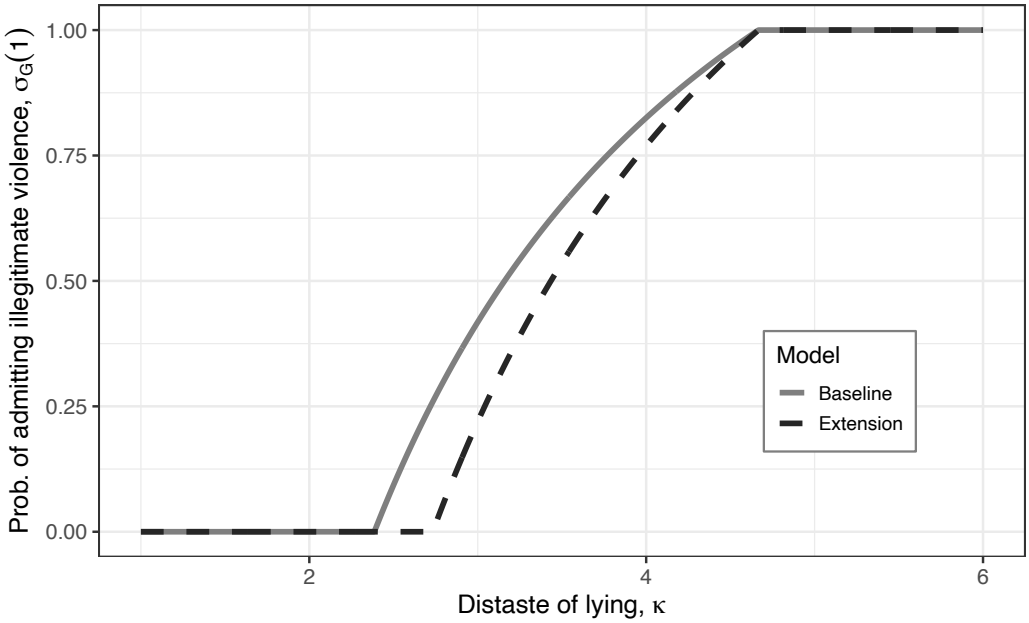
1. $\sigma_G(1) = x$ and $\sigma_G(0) = 0$;
2. $\mu_0 = \tilde{\mu}_0[x]$, $\mu_1 = \mu_1^0 = 1$, and $\mu_0^0 = \tilde{\mu}_0^0[x]$;
3. σ_N and σ_O follow the equilibrium conditions above.

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 H.1 and (b) $G(0)<0$ is equivalent to Equation H.2. The algebra to show (a) and (b) follows along similar lines as in the proof of Proposition 1.

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

Figure H.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 $\lambda = \frac{3}{4}$.

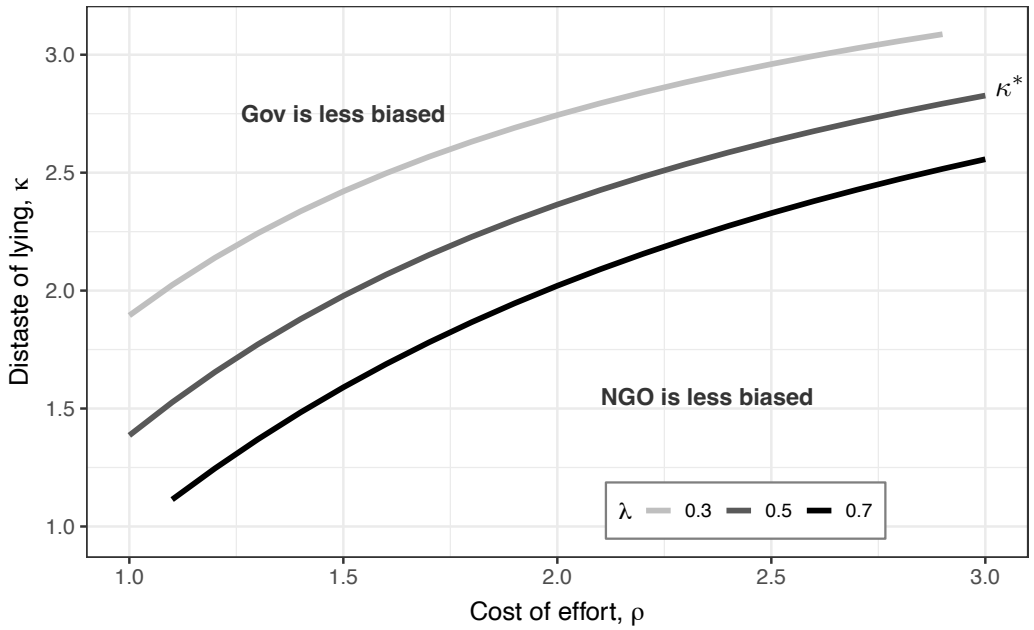
and one where $G(\sigma_G(1)) = 0$. Furthermore, by Claim 14 and Proposition 1 there exists an equilibrium in the baseline model (σ^b, μ^b) such that $\sigma^b_G(1) \in (0, 1)$. As we established previously, the probability of admitting illegitimate violence in the partially truthful equilibrium of the 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:

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

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 H.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 H.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 H.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 H.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.