Why Do Authoritarian Regimes Sign the Convention Against Torture? Signaling, Domestic Politics and Non-Compliance*

James R. Hollyer†
New York University

B. Peter Rosendorff‡
New York University

Current Version: August 2010

Abstract

Traditional international relations theory holds that states will join only those international institutions with which they generally intend to comply. Here we show when this claim might not hold. We construct a model of an authoritarian government's decision to sign the UN Convention Against Torture (CAT). Authoritarian governments use the signing of this treaty - followed by the willful violation of its provisions - as a costly signal to domestic opposition groups of their willingness to employ repressive tactics to remain in power. In equilibrium, authoritarian governments that torture heavily are more likely to sign the treaty than those that torture less. Signatory regimes are predicted to survive longer in office than non-signatories, enjoy less domestic opposition, and reduce their levels of repression - and we provide empirical support for these predictions. While the CAT reduces levels of torture in signatory states, it also prolongs authoritarian regimes' tenure in power.

*We would like to thank James Vreeland and Jennifer Gandhi for their generosity in providing access to their data, and Leslie Johns for her detailed comments and suggestions. We would also like to thank Bruce Bueno de Mesquita, David Stasavage, Jon Eguia, Fernando Martel Garcia, Art Stein, Johannes Urpelainen, Joanne Gowa, the participants in seminars at Claremont, Columbia, Georgetown, NYU, UCLA, UCSD and USC; the 2009 MPSA Panel on International Human Rights Agreements, and the 2009 APSA Panel on the Political Economy of International Regimes and the 3rd annual PEIO conference for helpful comments and suggestions. All remaining errors are our own.

†Corresponding author: New York University, Department of Politics. 19 W. 4th St., 2nd Floor. New York, NY 10012.

‡New York University, Department of Politics. 19 W. 4th St., 2nd Floor. New York, NY 10012.

1 Introduction

Sovereign states that sign international treaties, we are told, intend (most of the time) to comply with the obligations imposed by these treaties. The reasons for this claim are varied: International law (the Vienna Convention on the Law of Treaties in particular) declares that "every treaty ... is binding upon the parties." This declaration follows from the basic principle of international law *pacta sunt servanda* - treaties are to be obeyed. Downs, Rocke & Barsoom (1996) establish that those countries that are most likely to abide by the rules promulgated by an international institution are also those countries that are most likely to join in the first place. Failure to comply with treaty provisions is described as a 'managerial problem' (Chayes & Chayes 1993), or as a temporary aberration to be remedied by (re)negotiation (Koremenos 2005). Tolerated temporary escape (Bagwell & Staiger 2005, Rosendorff & Milner 2001), exchanges of information, and dispute resolution mechanisms are designed to complete gaps in treaty language or to generate better information about signatories' behavior (Rosendorff 2005) and to thus bring about treaty compliance. Where international treaties address issues of international externalities - such as trade, security, or the environment – the intent to comply is strengthened by the mutual gains associated with a predictable, stable, and cooperative international order. Similarly, states facing collective action problems may be inclined to forgo the temporary benefits of defection in order to remain within the society of cooperative nations, especially when future (relative to current) consumption is highly valued (Downs & Rocke 1995).

Recent scholarship has explored if these general findings also apply to the specific case of human rights treaties. Simmons (Forthcoming, 2009) argues that the major human rights treaties have been successful in reducing the prevalence of torture worldwide. She claims that countries accede to and ratify these treaties because they intend to comply with treaty provisions (p.42). She acknowledges that there are some states that ratify, but do not greatly

adjust their behavior. She describes these states as the "false positives,"the countries that sign human rights treaties and continue to torture. And she observes that these states tend to have relatively authoritarian regimes.

Simmons' observation reinforces the conclusions of Hathaway (2007), who finds a positive association between the practice of torture and the signing of human rights treaties by highly authoritarian regimes. Hafner-Burton & Tsutsui (2005, 2007) confirm that signing human rights treaties has little or no effect on the behavior of the world's worst repressors. As they put it, there is a "rising gap between states' propensity to join the international human rights regime and to bring their human rights practice into compliance" and this gap brings the efficacy of international law into fundamental question.

We thus have a puzzle: if states join agreements because they intend to comply with them, why do some states, particularly authoritarian states, sign and fail to comply with human rights treaties?

We argue that authoritarian states sign human rights treaties explicitly because they do *not* intend to comply. And it is important to those signatories that all observers understand that they have no intention of complying at the time of signing. The logic, while counter-intuitive, is straightforward: an elite facing threats from a domestic opposition can mitigate these threats by engaging in torture. If there is any additional cost to the elite of signing and then being found to torture, the act of signing the agreement signals to the opposition the strength of the elite's commitment to remaining in power. The signing of a human rights treaty is a signal to the opposition of the high value the elite places on holding onto power and its willingness to use torture if necessary. On observing the government's actions, the opposition - now better informed about the value the elite places on holding power - will rationally reduce its anti-regime activities. The government continues to torture, but will torture less. On the other hand a regime that doesn't sign shows itself to be vulnerable to the added costs associated with the use of torture. Thus, the opposition will increase efforts

to remove the regime on seeing that the government does not sign.

This logic leads to two conclusions: First, more repressive regimes (regimes with elites more willing to use force to hold onto power) will sign more frequently than less (or non-) repressive governments. Second, opposition political action falls in signatory states - yielding to reductions in the likelihood of regime collapse or transition. In the non-signatory states, opposition response actually rises, leading to more frequent regime failure.

The first finding is consistent with Hathaway's (2007) empirical results, and offers a theoretical explanation for the puzzle above. In order to check the veracity of the model, we test the second prediction: authoritarian regimes that sign the treaty will enjoy longer tenures in office than those that do not. This is true for two reasons: (1) a selection effect implies that those regimes that will fight most strongly to remain in power are the same regimes that sign the treaty; (2) an information effect implies that domestic opposition groups will engage in fewer activities designed to overthrow a signatory government. We test this claim using data on the signing of the UN Convention Against Torture (CAT) and find that it enjoys robust empirical support. Signatory regimes face a lower hazard rate than observationally similar non-signatories across a wide variety of empirical specifications.

We further test the claims that signing the CAT leads to a decline in levels of oppositional activity and of government repression. These claims follow from the *informational effect* of signing the CAT. Domestic opposition groups on witnessing the government sign the CAT conclude that it is a 'strong' type, likely to prevail in the contest for power. As a result, they reduce costly activities aimed at the government's overthrow. In response, the government reduces its repressive activities aimed at remaining in power. We find support for these claims using a variety of measures of domestic conflict and of torture.

Key to the causal logic of the argument is the notion that the CAT affects the costs to a member-state's elite of engaging in torture. We will argue that, aside from international opprobrium and withdrawals of concessions (or active sanctions) along other dimensions (such

as international trade) by the international community (Hafner-Burton 2005), signatories of the CAT must consider the role of "universal jurisdiction"and the extradition clauses of the CAT when determining whether or not to employ torture. These additional considerations serve to make torture more costly given accession to the CAT than not. However, these costs do not directly translate into higher levels of compliance by signatory states. Rather, they allow signing to act as a costly signaling mechanism, such that the states that sign are those that are most likely to defy their treaty obligations.

This paper makes contributions to three literatures. It first speaks to the literature on selection effects and international institutions. While it may generally be the case that governments join treaties by whose provisions they intend to abide; there may exist circumstances in which governments benefit by acceding to treaties whose provisions they intend to defy. Our model offers one instance in which this may take place.

Secondly, the paper speaks to the interaction of international regimes and domestic politics via an informational pathway. While the role of international institutions in generating information that facilitates cooperation among states is well recognized, here we identify a flow of information generated by the international institution that affects the domestic political conflict in significant and unexpected ways. Here, the information generated by signing the CAT leads to less domestic opposition and the preservation of torturing regimes in power.

This paper also contributes to the literature on human rights law. We provide a theory of when and why authoritarian governments are likely to join human rights treaties, and provide empirical evidence in support of this theory. We also explore an "unintended consequence" of increased legalization of the human rights agenda - these legal instruments provide signaling opportunities to domestic oppositions of the elite's intent *not* to abide by its obligations, and result in the increased survival in office of torturing regimes.

2 Autocracies and the CAT

The United Nations Convention Against Torture and Other Cruel, Inhuman and Degrading Treatment or Punishment (CAT) was adopted in December 1984, went into effect in June 1987. It has been ratified by 139 states. It forbids

> "any act by which severe pain or suffering, whether physical or mental, is intentionally inflicted on a person for such purposes as obtaining from him or a third person information or a confession, punishing him for an act he or a third person has committed or is suspected of having committed, or intimidating or coercing him or a third person, or for any reason based on discrimination of any kind, when such pain or suffering is inflicted by or at the instigation of or with the consent or acquiescence of a public official or other person acting in an official capacity."(CAT Article 1)

The CAT requires that each member-state passes appropriate domestic laws making torture a crime, and requires that each state asserts jurisdiction when the crime occurs within its own territory, or the offender or the victim is a national of that state, or if the offender is present in its territory (if the member-state does not for some reason extradite the offender).

The emergence of a set of human rights treaties has been heralded as a major shift in the international system,[1] and a measure of the success and efficacy of international law. While these agreements have been ratified by most states in the world, repressive state behavior has continued to rise over time. Hafner-Burton & Tsutsui (2005) report that in 2000, while the average state had ratified 80% of all available human rights treaties, 35% of states are

[1]The major human rights treaties (in addition to the CAT) are the International Convention on the Elimination of All Forms of Racial Discrimination (adopted 1965), the International Convention on Economic, Social, and Cultural Rights (1966), The International Convention on Civil and Political Rights (1966), the Convention on the the Elimination of All Forms of Discrimination Against Women (1979) and the Convention on the Rights of the Child (1989). In addition other treaties, such as many Preferential Trading Agreements have both soft and hard prohibitions against human rights abuses (Hafner-Burton 2005).

reported as having violated these agreements. States then are clearly willing to sign human rights agreements and continue to violate their treaty commitments.

The lack of compliance with human rights treaties is often viewed as stemming from a failure of enforcement. Enforcement of an international obligation has a number of prerequisites. First, failure to abide by the agreement must be observable. If violations are obscure, mixed in with noise or are otherwise difficult to observe or prove, enforcement is difficult and compliance unlikely. Second, there must exist a system of punishments to be imposed on a state or its elite in the event of a treaty violation to deter non-compliance (or rewards and incentives for compliance). And third, there must be some mechanism or process by with these costs are actually applied (or the benefits accrued). At the international level, this may be the withdrawal of concessions by a trading partner, or the application of sanctions (or enhancements to a state's trading or investment opportunities). At the domestic level, failure to abide by a ratified and implemented international agreement is likely to be a violation of domestic law and subject to sanction by domestic authorities currently or in the future.

Human rights treaties are viewed as being weak on all three dimensions. Violations are difficult to observe.[2] The costs of non-compliance are low, and any potential benefits stemming from compliance such as enhanced trade or investment flows are absent (Nielsen & Simmons 2009). There are few mechanisms for enforcing the agreement.[3] If non-compliance

[2]The CAT does establish a monitoring committee, but it can only investigate and file a report of torture if the torturing state has explicitly accepted Article 21 and or Article 22; otherwise such allegations must be ignored, even if the state is a signatory to the rest of the CAT.

[3]A number of scholars have argued that even weak enforcement regimes can influence state behavior - by socialization into norms of appropriateness (Finnemore 1996) or cascades, where states feel pressured to conform (Keck & Sikkink 1998). Others have argued that the international regimes create openings for non-governmental actors (NGOs) to engage in information gathering, political action, legal maneuvering etc. that influence state behavior (Neumayer 2005, Simmons Forthcoming, 2009). Moravcsik (2000) suggests that unstable democracies can "lock-in" human rights norms by treaty accession. Gilligan & Nesbitt (2007) argue that these norm-based arguments for the adoption of the CAT have not had any noticeable effect on torture levels. Nielsen & Simmons (2009) further find no evidence that signatory governments receive even praise from the US State Department on signing the CAT. If states exert pressure to sign the CAT and conform to its provisions, there is little evidence of this pressure in press statements.

costs are in fact quite low, following Downs, Rocke & Barsoom (1996), we would expect most or all states to sign the CAT, and that state behavior on signing would be little (or un-)changed. The pattern of accession and compliance is somewhat different however. Many governments do not accede to the CAT (in our sample of 129 authoritarian regimes between 1985 and 1996, 74 regimes were never signatories); others sign and reduce torture levels, while still others sign and continue to torture.[4]

Scholars have focused on domestic enforcement mechanisms to explain the observed variation in accession patterns and torture behavior. Hathaway contends that since domestic mechanisms to enforce compliance - such as an independent judiciary or an opposition party - are absent in autocracies, they find accession to human rights treaties essentially costless. In democracies, on the other hand, treaty violations are likely to impose costs on the incumbent government in the form of legal penalties or opposition attacks. Therefore, democracies are only likely to accede to human rights treaties if they are in compliance with these treaties' provisions *before* signing. Autocracies, however, should be willing to enter such treaties regardless of prior compliance. Both autocratic torturers and non-torturers will accede to the CAT. Contrary to this claim, however, Hathaway's (2007) empirical findings indicate that, amongst autocracies, there is a positive association between torture and the signing of the CAT.

Vreeland (2008) explores the domestic political and institutional dynamics of autocracies, and offers an explanation for Hathaway's puzzling finding. He contends that the positive association between levels of torture and accession to the CAT stems from omitted variable bias. More precisely, Vreeland argues that the presence of domestic opposition parties

[4]See Vreeland (2008) for a description of the variation in torture levels among dictatorships. Neumayer (2005) shows that torture levels fall in democracies and in other polities with richer civil society. Simmons (Forthcoming, 2009) argues that the CAT reduces torture in all but the most stable democracies and autocracies, due to the presence of NGOs and other civil society actors. Powell & Stanton (2009) demonstrate that an average of 83 percent of CAT signatories violate some CAT provisions each year, and 42 percent of signatories systematically violate CAT provisions.

both causes autocrats to torture more heavily and forces the these governments to sign human rights treaties. When opposition parties exist, there must be some freedom to engage in speech and activities that contradict the will of the incumbent government. In such a situation, opposition activists are likely to "cross the line" in their criticisms, leading the government to employ torture to maintain its control. Moreover, these opposition parties will pressure the government to enter into human rights agreements. Since Hathaway's regressions do not control for the presence of opposition parties, she finds a spurious association between torture and accession. When the presence of such parties is controlled for, the association between torture and the signing of the CAT drops to insignificance.

Vreeland's theory appears to rely on out-of-equilibrium behavior. If, as Hathaway claims, human rights treaties do not constrain autocratic governments, what is motivating the domestic opposition to push for treaty accession in the first place? Opposition groups are acting on out-of-equilibrium beliefs. If, on the other hand, human rights treaties do constrain autocratic governments, Vreeland does not articulate how these treaties do so. Nor is it clear why, if autocratic governments are willing to so tie their hands, a treaty is necessary to enforce cooperation between the government and opposition.[5]

It may be argued that the CAT acts as a commitment mechanism that constrains the government from acting against the opposition following an agreement exchanging reduced levels of repression by the government for reduced anti-regime activities by the opposition. However, commitment problems between a government and an opposition are two-sided. If the government is tying its hands by signing the CAT, how does the opposition similarly

[5]Empirically, the inclusion of a control for the presence of opposition parties causes the association between torture and the signing of the CAT to drop to insignificance only when a broad spectrum of other controls are also included in the Vreeland regressions. When only 'opposition parties' and 'torture levels' are used to predict signing, both are significant. Moreover, the inclusion of additional variables does not significantly reduce the magnitude of the coefficient on torture. Since there is a substantial amount of multicollinearity between these 'torture' and 'opposition' measures, one cannot determine whether the newfound insignificance of 'torture' is simply due to problems of estimation. Without a more convincing theory of why the presence of opposition parties leads an autocrat to sign human rights treaties, there seems little reason to suppose otherwise.

commit to refrain from future anti-regime activities once the CAT is signed? Presumably, if the government has tied its hands by signing the CAT, it opens itself to future oppositional efforts. Any theory postulating such a mechanism should fully specify the means by which the opposition can commit to a compromise agreement with the government, or else rely on *ad hoc* assumptions about the the credibility of opposition promises.

Hafner-Burton & Tsutsui (2007) also explore the link between autocratic accession to the CAT and torture levels. They argue (as in Hathaway 2007) that while there are vague political benefits from CAT membership ("window-dressing"); these treaties lack coercion and enforcement mechanisms, fail to make states internalize or acculturize international norms, and do not cause a domestic human rights institutional capacity to emerge. Autocracies are therefore unlikely to show any evidence of improvement in repressive behavior after accession. In fact, the effect of signing the CAT is least in those autocratic regimes that torture heavily *ex ante*. This result is robust conditioning on measures of civil society.

Yet explanations that stress the CAT's lack of enforcement would seemingly suggest that *all* authoritarian regimes will sign. As mentioned above, this is not empirically the case. Moreover, it is unclear from either Hafner-Burton & Tsutsui (2007) or Hathaway (2007) why there exists a positive association between torture levels and the signing of the CAT by authoritarian regimes. It therefore remains an open question as to why those states with the worst human rights records sign these agreements with the greatest frequency and then ignore their obligations.

In the theory developed below, we concur with Vreeland's focus on the role of the interaction between autocratic governments and their domestic opposition as it affects the signing of human rights treaties. But we view this interaction quite differently. We assume a game is played between an office-seeking government and an opposition party. Their interaction is characterized by attempts to maintain (seize) power: the government can undertake costly measures to repress the opposition even as the opposition can take costly actions to remove

the government. We assume that the opposition is imperfectly informed as to the costs repressive measures impose on the government. We demonstrate that, in such a game, the government may use the signing of human rights treaties as a signal to the domestic opposition that they can repress at low cost.[6] In such an equilibrium, those governments that sign the treaty would torture more heavily *ex ante* than those that do not. Moreover, we find that signatory governments are likely to survive longer in office than non-signatories.

This logic may at first appear highly counterintuitive. However, it is in keeping with common perceptions of governments' defiance of international actors in situations not pertaining to human rights. For instance, it is widely believed that North Korea's recent nuclear test - despite international disapprobation - was meant to reinforce the regime's control following Kim Jung-il's ill-health and designation of a successor.[7] Since a 'weak' regime would be unable to face the international pressures stemming from such a test, this action is a credible signal of the regime's strength to a domestic audience. Similarly, it is often argued that the Castro regime in Cuba enhanced its domestic stability by provoking the United States. One could view these actions as signals meant to intimidate domestic political opponents. Our theory suggests that the signing, and willful defiance, of human rights treaties might play a similar role.

3 Theory

Article 4 of the CAT states that "Each State Party shall ensure that all acts of torture are offences under its criminal law." Moreover "[e]ach State Party shall make these offences punishable by appropriate penalties." Article 5 requires that any State Party to the CAT

[6]Our theory is, in some ways, analogous to the literature on audience costs (see, for instance, Fearon 1994, Smith 1998). Whereas audience cost theories often presume that failure to comply by an international agreement reveals negative information about a government's type (e.g. a lack of ability), we demonstrate that the willful defiance of an international agreement may be used to signal the government's 'strength.'

[7]Fackler, Martin. "Test Delivers a Message for Domestic Audience." *The New York Times*. May 25, 2009. `http://www.nytimes.com/2009/05/26/world/asia/26northk.html` - accessed November 7, 2009.

take into custody any alleged offender that is present in its territory. And Article 6 requires that, if requested to do so, any State Party must extradite the alleged offender to any state with jurisdiction over the case, which may be defined by the nationality of the perpetrator or the victim. If no such extradition occurs, the State Party must try the offender domestically.[8] Finally Article 8 further requires signatories to treat violations of the prohibition on torture as extraditable offenses.[9]

The CAT does, therefore, make torture a more serious offense.[10] Consider an autocrat inclined to torture in order extract information from or to punish a domestic political opponent. Should the autocrat, at some point in the future, find himself (and always it is himself, not herself) out of power, deposed or otherwise overthrown, the consequences will differ depending on whether the state was a signatory to the CAT. The usual act of a falling autocrat is to abscond to another country, if he manages to remain alive or out of jail. Assume that the autocrat's country were a Party to the CAT. If the country to which he escaped were also a CAT signatory, the autocrat's successor can demand the autocrat's extradition for trial for human rights violations. No such obligation would necessarily exist of the country is not a signatory to the CAT. On this basis, we argue that signing the CAT will at least weakly increase the penalties an autocrat would suffer after being evicted from office.

If an autocrat flees into exile - and if his state is unable or unwilling to try the him domestically - the now host nation, if it is a CAT signatory, has an obligation to try the ex-

[8]This requirement is often referred to as establishing 'universal jurisdiction' for human rights offenses.

[9]United Nations Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment. `http://www.hrweb.org/legal/cat.html`

[10]Some scholars question whether torture and other human rights offenses are currently covered under customary international law, and perhaps even enforceable in domestic courts (see e.g. Klein (1988)). Our point is that the CAT increases the potential costs of engaging in torture over and above that which might be expected under customary international law. Moreover, of the seven "core" human rights treaties, it may be argued that the CAT possesses the most serious enforcement mechanism. Goodliffe & Hawkins (2006) argue the CAT was the first treaty to apply the principle of universal jurisdiction to human rights law - jurisdiction is based "on the nature of the crime rather than .. where the crime occurred or the nationality of the alleged perpetrator or victim" (p.2) . As such, they suggest its enforcement mechanisms are more coercive than those of other human rights treaties or customary international law alone.

dictator for human rights offenses. It is reasonable to think that, if the state of the offending dictator had signed the CAT too, the pressures for arrest and indictment would be higher than if it were not a CAT signatory. And finally, a third state can demand extradition from a fellow CAT member if the now host country fails to try the alleged perpetrator.

These provisions increase the expected costs of torture substantially. In the event an autocrat is removed from office, the danger of extradition may significantly limit his possible destinations for exile. The long term costs of this restriction on his movement would be considerable. Clearly therefore, and contrary to much of the scholarship on the CAT, there are post-tenure liabilities associated with engaging in torture. While these might be perceived to be unlikely to occur, or might happen only the distant future, the costs from treaty violation are non-trivial in expected value. Moreover the probability that these liabilities will be applied increases as more countries apply the principle of universal jurisdiction - Goodliffe, Hawkins & Vreeland (2009) finds that 109 states have incorporated universal jurisdiction in domestic law, 14 have tried cases under the principle and the courts have enforced the law in 12 of them. What matters for our argument is not that it is always applied but that it might, effectively raising the expected costs of engaging in torture.[11] We model these punitive mechanisms as increasing of the marginal cost of engaging in torture or repression.

The costs imposed by the CAT have most vividly been illustrated in the extradition proceedings in the British House of Lords against Augusto Pinochet in 1998. Famously, the Law Lords ruled that Pinochet may be extradited to face criminal charges in Spain. Offenses after 1988 were ruled as extraditable, as 1988 marked the year that the UK ratified the CAT

[11]In the costly signaling model developed below, as the costs of treaty violation go to zero, all governments pool on signing. As the costs increase, only the more repressive governments are likely to sign. As noted above, we do not witness all governments pooling on signing the CAT. Moreover, Hathaway's (2007) empirical findings are consistent with punitive costs that exceed the minimum threshold for separation. To the extent that Goodliffe & Hawkins (2006) are correct regarding the relatively punitive enforcement mechanisms of the CAT, we would be more likely to see this pattern of behavior in CAT accession than in the accession to other human rights treaties.

and passed domestic implementing legislation (Roht-Arriaza 2001). This finding allowed the prosecution of Pinochet to proceed despite a negotiated amnesty with his successor regime. Moreover, the Spanish prosecution (with the consent of the UK Law Lords) catalyzed Chilean courts to permit filings and as many as 170 complaints were subsequently brought in Chilean courts (Jonas 2004, Roht-Arriaza 2001). Similarly, former Chadian dictator Hissène Habré is under house arrest in Senegal for CAT violations. Despite findings from both the UN Commission on Torture and the African Union that Senegal is obliged under CAT provisions to either extradite or try Habré for torture that took place while he was in office, Senegal appears to be dragging its feet. Belgium has sued Senegal at the International Court of Justice arguing that Senegal is violating the CAT by neither prosecuting nor extraditing him, effectively giving Habré asylum. There is no ruling from the ICJ as yet, but this provides another example that the CAT is preventing Habré from escaping to a villa on the French Riveira, and raising the personal costs of exile (ICJ 2009).

Paradoxically, the increased cost signing the CAT places on repression ensures that those countries that torture heavily are most likely to sign.[12] Assume that autocrats vary in the costs they face from engaging in repression and further assume that opposition groups are unable to perfectly observe these costs.[13] Those governments that can repress cheaply will be willing to engage in more torture than those that face higher costs. However, since all governments would like to intimidate the opposition, no government can effectively communicate

[12]Here - and throughout - we concentrate on the signing rather than the ratification of human rights treaties. We do so for two reasons: (1) The signing of treaties is the prerogative of the executive. Ratification may or may not (depending on the authoritarian system in question) be subject to the approval of other actors. (2) Ratification of a treaty follows its signing. Hence we argue that the act of signing a treaty likely carries the most informational content about the executive's costs of torture, rather than its ratification.

[13]It may be objected that opposition groups are aware of the costs a regime faces from repression and its willingness to employ draconian methods to remain in office. While such groups no doubt have some information in this regard, this information is not always perfect. The veterans of many successful opposition movements often express surprise at their successes. And many failed opposition groups undertake costly activities in the vain hope of removing the regime. These actions are most readily explained by imperfect information. Theoretically, governments would only be able to perfectly reveal their willingness to employ repressive tactics if they had a continuous array of credible signals at their disposal. So long as some uncertainty exists, their remains an incentive for low cost governments to signal their type.

whether it is truly a 'strong' or 'weak' type.

Since signing a human rights treaty imposes a cost on autocrats who torture - and only sufficiently 'strong' types would be willing to bear such a cost - signing such a treaty may act as a credible signal to the domestic opposition of a government's type. If this is true, it is those governments that can repress at low cost who sign the treaty and continue to torture. High-cost governments do not sign. Such behavior would seem consistent with existing empirical findings.

Moreover, those autocrats who sign such treaties will survive in office longer than non-signatories. A *selection effect* implies that those regimes that will fight most strongly to remain in power are the same regimes that sign the treaty. And an *information effect* implies that opposition groups - on learning that the state has signed the treaty and is therefore a strong state - will engage in fewer activities designed to overthrow a signatory government.

3.1 Model

We model the signing of a human rights treaty as the outcome of an interaction between an autocratic government G and its domestic opposition D. Both are assumed to be office-seeking: i.e. each derives some value from holding power $R > 0$. In the contest for power, the government may - at some positive cost - engage in repressive measures entailing human rights violations against the opposition. Similarly, the opposition may undertake costly efforts to remove the government. The outcome of the contest for power will be determined, in part, by each party's respective choice of repression and effort level.

The sequence of the game is as follows: First, nature chooses the type of government $\theta \in [0, 1]$, where θ represents the cost of repression.[14] This variable is observed by the government, but not by the opposition. Second, the government chooses whether or not to sign a human rights treaty $s \in \{0, 1\}$. Third, the opposition and the government simultaneously choose

[14] All results would be preserved if the government's type determined the value it places on office.

$e \geq 0$ - the level of effort put into deposing the government, and $t \geq 0$ - the level of repression. The choice of t is made at the constant marginal cost θ if $s = 0$ and $k\theta$ if $s = 1,\ k > 1$. Opposition effort e is chosen at cost be where b is a constant $b > 0$. Fourth, nature determines whether the government survives - with probability $\pi(t, e)$ - or not. All payoffs are realized and the game ends.

$\pi(t, e)$ is a standard contest success function (Hirschleifer 1991, Skaperdas 1996): $\pi(t, e) = \frac{t}{t+e}$.[15] For simplicity, we let the distribution of government types be defined by the uniform distribution $f(\cdot)$ with support over the unit interval. This distribution is common knowledge.

Player utilities are defined by their expectation of holding office and the choice of $s,\ t$ and e by the autocrat and the opposition respectively. The autocrat's expected utility function is:

$$U_G(t, e, s; \theta) = \pi(t, e)R - [sk + (1 - s)]\theta t;$$

while the opposition's is defined as

$$U_D(t, e, s) = [1 - \pi(t, e)]R - be.$$

The government enjoys the rents from office $R > 0$ with probability $\pi(t, e)$ and pays a cost for repression equal to $k\theta t$ if $s = 1$ and θt otherwise. The opposition, on the other hand, obtains R with probability $1 - \pi(t, e)$ and pays a cost for its anti-regime efforts of be.

3.1.1 Equilibrium

The game is solved through generalized backwards induction using the perfect Bayesian equilibrium concept. Our first proposition establishes that there is an equilibrium in which the strong types sign the agreement, and the weak types do not. In Appendix A (where all

[15] The assumption that the probability of government survival is increasing in repression and decreasing in opposition activities is central to our results. As discussed above, this assumption is inappropriate for democracies. We thus restrict our analysis to autocracies - for which this assumption is far more reasonable.

the proofs can be found) we define an invertible function $\Psi(x)$ for any $x > 1$.

Proposition 1. *If* $k > \frac{3}{2}$ *and* $b > \frac{k}{3}\sqrt{\Psi^{-1}(k)}$ *then there exists a unique semi-separating equilibrium where for* $\widetilde{\theta} = \Psi^{-1}(k)$

- *If* $\theta < \widetilde{\theta}$, $s(\theta) = 1$ *and* $e(1) = \frac{Rk\tilde{\theta}}{9b^2}$, $t(1,\theta) = \frac{R}{3b}\left[\sqrt{\frac{\tilde{\theta}}{\theta}} - \frac{k\tilde{\theta}}{3b}\right]$
- *If* $\theta > \widetilde{\theta}$, $s = 0$ *and* $e(0) = \frac{R}{9b^2}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)^2}{\left(1-\widetilde{\theta}\right)^2}$, $t(0,\theta) = \frac{R}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\widetilde{\theta}\right)}\left(\sqrt{\frac{1}{\theta}} - \frac{1}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\widetilde{\theta}\right)}\right)$

In words, this equilibrium implies that, for any human rights treaty that makes repression sufficiently costly ($k > \frac{3}{2}$), and if the opposition faces sufficiently high marginal costs to anti-government efforts ($b > \frac{k}{3}\sqrt{\Psi^{-1}(k)}$), relatively low-cost autocratic regimes will sign while relatively high-cost regimes will not. *Contra* standard selection arguments, this equilibrium posits that it is precisely those regimes that are least likely follow the treaty's provisions absent any agreement that choose to sign the human rights treaty. It is, however, consistent with the empirical evidence on the entry of authoritarian regimes into the CAT - those authoritarian regimes that torture more *ex ante* are more likely to sign.

The logic for this finding is straightforward. All autocratic governments seek to convince their opposition that they face a low cost to repression, as this will serve to reduce the level of effort the opposition will put into removing the autocrat. Signing a human rights treaty acts as a costly signal to the opposition of the government's low cost to repression. In equilibrium, the opposition learns that the government is a tough tpe, and that the marginal benefiit of its opposition actvity is lower than they had thought. In order to raise the marginal benefit of its activity, the opposition will reduce the its anti-regime effort e (since effort has declining marginal benefit). This both benefits the government directly - as it faces a lower probability of removal from office $\pi(t,e)$ - and indirectly, as it can reduce its level of (costly) repression. However, whatever repression it continues to practice has become more costly for the government ($k > 1$).

The government must, therefore, weigh the costs of treaty accession against the benefits of lower opposition effort. If the penalty the treaty imposes on human rights violations is low ($k < \frac{3}{2}$), all autocrats pool on signing. If k goes to infinity, no government will sign (the threshold $\tilde{\theta}$ goes to zero making all types non-signers). For values of k between $\frac{3}{2}$ and infinity however, some governments choose to sign and others do not. Low-cost repressors benefit more from any reduction in opposition effort than high-cost repressors. A low marginal cost of repression θ implies that the government is highly responsive to any change in opposition effort levels. Thus, for any decline in e, the a low cost government reduces t more if θ is low than if it is high. Therefore, for any value of $k > \frac{3}{2}$, it is the low-cost governments ($\theta < \tilde{\theta}$) that benefit more from signing the treaty than high-cost ($\theta > \tilde{\theta}$) ones.

From the equilibrium levels of repression practiced by the autocrat and effort exerted by the opposition, we can determine the probability of regime survival in equilibrium. We state these probabilities in the following Lemma 1:

Lemma 1. *In the semi-separating equilibrium, survival probabilities are given by the expressions* $\pi(t(1,\theta), e(1)) = 1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta}$ *for signatories and* $\pi(t(0,\theta), e(0)) = 1 - \frac{\sqrt{\theta}(1-\tilde{\theta}^{\frac{3}{2}})}{3b(1-\tilde{\theta})}$ *for non-signatories.*

This leads directly to the following result:

Proposition 2. *In the semi-separating equilibrium, signatories will survive (weakly) longer in office than non-signatories.*

The survival effect stems from two causes. First, a selection effect is evident from Proposition 1 - autocrats who can repress more readily are more likely to sign the treaty than those for whom repression is more costly. So the treaty selects those autocrats who would survive longer even in a world absent the CAT. But there is an additional causal effect of

the CAT on regime survival: domestic opposition declines on signing the CAT, enhancing leader survival.

In the two results that follow, we compare the opposition's effort levels, and the government's torture levels in two states of the world: the counterfactual - a world in which the CAT is not available as a signaling device, and the world with the CAT.

Proposition 3. *If the CAT is absent, opposition effort is increasing in in the cost of effort, b and government torture is decreasing in the type, θ.*

If there is no CAT, there is no opportunity for signaling. The domestic opposition makes its best guess about the type of government it is facing; given this level of domestic opposition, the cheaper is torture, the more torture the elite will undertake.

Combining the insights of Propositions 1 and 3, we can compare the behavior of states with and without the CAT. Firstly, the states that torture the most absent the CAT are the states most likely to sign the CAT when it becomes available to them - the types for whom torture is less costly. We therefore now have a theoretical foundation for Hathaway's (2007) unexplained observation that the worst torturers are more likely to sign the CAT. Simmons (2009) makes a similar empirical finding of a positive association between torture and CAT signing. As a first test then, the model makes a prediction that is consistent with *extant* empirical work.

We are now able to explore the consequences of the CAT on the level of domestic opposition and on the levels of torture .

Proposition 4. *In signatory states, domestic opposition is lower with a CAT than without; in non-signatory states, domestic opposition is higher with the the CAT than without.*

The CAT conveys information to the domestic opposition about the toughness of the government. A signer of the CAT is signaling its toughness, causing the domestic opposition to alter its optimal behavior. On learning the government is a tough type (since it signed

the CAT), the domestic opposition has learned that the marginal productivity of its effort is lower than it previously believed. Its optimal response to this new information is to reduce the amount of effort it incurs in order to raise its marginal productivity. Hence signatory governments face less unrest than would be the case if there was no CAT.

On the other hand, a state that does not sign is signaling weakness; the domestic opposition raises its effort accordingly.

Empirically, this suggests that on signing the CAT, domestic opposition efforts will decline in absolute terms and relative to efforts in non-signatory states.

A strong government, on signing, now expects less domestic opposition. Since the actions of the players are strategic complements, the government can reduce its torture levels, given it is facing a less severe revolutionary threat. This leads to the final proposition.

Proposition 5. *In signatory states, torture levels are lower with a CAT than without; in non-signatory states, torture is higher with the CAT than without.*

On signing the CAT we expect torture in signatory states to fall. Overall, the CAT reduces torture and the revolutionary threat in signatory states, and increases the survival of those autocrats in office. Note that the CAT does have a causal effect here: the change in the behavior of the domestic opposition induced by the CAT affects leadership survival over and above the selection effect - where those leaders most likely to survive ex ante are the signers of the treaty.[16] The agreement has the effect of lengthening the term of office of the worst offenders of the human rights regime.

[16]Note that the selection problem produced in measuring the effect of the treaty is precisely the opposite of that discussed by Downs, Rocke & Barsoom (1996) in regards to most international institutions - where signers self-select into the agreement because they intend to comply; here they self-select exactly because they do not intend to comply

3.1.2 Model robustness

In Appendix B (attached here, but to be posted online), we demonstrate that semi-separating equilibria with similar properties exist for a number of alternative model specifications. Perhaps most significantly, we demonstrate that an analogous equilibrium exists when signing the treaty only results in punishment in the event the government steps down from office - i.e. punishments are strictly post-tenure. As is true when signing the treaty increases the government's marginal cost of repression, there exists a semi-separating equilibrium wherein only those governments that can repress at low cost sign (Proposition B1).

The application of post-tenure punishments does produce an additional effect not present in the baseline model. As the level of post-tenure punishment rises, the signatory governments grow increasingly willing to employ repression to remain in office and allocate more resources to torture in equilibrium (Proposition B2). We term this result the *commitment effect*. Knowing this, the opposition will devote less effort to removing signatory governments when post-tenure punishments are large. The difference in survival times between signatories and non-signatories is thus increasing in the level of post-tenure punishments.

In model extensions where post-tenure punishments are allowed, we also find that semi-separating equilibria will exist regardless of whether governments vary in their cost to repression (as in the baseline model) or in the value they place on office. In the appendix, we prove the existence of a semi-separating equilibrium in which governments vary in the value they place on office (and this is private information to the government). Only those governments that benefit greatly from remaining in power are willing to sign the treaty; those who benefit less are not so-willing to sign (Proposition B3). This is to be expected. From the baseline model it is clear that governments devote greater effort to remaining in office as costs decline. In this extension, we simply shift consideration from variation in the cost-side to variation in the benefits side of governments' decision calculus. Thus, as is true in the baseline model, it is those governments willing to fight hardest to remain in power

- and who thus practice the greatest levels of repression absent the treaty - who are most likely to sign.

3.1.3 Examples

The equilibrium described above predicts that (1) those authoritarian regimes that torture most heavily will be most likely to sign the CAT, (2) signatory regimes will survive longer in office than observationally similar non-signatories, (3) authoritarian governments will reduce their levels of torture on signing the CAT, and (4) domestic opposition in authoritarian states declines on CAT signing.

These predictions - and the informational logic behind CAT accession - run counter to most *prima facia* expectations. However, several examples of authoritarian regimes that sign the CAT fit this logic rather well. For instance, Chad became a CAT signatory on June 9, 1995. The Chadian regime - headed by Idriss Déby - faced extensive armed opposition at the time, which it repressed through the extensive use of torture.[17] The following year, the Déby regime unveiled a new constitution, which controversially granted sweeping powers to the presidency. This constitution was adopted on March 31, 1996 and presidential elections - in which there were reports of extensive irregularities - followed soon after.[18] According to our theory, Déby's decision to sign of the CAT acted as a signal to opposition forces of his intention to cling to power. Following Propositions 2 and 5, Déby would be predicted to survive in office and reduce torture levels, even as the opposition would be expected to reduce its efforts at bringing about his ouster.

In fact, Déby did remain in power following elections in 1996 and remains in power currently. Torture levels in Chad declined in 1996 following the signing of the CAT, as is

[17] In 1995, Chad had a value of 4 on Hathaway's (2007) 5 point torture scale, and a 3 on CIRI's (2007) 3 point scale. (Here, and throughout, the CIRI scale is inverted such that higher values on the CIRI index correspond to the more widespread use of torture.) See also: James, Odhiambo. "Human Rights: Pattern Changes but Violations Continue. *Africa News*. August, 1995.

[18] 'Background Note: Chad.' US Department of State. Feb. 2009. `http://www.state.gov/r/pa/ei/bgn/37992.htm`

in keeping with Proposition 5.[19] And, in 1997, several armed opposition groups ended their insurgency through negotiations with the government, consistent with Proposition 4.[20]

Following a similar logic, many authoritarian regimes signed the CAT immediately following or preceding a transition of power.[21] For instance, the Museveni regime in Uganda signed the CAT in the year it assumed power.[22] Similarly, the Stevens government in Sierra Leone signed the CAT on March 18, 1985, immediately before handing power over to Stevens' chosen successor - Joseph Saidu Momoh - on November 28th of that year.[23] The theory predicts that signing the CAT sends an informative signal of a regime's willingness to cling to office. Logically, the period immediately surrounding a change in the head of a regime would be a period of great uncertainty regarding the incoming elite's willingness and ability to cling to power. This is also likely to be the period during which domestic opposition is particularly determined. As such, the informational value of signing the CAT is particularly great during transitional periods.

Of course, such results hardly constitute definitive evidence of the informational value of signing the CAT, though they are suggestive. Indeed, case studies are unlikely to provide strong support for our theoretical claims. We, in essence, argue that authoritarian regimes sign the CAT as part of an effort to deter opposition groups from undertaking anti-regime activities. As is argued by Achen & Snidal (1989), the use of case study evidence is problematic for assessing the effectiveness of deterrence. Our preferred tests of our theory therefore consists of a large-N analysis examining the claims of Propositions 2, 4 and 5.

[19]These ranked at a level of 2 on both the Hathaway and CIRI scales.

[20]'Background Note: Chad.' US Department of State. Feb.2009.
`http://www.state.gov/r/pa/ei/bgn/37992.htm`

[21]In our sample, just under 15 percent of authoritarian regimes that joined the CAT did so in a year of transition. This was the most commonly observed time of CAT signing in our sample.

[22]Museveni assumed power on January 29, 1986 and the Museveni regime signed the CAT on November 3, 1986. Museveni currently remains in power.

[23]Momoh remained in power until April of 1992.

4 Empirics

Since case study evidence is unlikely to provide conclusive in either supporting or falsifying our theory, we instead test the implications of the theoretical model above using large-N analyses. The model makes 4 testable predictions: (1) it is the most severe torturers that sign the CAT (Propositions 1 and 3); (2) those authoritarian signatories that sign will survive in office with higher probability than observationally similar non-signatories (Proposition 2); (3) opposition effort declines on CAT signing (Proposition 4); and (4) signatories will torture less after signing the CAT (Proposition 5).

The first prediction - the most severe autocratic[24] torturers are more likely to sign the CAT - has been substantiated by previous work (Hathaway 2007, Simmons 2009 and Hafner-Burton and Tsutui 2007), and we do not replicate those findings here. Instead we focus on the remaining empirical predictions 2, 3 and 4 above.

4.1 Data

In all sets of regressions, we employ data from Vreeland's (2008) dataset on CAT accession, the Archigos database on political leaders and regime survival (Goemans 2006), and Gandhi and Przeworski's (2007) data on the longevity of authoritarian regimes.[25]

Our measures of domestic unrest are derived from UCDP/PRIO Battle Deaths Dataset Version 3.0 (Lacina & Gleditsch 2005). We make use of annually observed observations of battle deaths resulting from civil wars, which can be viewed as a function both of opposition efforts to remove the government and of government repression of the opposition. We also

[24] Autocratic regimes are identified following the definition advanced in Przeworski et al. (2000). Thus, autocracies are states which either lack executive or legislative elections, in which there exists either one party or no parties, or in which the ruling party has never been removed from power.

[25] Throughout our analysis, we define a 'regime' as a single leader's tenure in the Archigos dataset. In this our analysis differs from both Hathaway's and Vreeland's. Both of these authors use countries as subjects and the country-year as the unit of observation. We use both the regime and the regime-year in sections 4.2.1 and 4.2.2 respectively.

rely on data from Banks' Cross-National Time-Series Data Archive (Banks 1979) (drawn from the dataset made available by Bueno de Mesquita et al. (2003)) to measure riots, strikes, revolutions, demonstrations, and other anti-government activities. These variables we view as measures of oppositional activity aimed at displacing the government.

4.2 Leader Survival

We first subject the claim that signatories of the CAT enjoy more secure tenures in office than non-signatories - derived in Proposition 2 above - to empirical scrutiny. Note that this claim is *not* equivalent to the statement that signing the CAT causes an authoritarian regime to survive longer in office. Our prediction is that signatories face a lower probability of removal due, in part, to a selection effect. Signatories face a systematically lower cost to repression than non-signatories. Signatories also benefit from the revelation of information to opposition groups. Precisely because only low-cost repressors choose, in equilibrium, to sign the CAT, these regimes face a less restive domestic opposition. Domestic opposition groups are less willing to exert costly effort against regimes that reveal that they are low cost types by signing the CAT. Proposition 2 holds as a result of the cumulative force of these selection and information effects. Our tests of this proposition cannot differentiate between the two effects.

4.2.1 Single Record Cox Estimates

To test the association between the signing of the CAT and the survival time of regimes, we first run a Cox proportional hazards model of the probability of regime failure. The unit of observation in this model is the regime. Time is defined by sumten the sum total number of days a given regime served in office, as taken from the Archigos dataset. The key explanatory variable is Eversign, a binary indicator variable that takes the value of 1 if a given regime is

ever a signatory of the CAT.[26] 129 authoritarian regimes are observed during the 1985-1996 period.[27],[28]

The Cox model provides an estimate of the hazard rate of a given regime i (the probability regime i collapses at time t given that it has survived until time t) conditional upon observed covariates: $h_i(t) = h_o(t)e^{X_i\beta}$, where $h_o(t)$ is the baseline hazard function. The Cox model makes no assumptions about the parametric form of the baseline hazard function, which is estimated non-parametrically based on the exit times of regimes in the dataset. As our theory makes no predictions regarding the effect of time-in-office on regime survival, we treat time as - in effect - a nuisance parameter.[29] The results of this regression are reported in Table 1.

The controls include the variables identified as significant to the survival of authoritarian regimes by Gandhi & Przeworski (2007) - the number of changes in the executive during a given authoritarian spell, whether or not the executive has a civilian background, whether or not the government inherited one or more opposition parties, and an indicator for resource dependence.[30] We also control for variables relevant to the signing the CAT identified by Vreeland (2008) and Hathaway (2007), notably the average level of torture employed by the

[26]The use of a similar variable Everratify produces estimates of similar magnitude and direction; though the coefficient is not significant at conventional levels.

[27]These data are subject to a left-censoring problem. The sumten variable takes on values such that some regimes begin life before they come under observation. This censoring is unlikely to effect our results, as regimes that sign the CAT do not significantly vary in their age (at the time of signing) from those that do not. There is also a right-censoring issue, insofar as we do not observe all regimes that will sign the CAT. Since regimes that go on to sign the CAT after the sample ends are expected to have a lower hazard rate than those that will never sign the CAT, this would be expected to bias our results downwards.

[28]The period under observation begins when the CAT comes into force. The 1985-1996 period is identical to that covered in the Vreeland (2008) dataset.

[29]The Cox model does assume hazard rates are proportional - that the shape of the hazard function does not differ across units. We test this assumption using the Grambsch-Therneau of the Schoenfeld residuals (Box-Seffensmeier & Jones 2004). We do not reject the null that the hazard functions are proportional. However, Harrell's rho statistics, testing individual covariates for violations of the proportional hazards assumption, suggest that the measures of torture used should be interacted with time. We include these interactions in our specifications below. Results are substantively unchanged if these interactions are not included.

[30]This indicator takes the value 1 if primary commodity exports exceed 50 percent of total exports.

regime,[31] an indictor variable that takes the value one if multiple parties are allowed to exist, and the population (in millions). Finally, we add a control for the military capabilities index compiled by the Correlates of War project, averaged over each regime.(Singer 1987) As can be readily seen in Table 1, regimes that are signatories of the CAT have lower hazard rates than those that do not.[32] This difference is apparent even after controlling for other factors related to regime survival and for factors related to CAT accession.

It may be objected that our definition of the explanatory variable of interest in these models is problematic. The Eversign variable takes a value of 1 for *all* regimes that are signatories of the CAT, even those that inherit their signatory status from a predecessor government. It may reasonably be argued that such regimes are unlikely to withdraw from the CAT, regardless of their willingness to employ repression to stay in power. Choosing to remain in the CAT sends a very different signal than choosing to sign the CAT. If it is true that regimes that inherit their signatory status are unwilling to remove themselves from the treaty, then our results in Table 1 will be biased downwards. We therefore drop all regimes that inherit their signatory status and rerun the model above. Results are reported in Table 2.

The sign on the Eversign variable remains negative and increases in value and in significance. These results are consistent both with Proposition 2 and with the claim that regimes that inherit their signatory status are unlikely to remove themselves from the CAT.[33]

[31] We employ both the Hathaway (Hathaway 2007) and CIRI (Cingranelli & Richards 2007) measures of torture. Both are based on data made available by the US Department of State and by Amnesty International. The Hathaway measure is an ordinal index running from 1 to 5 with higher values indicating the more extensive practice of torture. The CIRI index is an analogous measure with values running from 1 to 3. Both are drawn from the dataset used in Vreeland (2008).

[32] The reported coefficients are not hazard ratios. A coefficient of zero implies that the variable in question has no effect on the hazard rate, negative coefficients imply that an increase in the variable reduces the hazard rate, and positive coefficients imply the reverse. We have also run models wherein the **Eversign** variable is interacted with the indicators for regime-type. These estimates do not indicate any significant difference in the relationship between CAT accession and survival between military and non-military regimes.

[33] We have also run robustness checks on the single record model by (1) dropping all regimes that survive for less than either one or two years from the dataset, and (2) running a propensity score matching algorithm to ensure that the data are balance with respect to CAT signing. These results do not change the direction

Table 1: Coefficient Estimates from a Single Record Cox Model

	Hath1	**CIRI1**	**Hath2**	**CIRI2**
Eversign	-.662	-.715	-.661	-.699
	(.305)**	(.305)**	(.282)**	(.293)**
Torture	.232	-.299	.361	.889
	(.233)	(.261)	(.228)	(.477)*
Torture*time	-.0001	-.0002	-.0001	-.0002
	(.00004)***	(.00007)**	(.00003)***	(.00006)***
No. Changed in Exe.	.146	.136	.133	.117
	(.032)***	(.031)***	(.034)***	(.037)***
Civilian Exe.	.042	.07	-.033	-.054
	(.271)	(.26)	(.289)	(.297)
Inherited Opposition Party	.303	.321	.4	.387
	(.23)	(.224)	(.25)	(.242)
Multiple Parties	-.536	-.662	-.555	-.633
	(.336)	(.33)**	(.346)	(.334)*
Resource Dependence	.416	.462	.203	.158
	(.422)	(.441)	(.478)	(.463)
Population	.004	.006	.	.
	(.007)	(.008)		
COW Capabilities Index	40.132	12.679	.	.
	(67.525)	(68.319)		
GDP *per capita*	-.091	-.074	.	.
	(.045)**	(.04)*		
No. Subjects	129	129	129	129
No. Failures	71	71	71	71

Results from a single record Cox Survival analysis of regime tenure. Hazard rates estimates based on number of days in office as measured by the Archigos dataset. Estimates are constructed for a sample of 129 authoritarian regimes in the years 1985-1996. Coefficient estimates are of the form $h_i(t) = h_0(t)e^{\mathbf{X}\beta}$. **Hath** models use Hathaway's (2007) torture index; **CIRI** models use the CIRI (2007) index. All standard errors are clustered by country. * denotes significance at the 90 percent level, ** at the 95 percent level, and *** at the 99 percent level.

Table 2: Coefficient Estimates Dropping Regimes that Inherited Signatory Status

	Hath1	**CIRI1**	**Hath2**	**CIRI2**
Eversign	-1.004	-1.076	-.985	-1.032
	(.372)***	(.352)***	(.337)***	(.328)***
Torture	.398	1.31	.481	1.323
	(.291)	(.586)**	(.268)*	(.524)**
Torture*time	-.0001	-.0002	-.0001	-.0002
	(.00003)***	(.00006)***	(.00003)***	(.00006)***
No. Changes in Exe.	.142	.122	.125	.103
	(.039)***	(.039)***	(.045)***	(.046)**
Civilian Exe.	-.028	-.028	-.122	-.156
	(.314)	(.289)	(.343)	(.319)
Inherited Opp. Party	.238	.236	.313	.329
	(.261)	(.261)	(.277)	(.269)
Multiple Parties	-.552	-.774	-.515	-.658
	(.375)	(.382)**	(.368)	(.359)*
Resource Dependence	.309	.227	-.06	-.127
	(.454)	(.463)	(.508)	(.498)
Population	.006	.011	.	.
	(.008)	(.009)		
COW Capabilities Index	18.847	-55.968	.	.
	(87.747)	(96.877)		
GDP *per capita*	-.111	-.079	.	.
	(.054)**	(.044)*		
No. Subjects	108	108	108	108
No. Failures	62	62	62	62

Results from a Cox Proportional Hazards estimate of regime tenure. Hazard rates estimates based on number of days in office as measured by the Archigos dataset. All regimes that inherited their status as a signatory government from a previous regime are dropped from the sample. Estimates are constructed for a sample of 108 authoritarian regimes in the years 1985-1996. Coefficient estimates are of the form $h_i(t) = h_0(t)e^{\mathbf{X}\beta}$. **Hath** models use Hathaway's (2007) torture index; **CIRI** models use the CIRI (2007) index. All standard errors are clustered by country. * denotes significance at the 90 percent level, ** at the 95 percent level, and *** at the 99 percent level.

These results *cannnot* be interpreted causally. It is very possible that our model results are being driven by a selection effect - particularly selection on unobserved covariates. However, Proposition 2 posits selection on covariates unobserved to the opposition party. The simple binary relationship - that those regimes that sign the CAT survive in office longer than those that do not - is consistent with model predictions.

4.2.2 Multiple Record Survival Analysis

In addition to the single-record model analyzed above, we run a multiple-record Cox survival analysis of the relationship between CAT signing and authoritarian regime survival. The unit of analysis in this model is the regime year. Time is defined by the difference between the year in which a regime took office and the current year.[34] The explanatory variable of interest is Lagged CAT Signing, a binary indicator variable that takes the value 1 in the year following a CAT signing. The model thus tests if signing the CAT in time t increases regime stability in year $t+1$. 116 authoritarian regimes are observed over the 1985-1996 period.

As in section 4.2.1, we run a Cox proportional hazard model of the form $h_i(t) = h_0(t)e^{\mathbf{x}\beta}$. We have tested the proportional hazards assumption using Grambsch-Therneau tests of the Schoenfeld residuals. The tests do not reject the null hypothesis that the proportional hazards assumption holds.[35]

Table 3 reports coefficient estimates from this model. Columns marked **Hath** control for the Hathaway (2007) measure of torture, whereas those marked **CIRI** control for the CIRI (2007) measure. In all models the coefficient on Lagged CAT Signing is negative, implying that signing the CAT in year t reduces a regime's hazard rate in year $t+1$. These estimates are

of the coefficient estimates reported above, and all results are significant at the 10 percent level or above. Results are available from the authors on request.

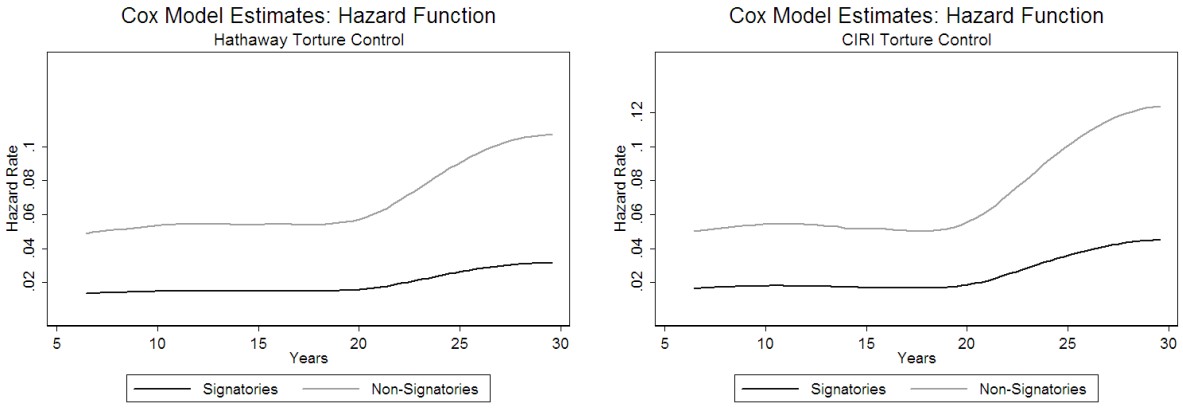
[34] Data is stset in Stata to adjust for problems of both left and right-censoring. See Box-Seffensmeier & Jones (2004) for more details on censoring.

[35] Harrell's rho statistics on parameter specific violations of the proportional hazards assumption are bordering on significant for `cinc`, `war` and `population` across both specifications. Including time interactions with these terms does not substantively affect the coefficients of interest. Results are available from the authors on request.

significant in all but one specification.[36] These results offer robust support for Proposition 2.

Since the Cox model is non-linear, the coefficients reported in Table 3 can be difficult to interpret. To give a better sense of the strength of the association implied by these estimates, we plot the baseline hazard function when Lagged CAT Signing is equal to 0 and when it is equal to 1 in Figure 1.

Figure 1: Hazard Function Estimates



Hazard function plots based on the models reported in Table 3 above. Plots using the Hathaway measure of torture as a control variable are plotted to the left, those using the Ciri measure of torture are plotted to the right. Hazard rates are depicted on the y-axis, while time in office (in years) is plotted on the x-axis. Signatories are depicted by the more darkly shaded line, non-signatories by the more lightly shaded line. All other covariates are set to their mean levels to produce this estimate.

It is also worth noting that the coefficient on Torture is negative and significant at the 10 percent level in all multiple record specifications. These results are consistent with the assumptions of the model - that regimes employ repressive tactics to maintain their hold on power.

[36]The level of significance in this specification $p = 0.13$ borders on significance.

Table 3: Coefficient Estimates from a Multiple Record Cox Model

	Hath 1	**Hath 2**	**CIRI 1**	**CIRI 2**
Lagged CAT Signing	-1.328	-1.418	-1.142	-1.248
	(.696)*	(.667)**	(.753)	(.742)*
Torture	-.385	-.349	-.493	-.468
	(.221)*	(.194)*	(.266)*	(.25)*
No. of Changes in Exe.	.159	.169	.139	.149
	(.045)***	(.042)***	(.041)***	(.042)***
Civilian Exe.	-.327	-.073	-.386	-.151
	(.369)	(.317)	(.389)	(.325)
Inherited Opp. Party	.264	.	.269	.
	(.28)		(.29)	
Multiple Parties	-.272	.	-.271	.
	(.383)		(.377)	
Resource Dependence	.016	.	.101	.
	(.484)		(.478)	
Growth	-.018	-.018	-.019	-.019
	(.008)**	(.008)**	(.009)**	(.008)**
GDP *per capita*	-.007	.	-.023	.
	(.04)		(.04)	
COW Capabilities Index	79.097	34.521	71.092	25.545
	(75.767)	(44.924)	(75.674)	(47.662)
War	.869	.758	.987	.796
	(.398)**	(.329)**	(.454)**	(.376)**
Population	-.006	.	-.007	.
	(.009)		(.009)	
No. of Subjects	113	116	112	116
No. Failures	41	43	41	42

Results from a Cox Proportional Regime survival function estimates. Survival function estimates estimates based on number of years in office as measured by the Archigos dataset. Estimates are constructed for a sample of 90 authoritarian regimes in the years 1985-1996. Coefficient estimates are of the form $h_i(t) = h_0(t)e^{\mathbf{X}\beta}$. **Hath** models use Hathaway's (2007) torture index; **CIRI** models use the CIRI (2007) index. All standard errors are clustered by country. * denotes significance at the 90 percent level, ** at the 95 percent level, and *** at the 99 percent level.

4.3 Opposition Effort and Government Repression

While the above results offer strong support for a novel prediction generated by our model, they do not constitute direct tests of the mechanisms driving our results. We claim that an authoritarian government's decision of whether or not to sign the CAT can be modeled as a costly-signaling game. Since regimes that need to fear loosing their grip on power (i.e. those regimes that face a high cost to repression) face the greatest expected cost of punishment under the CAT, these regimes will opt not to sign. Low-cost repressors, on the other, on the other hand, will be willing to sign. As a result, the opposition will cue off of the government's signatory status, and will reduce their costly anti-regime efforts on witnessing a government sign the CAT. As opposition groups reduce their efforts at removing a signatory government, the government lessens its costly attempts to repress the opposition.

It therefore follows that signing the CAT should be associated with a reduction both in opposition efforts to remove the government and in government efforts to suppress the opposition. However, both of these terms are extremely difficult to accurately measure. As a proxy, we first turn to the UCDP/PRIO estimates of battle deaths suffered during civil wars. While battle deaths are not a direct measure of opposition effort, we can infer something about opposition effort by interpreting battle deaths suffered during civil wars as a joint function of both government repression and opposition activity that is increasing in both arguments. More opposition activity to remove an autocrat from office is likely to be correlated with more battle deaths during civil wars.

The UCDP/PRIO dataset contains estimates of the number of battle deaths suffered in all wars with at least 25 fatalities between 1946 and 2005. Our data contain battle deaths estimates from wars UCDP/PRIO classifies as civil wars (types 3 and 4 in the PRIO data). We rely only on observations for which annual battle deaths estimates are available.[37] And

[37]In some instances UCDP/PRIO estimates the total battle deaths from a given war and divides these deaths evenly over the period during which the war took place. Such observations are treated as missing values in our dataset.

we code all observations wherein no civil war was active (i.e. fewer than 25 battle deaths) as having zero battle deaths.

Our causal claim is that the signing of the CAT should lead to a decline in the number of battle deaths experienced in the following year. To test this claim, we employ a difference-in-differences specification $\Delta\ battledeaths_{i,t} = \gamma \Delta\ CATsignatory_{i,t-1} + \mathbf{\Delta X_{i,t}}\beta + \epsilon_{i,t}$ where i denotes country i, t denotes year t and Δ is the difference operator. The difference operator ensures that all unit specific effects that are constant between year t and $t-1$ are controlled for in the regression estimates.[38] All specifications include controls for military capabilities, GDP *per capita*, and economic openness as well as for a cubic polynomial of time. We estimate this model using both the high and low estimates for battle deaths from the UNCP/PRIO dataset, and employ a seemingly unrelated regressions model, given the likely correlation of the error term across these estimates.

In addition to estimating this model on the full time-series cross-section of country-years in our dataset, we also pre-process our data using propensity score matching. Pre-processing the data in this manner is recommended as a measure to ensure covariate balance and reduce model dependence, particularly when working with a binary regressor (Ho et al. 2007). Since we are attempting to match panels - rather than individual observations - we employ the method of Simmons & Hopkins (2005). For countries that eventually sign the CAT, we collapse all pre-signing observations. For countries that do not, we collapse all observations over the full 1985-1996 period. We then estimate the probability that a given country ever signs the CAT based upon the covariates in the collapsed dataset. Panels are matched to one another using the MatchIt package (Ho et al. 2004) run from R 2.7.1. We employ the genetic matching algorithm of Diamond & Sekhon (2006). The resultant weights are then merged with the full dataset.[39]

[38] For this reason, we do not include the Gandhi & Przeworski (2007) controls in these specifications, as these measures are largely constant over time.

[39] Matching diagnostics are available from the authors on request.

Table 4: CAT Signing and Civil War Battle Deaths: Seemingly Unrelated Regression Model

	Matched Dataset		Full Dataset	
	Δ PRIO Low Battledeaths Estimate			
Lagged CAT Signing	-623.486 (183.765)***	-610.555 (183.84)***	-582.804 (142.733)***	-479.823 (124.52)***
change Military Capabilities	39.279 (149.752)	-35.824 (140.141)	10.625 (87.904)	-20.987 (74.277)
change GDP *per capita*	-171808 (208743)	-150511.5 (208494.4)	-34204.07 (54874.4)	-2203.521 (17050.05)
change growth	-6.617 (4.753)	.	-3.419 (2.805)	.
change openness	1.161 (3.86)	.	.847 (2.212)	.
Cubic Time Polynomial	✓	✓	✓	✓
	Δ PRIO High Battledeaths Estimate			
Lagged CAT Signing	-2092.317 (864.664)**	-2038.021 (863.805)**	-1872.469 (715.704)***	-1470.178 (645.406)**
change Military Capabilities	97.228 (704.625)	-167.57 (658.477)	-189.672 (440.777)	-347.977 (384.989)
change GDP *per capita*	-517324.4 (982191.6)	-432419.2 (979650)	-107367.3 (275156.3)	-2685.523 (88373.34)
change growth	-23.354 (22.365)	.	-19.158 (14.067)	.
change openness	-.298 (18.162)	.	-9.189 (11.094)	.
Cubic Time Polynomial	✓	✓	✓	✓
N	333	333	493	548

Results from a difference-in-differences model run using the UNCDP/PRIO measures of battle deaths. Estimates are from a seemingly unrelated regressions model. Coefficients to the left are estimates from models run on the matched dataset; those on the right are estimates from models run on the full dataset. * denotes significance at the 90 percent level, ** denotes significance at the 95 percent level, and *** denotes significance at the 99 percent level.

The results of these models - run on both the matched and unmatched datasets - are reported in Table 4. The coefficient on the lagged change in CAT signatory status is negative and highly significant (either at the 1 percent or 5 percent level) in all regressions. The association is also substantively large. Changes in the low estimates of battle deaths in the full sample had a mean of approximately -32 and a standard deviation of approximately 856. Thus, the signing of the CAT is associated with a roughly 0.7 standard deviation decline in the low estimates of battle deaths.[40]

As a robustness check of our estimates reported in Table 4, we run a similar model using the Banks (1979) measures of regime instability. Included in these measures are the number of assassinations, strikes, government crises, riots, revolutions, opposition demonstrations, and guerrilla movements in a given country in a given year. The models are identical to the above in all but the regressand.

The Banks data, it is important to note, are derived from reports in the *New York Times*. As a result, they are likely subject to a substantial degree of measurement error of a particular kind. Since the *Times* is unlikely to report on a strike or demonstration that did not in fact take place, it is reasonable to assume that any measurement error is likely to result in an *underreporting* of events. Such underreporting is likely to bias our difference-in-differences estimator towards zero.[41] As such is the case, it would be particularly surprising to find any

[40]The identification of these effects relies on a very small pool of countries that were experiencing a civil war at the time they signed the CAT. The results are thus sensitive to the exclusion of outliers. To adjust for this problem, we ran a fixed-effects OLS specification, with the regressor of interest a dummy that takes the value of 1 if the country-year is a CAT signatory. These results are consistent with the difference-in-differences estimate and indicate that CAT signatory status is associated with a substantial and significant decline in battle deaths. Moreover, these results are not sensitive to outlier observations (dropping a single observation never causes the results to decline in significance below the 95 percent level).

[41]Assume, by way of example, that every event has a 95 percent chance of appearing in the *Times* data, and that the probability of observation is independent across events. Let us say that there are 5 riots in a signatory country in year $t-1$ and this number is reduced to 3 riots in the year after signing the CAT . The expected observed difference is given by $.95 * 5 - .95 * 3 = 1.9$, which is strictly less than the true difference of 2. Conversely, a non-signatory government that experience the same number of riots in year $t-1$ and did not experience a reduction in coups stemming from signing the CAT would have an expected observed difference of 0.

systematic relationship between CAT signing and opposition activity using this data.

Our estimates from these difference-in-differences models are reported in Table 5. As was true with the battle deaths data, we employ a seemingly unrelated regressions model to adjust for the likely correlation in the error terms across these specifications.

As might be expected, our coefficient estimates are rarely significant (the lone exceptions are for two models testing the association between CAT signings and changes in the number of strikes). However, the coefficient estimates are nearly uniformly (23 of 28) negative - i.e. in the direction postulated by our theory. Moreover, the coefficients are often substantively large. For instance, the coefficient on strikes suggests that CAT signing is associated with a 0.5 standard deviation decline in the number of strikes from one year to the next. The coefficient on revolutions suggests a 0.2 standard deviation decline in the change in the number of revolutions. Our interpretation of these results, therefore, is that the data weakly support our theory. Given the substantial problems with these data, consistently negative sign and non-trivial magnitude of our coefficient estimates are unlikely to be purely the result of chance.

4.4 Change in Torture Levels

As a final test of the observable implications of our theory, we examine the claim that torture levels fall in CAT signatories relative to prevailing levels before the CAT is signed. This claim is advanced in Proposition 5 of the model. If signing the CAT acts as a signal that the government is a 'strong' type, a rational opposition will reduce costly anti-regime efforts in response to this information. Since repression and oppositional effort are strategic complements, levels of repression are predicted to similarly decline - though remain greater than zero.

To conduct our empirical tests of this claim, we rely on the torture indexes created by Hathaway (2007) and Ciri (2007). The former is an ordinal index ranging from 1 to 5, with

Table 5: CAT Signing and Levels of Unrest: Seemingly Unrelated Regression Model

	Matched Dataset		Full Dataset	
	Controls	No Controls	Controls	No Controls
Δ Assassinations	.067	-.002	.044	.012
	(.205)	(.193)	(.168)	(.12)
	Controls	No Controls	Controls	No Controls
Δ Strikes	-.275	-.219	-.204	-.157
	(.157)*	(.148)	(.126)	(.085)*
	Controls	No Controls	Controls	No Controls
Δ Gov't Crises	-.048	-.035	-.027	-.016
	(.098)	(.097)	(.088)	(.071)
	Controls	No Controls	Controls	No Controls
Δ Riots	-.112	-.141	.044	-.01
	(.381)	(.366)	(.318)	(.222)
	Controls	No Controls	Controls	No Controls
Δ Revolutions	-.15	-.162	-.146	-.147
	(.205)	(.19)	(.205)	(.146)
	Controls	No Controls	Controls	No Controls
Δ Demonstrations	-.023	-.045	.094	-.134
	(.414)	(.405)	(.353)	(.281)
	Controls	No Controls	Controls	No Controls
Δ Guerrilla Movements	-.077	-.065	-.063	-.064
N	328	365	497	743

Coefficient estimates on a lagged CAT signing indicator from a difference-in-differences model run using the Banks measures of domestic unrest. Estimates are from a seemingly unrelated regressions model. Regressands are reported in the left column, coefficient estimates and standard errors in the four other columns. Coefficients to the left are estimates from models run on the matched dataset; those on the right are estimates from models run on the full dataset. * denotes significance at the 90 percent level, ** denotes significance at the 95 percent level, and *** denotes significance at the 99 percent level. Columns marked 'Controls' include first-differenced measures of the growth rate of GDP, the level of GDP, the level of military capabilities, and the degree of trade openness. All estimates include controls for a cubic polynomial of time.

higher values corresponding to the more widespread use of torture. The latter is a similar index with values ranging from 1 to 3. Both indexes rely on information from Amnesty International and the US Department of State to derive their torture scales.

Of course, the extent of torture is inherently difficult to measure. The practice of torture is unlikely to be fully documented and highly repressive regimes are unlikely to be to be particularly transparent to either Amnesty International or the US Department of State. Both the Hathaway and Ciri data thus no doubt suffer from a good deal of measurement error.

Since these indexes have minimal and maximal values - and vary only over a restricted range - this measurement error *cannot* be normally distributed with mean zero. A country-year that should be classified as a 1 on the Ciri index can *only* be misclassified as having a higher level of torture than actually practiced and, conversely, a country-year that should be classified as a 3 can only be misclassified as having a lower level of torture than actually practiced. As was true of measurement error in the Banks dataset, errors in the Hathaway and Ciri measures will bias any results towards zero.

As with the analyses above, we run a difference-in-differences estimation to assess whether CAT signings are associated with a decline in the prevalence of torture relative to prevailing existing practices $\Delta torture_{i,t} = \gamma \Delta CATsignatory_{i,t-1} + \mathbf{\Delta X_{i,t}} \beta + \epsilon_{i,t}$ where i denotes country i, t denotes year t, and Δ is the difference operator.[42] Controls include the change in the military capabilities index, the change in GDP *per capita*, the change in the rate of economic growth, the change in levels of economic openness, and a cubic polynomial of time. These models are run both on the full panel of observations and on a panel consisting only of countries that eventually sign the CAT. The latter estimates rely only on time-series

[42]In using a first difference operator on the torture indexes, we implicitly assume that these are interval level measures - i.e. that movement from a 1 to a 2 is equivalent to movement from a 2 to a 3 on either index. We run analogous specifications using a trichotomous $\{-1, 0, 1\}$ variable indicating a decline, no change, or an increase in the level of torture, respectively. All results are similar to those reported below; though the level of significance declines.

Table 6: CAT Signing and Torture Levels

	Signatories		Full Dataset	
	Δ Tort. Hathaway	Δ Tort. Ciri	Δ Tort. Hathaway	Δ Tort. Ciri
Lagged CAT Signing	-.503 (.301)*	.14 (.277)	-.5 (.275)*	-.034 (.257)
change Military Capabilities	-.096 (.205)	.11 (.194)	-.073 (.102)	.111 (.115)
change GDP *per capita*	51.513 (52.163)	35.232 (61.851)	16.142 (60.291)	58.678 (63.075)
change growth	-.003 (.006)	.0007 (.007)	-.001 (.006)	-.002 (.004)
change openness	-.008 (.004)**	-.018 (.005)***	-.002 (.005)	-.008 (.004)**
Cubic Time Polynomial	✓	✓	✓	✓
N	262	259	514	493

Coefficient estimates from an OLS regression of changes in torture levels on changes in CAT signatory status (a difference-in-differences specification). Models to the left are run on a panel of country years containing countries that eventually sign the CAT. Models to the right are run on the full panel of country-years. * denotes significance at the 90 percent level, ** denotes significance at the 95 percent level, and *** denotes significance at the 99 percent level.

variation amongst CAT signatories - signing the CAT is assumed to lower torture levels in the year after signing and have no effect on the rate of change thereafter. Reports from all specifications are reported in Table 6.

The sign on the lagged change in CAT signatory status is negative in three of the four specifications, and is significant at the 90 percent level when the more fine-grained Hathaway measure is used. The magnitude of the coefficients is fairly consistent across both the panel of signatories and the full dataset. These findings are weakly supportive of the claim that signing the CAT leads to a reduction in levels of torture. Given the measurement problems documented above, it is surprising that any relationship is visible, yet alone that it rises to the level of significance when the Hathaway measures are used. Given that the Hathaway

index contains more information than the Ciri, it is not surprising that the former measure produces more significant results than the latter.

The empirical results above are thus consistent with theoretical predictions. There is robust evidence that signing the CAT is associated with an increase in the survival time of authoritarian regimes. Our results also suggest that opposition efforts decline as a result of CAT signings - consistent with the information effect postulated in the model. Our findings are particularly strong with respect to battle deaths; though the Banks unrest data are also weakly supportive of our theoretical claims. Finally, we find suggestive evidence that signing the CAT leads to a reduction in levels of torture. Given the substantial problems of measurement of many of these concepts, these results are strikingly consistent. Moreover, our theory is fully consistent with the earlier findings of Hathaway (2007) and Vreeland (2008) that authoritarian regimes that torture more frequently are more likely to sign the CAT than those that torture less.

5 Alternative Explanations for the Findings

If the CAT does indeed raise the costs of engaging in torture, consider an alternative story: the treaty acts instead as a mechanism for tying the hands of a repressive government - the treaty is a credible commitment by the government to reduce torture in exchange for reduced opposition efforts by the opposition. Government trades away torture in exchange for increased survival in office.

There are two problems with this interpretation. Firstly, as mentioned above, there is no mechanism for the domestic opposition to commit to reducing its efforts when the government ties its hands with the treaty. What is tying the hands of the domestic opposition? Secondly, in such a model, it is the weaker types, the types with the highest costs of torture that have the most to gain from such an arrangement. So this story would require the weaker types

to be most likely to sign the treaty. Instead we find the that it is the most severe torturers that sign, the states with the lowest costs of engaging in torture.

Notice that the argument made here is also a "tying-of-the-hands" story. The treaty raises the costs of violations, and hence is a credible commitment of the tough type's willingness to hold onto office - hence it is the tough types that sign the treaty.

A second alternative explanation questions the utility of the CAT to act as a signaling device. For instance, do we need the CAT in order for the elite to successfully signal its toughness? A tough autocrat could torture in excessive amounts in early periods, attempting to communicate its toughness to all observers; the domestic opposition might learn it is facing a tough opponent and reduce its efforts accordingly. Hence the the torture itself is a credible signal of the government's type. This would be consistent with the observation that it is the tough types that survive in office longer.

This claim rests on the notion that early period torture levels can separate out the types; but there will be incentives for even weak types to try mimic the strong types in the early periods in order to try to convince the domestic opposition that they are indeed tough. Hence, as in the canonical entry deterrence game, all types would pool on the same signal. Hence early period torture levels are unlikely to act as a signal that would permit separation of types.

Undoubtedly alternative signaling mechanisms to exist in order for governments to credibly signal their toughness. We do not argue that the CAT is unique in this regard; we do argue however, that the CAT does play this role, with unintended consequences.

6 Conclusion

Autocracies that torture more are more likely to sign the CAT than those that torture less; autocracies that sign the CAT continue to torture; and overall the CAT reduces torture

levels in signatory states. Furthermore, those autocrats who sign the CAT survive longer in office than those that do not, and that oppositional activities in signatory states fall when the CAT is signed.

If authoritarian governments use the signing - and violation - of human rights agreements as a signal of their willingness to repress domestic opponents, those regimes that practice repression *ex ante* are most likely to sign. Moreover, in equilibrium, those states that sign continue to torture after signing. The informational effect acts as a threat: signing the treaty signals strength and a willingness to torture if necessary. A rational opposition reduces its political activity in response. Since torture and opposition effort are strategic complements, torture levels in these states fall. So while the treaty is signed with an intent to defy its provisions, torture levels do fall relative to what they would have been absent signing. They do not, however, go to zero.

Torture falls in signatory states, but it rises in non-signatories relative to a world without the CAT. Signing the CAT prolongs the tenure in office of the worst torturers relative to the non-signers who are lesser torturers ex ante.

What then to make of the CAT? Firstly, while the CAT may reduce torture in the most autocratic of states, those states sign precisely because they intend to continue torturing. Torture levels do fall, but those regimes that sign become more secure. The good intentions of the international community may have the unintended consequence of strengthening undemocratic regimes around the world. Secondly, since torture falls in signatory states, but rises in non-signatories, the CAT appears to "redistribute" torture - away from the worst offenders to those lesser torturers who do not sign the CAT.

The aggregate welfare implications of the CAT remain to be established. It may be that the additional torture that occurs in non-signatory states is less than the reduction in the torture that occurs in signing states, in which case the CAT reduces torture in the aggregate. Similarly, the tenure in office of the worst torturers rises, but the tenure in office falls for

less oppressive autocrats; the overall tenure in office of autocrats may rise or fall will the introduction of the CAT. Finally, integrating the torture across time and across autocracies would yield insights as to the overall benefits, or lack thereof, of the CAT. These results remain to be established, and space constraints prohibit their exploration here, and remain a matter for further research.

Our model indicates that the CAT presents an opportunity for authoritarian regimes that value office very highly to signal their intent to hold firmly onto that office. The signal associated with signing the agreement is informative in the sense that signing acts as a credible threat of a willingness to exert effort and incur high costs in order to remain in power. While the treaty designers probably had no intention for the treaty to play this role, hard-core authoritarian regimes appear to have taken advantage of this mechanism. This suggests that these autocrats have an incentive to try to find other mechanisms to signal their type in a credible fashion. While such signaling devices may exist, few are credible in their capacity to separate out types. So long as some degree of uncertainty over governments' willingness to employ repression exists, low-cost types will have an incentive to signal their status. Thus, they will resort to a variety of signaling mechanisms, including - it seems - the signing of the CAT.

The results here offer a response to those that argue for a strengthening of the international human rights regime. A stronger regime means greater penalties for engaging in torture, raising the costs of compliance. Standard accounts suggest that the lower a country's cost of compliance, the greater its probability that it joins an international organization. In this model, the prediction is quite the opposite. As compliance costs rise, the pool of signatories becomes increasingly dominated by those that intend to defy the treaty. For the higher are the costs, the greater the separation of high- from low-cost types, and the more potent the signal sent to the domestic opposition. Higher compliance costs in this case may be more effective at protecting a high-torturing regime.

The findings of this research suggest an under-appreciated element of international institutional design. Agreements that focus on the nation state as a unitary actor, and ignore the effect of the institution on domestic politics - and in particular domestic conflict - may generate unanticipated and adverse effects. Policymakers, engaged in negotiations at the international level over the design of international institutions need to anticipate the effect of these agreements on the domestic polity. Agreements may come into effect exactly because they bolster the political survival of those leaders that sign them. When these leaders are autocratic, it is likely that they will use participation in international agreements to help prevent democratic reform.

How is possible that such a simple model yields such counterintuitive results? We believe this is a consequence of taking two aspects of international politics more seriously. First, domestic political contestation matters when it comes to a state's decision to accede or not to an international obligation. Second, international institutions generate information (if only by states' accession) that will affect the political calculus of the domestic groups engaged in political competition. By combining the information generated by the international institution and an explicit political contest at the domestic level, we generate results that depart somewhat from the standard canon: countries may accede to treaties they intend to defy.

References

Achen, Christopher H. & Duncan Snidal. 1989. "Rational Deterrence Theory and Comparative Case Studies." *World Politics* 41(2):143–169.

Bagwell, Kyle & Robert Staiger. 2005. "Enforcement, Private Political Pressure and the GATT/WTO Escape Clause." *The Journal of Legal Studies* 34(2):471–513.

Banks, Arthurs S. 1979. "Cross-National Time-Series Data Archive." Center for Social Analysis, State University of New York at Binghampton.

Box-Seffensmeier, Janet M. & Bradford S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. Cambridge University Press.

Bueno de Mesquita, Bruce, Alastair Smith, Randolph M. Siverson & James D. Morrow. 2003. *The Logic of Political Survival*. Cambridge, MA: The MIT Press.

Chayes, Abram & Antonia Handler Chayes. 1993. "On Compliance." *International Organization* 47(2):175–205.

Cingranelli, David L. & David L. Richards. 2007. "The Cingranelli-Richards (CIRI) Human Rights Dataset.".

Diamond, Alexis & Jasjeet S. Sekhon. 2006. "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies.".

Downs, George W. & David M. Rocke. 1995. *Optimal Imperfection?* Princeton University Press.

Downs, George W., David M. Rocke & Peter N. Barsoom. 1996. "Is the Good News About Compliance Good News About Cooperation?" *International Organization* 50(3):379–406.

Fearon, James D. 1994. "Domestic Political Audiences and the Escalation of International Disputes." *The American Political Science Review* 88.3:577–592.

Finnemore, Martha. 1996. *National Interests in International Society, Cornell Studies in Political Economy*. Cornell University Press.

Fudenberg, Drew & Jean Tirole. 1991. *Game Theory*. The MIT Press.

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

Gilligan, Michael J. & Nathaniel H. Nesbitt. 2007. "Do Norms Reduce Torture?".

Goemans, Hein. 2006. Archigos: A Database on Political Leaders. Paper Presented at the Annual Meeting of the American Political Science Association.

Goodliffe, Jay & Darren G. Hawkins. 2006. "Explaining Commitment: State and the Convention Against Torture." *Journal of Politics* 68:358–371.

Goodliffe, Jay, Darren Hawkins & James Raymond Vreeland. 2009. "Identity and Norm Diffusion in the Convention Against Torture.".

Hafner-Burton, Emilie M. 2005. "Trading Human Rights: How Preferential Trade Agreements Influence Government Repression." *International Organization* 59(3):593–629.

Hafner-Burton, Emilie M. & Kiyoteru Tsutsui. 2005. "Human Rights in a Globalizing World: The Paradox of Empty Promises." *American Journal of Sociology* 110(5):1373?1411.

Hafner-Burton, Emilie M. & Kiyoteru Tsutsui. 2007. "Justice Lost! The Failure of International Human Rights Law to Matter Where Its Needed Most." *Journal of Peace Research* 44(4):407–425.

Hathaway, Oona. 2007. "Why Do Countries Commit to Human Rights Treaties?" *The Journal of Conflict Resolution* 51(4):588–621.

Hirschleifer, Jack. 1991. "The Paradox of Power." *Economics and Politics* 3:177–200.

Ho, Daniel E., Kosuke Imai, Gary King & Elizabeth A. Stuart. 2007. "Matching as Prepossesing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* (forthcoming).

Ho, Daniel, Kosuke Imai, Gary King & Elizabeth Stuart. 2004. "MatchIt: Matching as Nonparametric Preprocessing for Parametric Causal Inference.". http://gking.harvard.edu/matchit/.

ICJ. 2009. Questions Relating to the Obligation to Prosecute or Extradite (Belgium v. Senegal). Summary of Order 2009/3 International Court of Justice.

Jonas, Stacie. 2004. "The Ripple Effect of the Pinochet Case." *Human Rights Brief* 11(3):36–38.

Keck, Margaret E. & Kathryn Sikkink. 1998. *Activists Beyond Borders: Advocacy Networks in International Politics*. Cornell University Press.

Klein, David F. 1988. "A Theory for the Application of the Customary International Law of Human Rights by Domestic Courts." *Yale Journal of International Law* 13:332–365.

Koremenos, Barbara. 2005. "Contracting Around International Uncertainty." *The American Political Science Review* 99:549–565.

Lacina, Bethany & Nils Petter Gleditsch. 2005. "Monitoring Trends in Global Combat: A New Dataset of Battle Deaths." *European Journal of Population* 21:145–166.

Moravcsik, Andrew. 2000. "The Origins of Human Rights Regimes: Democratic Delegation in Postwar Europe." *International Organization* 54(2):217–252.

Neumayer, Eric. 2005. "Do International Human Rights Treaties Improve Respect for Human Rights?" *The Journal of Conflict Resolution* 49(6):925–953.

Nielsen, Richard & Beth A. Simmons. 2009. "Rewards for Rights Ratification? Testing for Tangible and Intangible Benefits of Human Rights Treaty Ratification.".

Powell, Emilia Justyna & Jeffrey K. Stanton. 2009. "Domestic Judicial Institutions and Human Rights Treaty Violation." *International Studies Quarterly* 53(1):149–174.

Przeworski, Adam, Michael E. Alvarez, José Antonio Cheibub & Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990*. Cambridge University Press.

Roht-Arriaza, Naomi. 2001. "The Pinochet Precedent and Universal Jurisdiction." *New England Law Review* 35(2):311–320.

Rosendorff, B. Peter. 2005. "Stability and Rigidity: Politics and the Design of the WTO's Dispute Resolution Procedure." *The American Political Science Review* 99(3):389–400.

Rosendorff, B. Peter & Helen V. Milner. 2001. "The Optimal Design of International Trade Institutions: Uncertainty and Escape." *International Organization* 55(4):829–857.

Simmons, Beth A. Forthcoming, 2009. *Mobilizing for Human Rights: International Law in Domestic Politics*. Princeton University Press.

Simmons, Beth A. & Daniel J. Hopkins. 2005. "The Constraining Power of International Treaties: Theory and Methods." *The American Political Science Review* 99(4):623–631.

Singer, J. David. 1987. "Reconstructing the Correlates of War Dataset on Material Capabilities of State, 1816-1985." *International Interactions* 14:115–132.

Skaperdas, Stergios. 1996. "Contest Success Functions." *Economic Theory* 7:283–290.

Smith, Alastair. 1998. "International Crises and Domestic Conflicts." *The American Political Science Review* 92.3:623–638.

Vreeland, James Raymond. 2008. "Political Institutions and Human Rights: Why Dictatorships Enter into the United Nations Convention Against Torture." *International Organization* 62:65–101.

A Appendix: Proofs

A perfect Bayesian equilibrium of a signaling game consists of a strategy profile and a system of beliefs such that (1) the sender chooses her strategy to maximize her utility subject to the receiver's strategy; (2) the receiver chooses her strategy to maximize her utility subject both to the sender's strategy and to her beliefs conditional upon the sender's message; and (3) the receiver's beliefs are updated according to Bayes' Rule, whenever possible (Fudenberg &

Tirole 1991).

Definition: Define a pair of strategies $\{(s,t),e\}$ where $s:[0,1]\rightarrow\{0,1\},\ t:\{0,1\}\times[0,1]\rightarrow\mathbb{R}_{+},\ e:\{0,1\}\rightarrow\mathbb{R}_{+}$.

Definition: Define a function $\Psi(x)=\frac{1-x^{\frac{3}{2}}}{(1-x)\sqrt{x}}$. Note that since $\Psi(\cdot)$ is monotonic and decreasing for $x>0$, it is invertible. Denote the inverse of $\Psi(x)$ as $\Psi^{-1}(\cdot)$.

Proof of Proposition 1:

The proposition states that if $k>\frac{3}{2}$ and $b>\frac{2k}{3}\sqrt{\Psi^{-1}(k)}$ then there exists a unique semi-separating equilibrium where for $\widetilde{\theta}=\Psi^{-1}(k)$, where $s(\theta)=\begin{cases}1 & if \quad \theta<\widetilde{\theta}\\ 0 & if \quad \theta>\widetilde{\theta}\end{cases}$ and $e(s)=\begin{cases}\frac{Rk\tilde{\theta}}{9b^{2}} & if \quad s=1\\ \frac{R}{9b^{2}}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)^{2}}{\left(1-\tilde{\theta}\right)^{2}} & if \quad s=0\end{cases}$ and $t(s,\theta)=\begin{cases}\frac{R}{3b}\left[\sqrt{\frac{\tilde{\theta}}{\theta}}-\frac{k\tilde{\theta}}{3b}\right] & if \quad \theta<\widetilde{\theta}\\ \frac{R}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\tilde{\theta}\right)}\left(\sqrt{\frac{1}{\theta}}-\frac{1}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\tilde{\theta}\right)}\right) & if \quad \theta>\widetilde{\theta}\end{cases}$

We prove this first by checking, given any signal, that each player is playing a best response. Then we check that given this behavior, no type has an incentive to send another signal. Thirdly, we specify the conditions for the threshold type $\widetilde{\theta}$ to be interior to the type space. Finally we establish uniqueness of the semi-separating equilibrium.

Firstly, suppose $s=1$, $\frac{\partial U_G}{\partial t}=\pi_t(t,e)R-k\theta=0$ yields a reaction function $t(e,1,\theta)=\sqrt{\frac{eR}{k\theta}}-e\ \Rightarrow\ t(\frac{Rk\theta}{9b^2},1,\theta)=\frac{R}{3b}[\sqrt{\frac{\tilde{\theta}}{\theta}}-\frac{k\tilde{\theta}}{3b}]$. Therefore, autocratic signatory governments are playing a best response to their opposition when $t(1,\theta)=\frac{R}{3b}[\sqrt{\frac{\tilde{\theta}}{\theta}}-\frac{k\tilde{\theta}}{3b}]$.

Suppose $s=0$, $\frac{\partial U_G}{\partial t}=\pi_t(t,e)R-\theta=0$ yields a reaction function $t(e,0,\theta)=\sqrt{\frac{eR}{\theta}}-e\Rightarrow t(\frac{R(1-\tilde{\theta}^{\frac{3}{2}})^2}{9b^2(1-\tilde{\theta})^2},0,\theta)=\frac{R(1-\tilde{\theta}^{\frac{3}{2}})}{3b(1-\tilde{\theta})}(\sqrt{\frac{1}{\theta}}-\frac{1-\tilde{\theta}^{\frac{3}{2}}}{3b(1-\tilde{\theta})})$. Therefore, non-signatory autocratic governments are playing a best response to their opposition when $t(0,\theta)=\frac{R(1-\tilde{\theta}^{\frac{3}{2}})}{3b(1-\tilde{\theta})}(\sqrt{\frac{1}{\theta}}-\frac{1-\tilde{\theta}^{\frac{3}{2}}}{3b(1-\tilde{\theta})})$.

The opposition's problem, if the opposition observes $s=1$ is to maximize $EU_D=\int_0^{\tilde{\theta}}[(1-\pi(t(e,1),e))R-be]f(1)d\theta$, where $f(1)$ is the posterior distribution of θ over updated support $\left[0,\tilde{\theta}\right]$, conditional on signal $s=1$.

From the reaction function above, $\pi(t(e,1),e) = \frac{t(e,1,\theta)}{t(e,1,\theta)+e(1)} = 1 - \sqrt{ek\theta/R}$. Then:

$$\frac{\partial U_D}{\partial e} = \int_0^{\tilde{\theta}} [\frac{1}{2}\sqrt{\frac{k\theta R}{e}} - b] f(1) d\theta = 0$$
$$\Leftrightarrow e(1) = \frac{Rk\widetilde{\theta}}{9b^2}$$

Similarly, if the opposition observes $s = 0$, $EU_D = \int_{\tilde{\theta}}^1 [(1 - \pi(t(e,0),e))R - be] f(0) d\theta$, where $f(0)$ is the posterior distribution of θ over updated support $\left[\tilde{\theta}, 1\right]$, conditional on signal $s = 0$. From the reaction function above, $\pi(t(e,0),e) = 1 - \sqrt{e\theta/R}$. Then:

$$\frac{\partial U_D}{\partial t} = \int_{\tilde{\theta}}^1 [\frac{1}{2}\sqrt{\frac{R\theta}{e}} - b] f(0) d\theta$$
$$\Leftrightarrow e(0) = \frac{R(1 - \tilde{\theta}^{\frac{3}{2}})^2}{(9b^2)(1-\tilde{\theta})^2}$$

In both cases the opposition is playing a best response to the government's action in the equilibrium specified.

Secondly, we must check whether autocratic governments of type $\theta < \tilde{\theta}$ have an incentive to deviate by setting $s = 0$, or if governments of type $\theta > \tilde{\theta}$ have an incentive to deviate by setting $s = 1$.

Consider a type $\theta \in (0, \tilde{\theta})$. If $s = 1$, $U_G = \pi(t(1,\theta),e(1))R - k\theta t(1,\theta) = (1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta})R - \frac{Rk\theta}{3b}[\sqrt{\frac{\tilde{\theta}}{\theta}} - \frac{k\tilde{\theta}}{3b}] = R(1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta})^2$. If, on the other hand, $s = 0$, $U_G = \pi(t(0,\theta),e(0))R - \theta t(0,\theta) = R(1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})})^2$. Defection then occurs iff $\pi(t(1,\theta),e(1))R - k\theta t(1,\theta) < \pi(t(0,\theta),e(0))R - \theta t(0,\theta) \Rightarrow R(1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta})^2 < R(1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})})^2$. This inequality holds iff $k > \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$or iff $\theta > \Psi^{-1}(k) = \tilde{\theta}$. But $\theta < \tilde{\theta}$, so there is no incentive to defect.

Consider a type $\theta \in (\tilde{\theta}, 1)$. If $s = 0$, $U_G = \pi(t(0,\theta),e(0))R - \theta t(0,\theta) = R(1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})})^2$. If, on the other hand, $s = 1$, $U_G = \pi(t(1,\theta),e(1))R - k\theta t(1,\theta) = R(1 - \frac{k}{3b}\sqrt{\tilde{\theta}\theta})^2$. Defection takes place iff $\pi(t(0,\theta),e(0))R - \theta t(0,\theta) < \pi(t(1,\theta),e(1))R - k\theta t(1,\theta) \Rightarrow R(1 - \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})}^2) <$

$R(1-\frac{k}{3b}\sqrt{\tilde{\theta}\theta})^2$. This inequality holds iff $k < \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$ or iff $\theta < \Psi^{-1}(k) = \tilde{\theta}$. But $\theta > \tilde{\theta}$, so there is no incentive to defect.

Thirdly, for a unique interior $\tilde{\theta}$ to exist, we must show that $\Psi^{-1}(k) \in (0,1)$. Note that: $\Psi(\theta_+ \to 0) \to \infty$ and $\Psi(\theta_- \to 1) = 1.5$. Since $\Psi(\cdot)$ is monotonic, there is a unique threshold type $\tilde{\theta}$, iff $1.5 < k$. For non-negative levels of torture and effort in equilibrium, we require $1-\frac{k}{3b}\sqrt{\tilde{\theta}\theta} > 0$ for $\theta \in (0,\tilde{\theta})$ and $1-\frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})} > 0$ for $\theta \in (\tilde{\theta},1)$. Restating, we require $b \geq \frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3(1-\tilde{\theta})}$ for all $\theta \in (\tilde{\theta},1)$ and $b \geq \frac{k}{3}\sqrt{\tilde{\theta}\theta}$ for all $\theta \in (0,\tilde{\theta})$. Now, since $k = \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$, $\frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3(1-\tilde{\theta})} = \frac{k}{3}\sqrt{\tilde{\theta}\theta}$. That is the two constraints for interior torture levels are the same. That is we require $b \geq \frac{k}{3}\sqrt{\tilde{\theta}\theta}$ for all θ. Then $b \geq \frac{k}{3}\sqrt{\tilde{\theta}\theta}$ for all θ iff $b \geq \frac{k}{3}\sqrt{\tilde{\theta}} = \frac{k}{3}\sqrt{\Psi^{-1}(k)}$. By assumption, $b > \frac{2k}{3}\sqrt{\Psi^{-1}(k)} > \frac{k}{3}\sqrt{\Psi^{-1}(k)}$ which is the condition for the equilibrium. □

Proof of Lemma 1:

$t(e,1,\theta) = \sqrt{\frac{eR}{k\theta}} - e;\ \pi(t(e,1,\theta),e(1)) = 1-\sqrt{ek\theta/R} = 1-\sqrt{\frac{Rk\tilde{\theta}}{9b^2}k\theta/R} = 1-\frac{k}{3b}\sqrt{\tilde{\theta}\theta}.$

$t(e,0,\theta) = \sqrt{\frac{eR}{k\theta}} - e;\ \pi(t(e,0,\theta),e(0)) = 1-\sqrt{e\theta/R} = 1-\frac{(1-\tilde{\theta}^{\frac{3}{2}})\sqrt{\theta}}{3b(1-\tilde{\theta})}.$ □

Proof of Proposition 2:

$\pi(t(1,\underline{\theta}),e(1)) = 1-\frac{k}{3b}\sqrt{\tilde{\theta}\underline{\theta}}\ \forall\ \underline{\theta} \in (0,\tilde{\theta});\ \pi(t(0,\bar{\theta}),e(0)) = 1-\frac{\sqrt{\bar{\theta}}(1-\tilde{\theta}^{\frac{3}{2}})}{3b(1-\tilde{\theta})}\ \forall\ \bar{\theta} \in (\tilde{\theta},1)$. By contradiction: Consider $\underline{\theta} \in (0,\tilde{\theta})$ and $\bar{\theta} \in (\tilde{\theta},1)$, a signer and non-signer respectively. Assume the contrary: $\pi(t(1,\underline{\theta}),e(1)) < \pi(t(0,\bar{\theta}),e(0))$. Then $k\sqrt{\underline{\theta}} > \frac{1-\tilde{\theta}^{\frac{3}{2}}}{\sqrt{\tilde{\theta}}(1-\tilde{\theta})}\sqrt{\bar{\theta}} \Rightarrow k\sqrt{\underline{\theta}} > k\sqrt{\bar{\theta}} \Rightarrow \sqrt{\underline{\theta}} > \sqrt{\bar{\theta}}$ which contradicts our initial assumption that $\underline{\theta} < \bar{\theta}$. □

Proof of Proposition 3:

The government's utility from torture t is

$$\begin{aligned} EU_G(t|e,\theta) &= \frac{t}{t+e}R - \theta t \\ \Rightarrow \frac{\partial EU_G}{\partial t} &= \frac{e}{(t+e)^2}R - \theta = 0 \Leftrightarrow t(e,\theta) = \left(\frac{eR}{\theta}\right)^{\frac{1}{2}} - e \end{aligned}$$

Given the government's reaction function, the opposition's expected utility from effort e is:

$$\begin{aligned} EU_D &= \int_0^1 \left[\frac{e}{t(e,\theta)+e}R - be\right] d\theta = \int_0^1 \left[(eR\theta)^{\frac{1}{2}} - be\right] d\theta \\ \frac{\partial EU_D}{\partial e} &= \int_0^1 \left[\frac{1}{2}\left(\frac{R\theta}{e}\right)^{\frac{1}{2}} - b\right] d\theta = 0 \iff \frac{1}{2}\left(\frac{R}{e}\right)^{\frac{1}{2}} \int_0^1 \theta^{\frac{1}{2}} d\theta - b = 0 \\ &\Leftrightarrow \frac{1}{2}\left(\frac{R}{e}\right)^{\frac{1}{2}} \left[\frac{2}{3}\theta^{\frac{3}{2}} + c\right]_0^1 = b \iff \frac{1}{3}\left(\frac{R}{e}\right)^{\frac{1}{2}} = b \iff e = \frac{R}{9b^2} \end{aligned}$$

So effort is declining in b. The equilibrium torture level is $t(\theta) = \frac{R\left(3b-\theta^{\frac{1}{2}}\right)}{9b^2\theta^{\frac{1}{2}}} \geq 0$ for $\frac{1}{3} \leq b$ which is decreasing in θ. □

Proof of Proposition 4:

$$\begin{aligned} e < e(0) &\Leftrightarrow \frac{R}{9b^2} < \frac{R\left(1-\widetilde{\theta}^{\frac{3}{2}}\right)^2}{9b^2\left(1-\widetilde{\theta}\right)^2} \iff 1 < \frac{\left(1-\widetilde{\theta}^{\frac{3}{2}}\right)^2}{\left(1-\widetilde{\theta}\right)^2} \iff 1 < \frac{1-\widetilde{\theta}^{\frac{3}{2}}}{1-\widetilde{\theta}} \\ &\Leftrightarrow 1-\widetilde{\theta} < 1-\widetilde{\theta}^{\frac{3}{2}} \iff \widetilde{\theta}^{\frac{3}{2}} - \widetilde{\theta} = \widetilde{\theta}\left(\widetilde{\theta}^{\frac{1}{2}} - 1\right) < 0 \text{ for all } \tilde{\theta} \in (0,1) \end{aligned}$$

$$\begin{aligned} e(1) < e &\Leftrightarrow \frac{Rk\widetilde{\theta}}{9b^2} < \frac{R}{9b^2} \iff k\widetilde{\theta} < 1 \\ &\Leftrightarrow \widetilde{\theta} = \Psi^{-1}(k) < \frac{1}{k} \iff k > \Psi\left(\frac{1}{k}\right) = \frac{1-\left(\frac{1}{k}\right)^{\frac{3}{2}}}{\left(\frac{1}{k}\right)^{\frac{1}{2}} - \left(\frac{1}{k}\right)^{\frac{3}{2}}} = \frac{k^{\frac{3}{2}}-1}{k-1} \\ &\Leftrightarrow 0 < k^2 - k^{\frac{3}{2}} - k + 1 \text{ which holds for } k > \frac{3}{2} \end{aligned}$$

□

Proof of Proposition 5:

$t(1,\theta) \leq t(\theta) \Leftrightarrow \frac{R}{3b}\left(\frac{\sqrt{\tilde{\theta}}}{\sqrt{\theta}} - \frac{k\tilde{\theta}}{3b}\right) \leq \frac{R}{3b}\left(\frac{1}{\sqrt{\theta}} - \frac{1}{3b}\right) \Leftrightarrow b \geq \frac{1}{3}\left[\frac{(1-k\tilde{\theta})\sqrt{\theta}}{1-\sqrt{\tilde{\theta}}}\right]$. Now by assumption, $b > \frac{b}{2} > \frac{k}{3}\sqrt{\Psi^{-1}(k)}$. Then if $\frac{k}{3}\sqrt{\Psi^{-1}(k)} \geq \frac{1}{3}\left[\frac{(1-k\tilde{\theta})\sqrt{\theta}}{1-\sqrt{\tilde{\theta}}}\right]$ we are done. Now $\frac{k}{3}\sqrt{\Psi^{-1}(k)} = \frac{k}{3}\sqrt{\tilde{\theta}}\sqrt{\theta}$. Then

$$\begin{aligned} t(1,\theta) \leq t(\theta) &\Leftarrow \frac{k}{3}\sqrt{\tilde{\theta}}\sqrt{\theta} \geq \frac{1}{3}\left[\frac{(1-k\tilde{\theta})\sqrt{\theta}}{1-\sqrt{\tilde{\theta}}}\right] \\ &\Leftarrow k \geq \frac{1}{\tilde{\theta}} - \frac{1-\tilde{\theta}}{\sqrt{\tilde{\theta}}} \end{aligned}$$

Recall that $k = \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}$. Then

$$\begin{aligned} t(1,\theta) \leq t(\theta) &\Leftarrow \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}} \geq \frac{1}{\tilde{\theta}} - \frac{1-\tilde{\theta}}{\sqrt{\tilde{\theta}}} \\ &\Leftarrow 0 \geq \tilde{\theta}^{\frac{1}{2}} - \tilde{\theta}^2 - 2 \end{aligned}$$

which is true for all $\tilde{\theta} \in (0,1)$.

$t(0,\theta) \geq t(\theta) \Leftrightarrow \frac{R}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\tilde{\theta}\right)}\left(\sqrt{\frac{1}{\theta}} - \frac{1}{3b}\frac{\left(1-\tilde{\theta}^{\frac{3}{2}}\right)}{\left(1-\tilde{\theta}\right)}\right) \geq \frac{R}{3b}\left(\frac{1}{\sqrt{\theta}} - \frac{1}{3b}\right) \Leftrightarrow b \geq \frac{2-\tilde{\theta}^{\frac{3}{2}}-\tilde{\theta}}{3(1-\tilde{\theta})}\sqrt{\theta}$. As in the previous case, if we can prove $\frac{2k}{3}\sqrt{\tilde{\theta}}\sqrt{\theta} \geq \frac{2-\tilde{\theta}^{\frac{3}{2}}-\tilde{\theta}}{3(1-\tilde{\theta})}\sqrt{\theta}$ then we are done. Then

$$\begin{aligned} t(0,\theta) \geq t(\theta) &\Leftarrow \frac{2k}{3}\sqrt{\tilde{\theta}}\sqrt{\theta} \geq \frac{2-\tilde{\theta}^{\frac{3}{2}}-\tilde{\theta}}{3(1-\tilde{\theta})}\sqrt{\theta} \\ &\Leftarrow \frac{1-\tilde{\theta}^{\frac{3}{2}}}{(1-\tilde{\theta})\sqrt{\tilde{\theta}}}\frac{2\sqrt{\tilde{\theta}}}{3} \geq \frac{2-\tilde{\theta}^{\frac{3}{2}}-\tilde{\theta}}{3(1-\tilde{\theta})} \\ &\Leftarrow 2(1-\tilde{\theta}^{\frac{3}{2}}) \geq 2-\tilde{\theta}^{\frac{3}{2}}-\tilde{\theta} \\ &\Leftarrow \tilde{\theta}\left(1-\tilde{\theta}^{\frac{1}{2}}\right) \geq 0 \end{aligned}$$

which is true for all $\tilde{\theta} \in (0,1)$. □

B Hollyer and Rosendorff: Model Robustness

Not for publication, but for posting online.

B.1 Post-Tenure Punishments

Assume that signing the CAT makes an autocratic leader vulnerable to post-tenure liability. That is, signing the CAT renders the leader open to punishment in the event that he is removed to office. Further assume that this additional punishment is a constant p. (In equilibrium, all autocratic governments repress, at least to some extent. The assumption that the punishment is a constant allows us to incorporate the possibility of post-tenure punishments without burdening the model with undue mathematical complexity.)
Denote the Government's (G's) utility as:

$$U_G(t, e, s; \theta) = \pi(t, e)[R + sp] - \theta t - sp$$

where θ denotes the marginal cost of repression, $s \in \{0, 1\}$ takes the value of 1 in the event G signs the CAT, and $\pi(t, e)$ is the contest success function $\pi(t, e) = \frac{t}{t+e}$. D's utility function is identical to that in the main model:

$$U_D(t, e, s) = [1 - \pi(t, e)]R - be$$

Proposition 1. *If $p > 0$, there exists semi-separating equilibrium where all governments with low costs to repression $\theta \leq \widehat{\theta}$ sign the treaty, and all governments with high costs $\theta > \widehat{\theta}$ do not sign for some $\widehat{\theta} \in (0, 1)$.*

Proof: Firstly, the best responses are $t(e, 1; \theta) = \sqrt{\frac{e(R+p)}{\theta}} - e$ and $t(e, 0; \theta) = \sqrt{\frac{eR}{\theta}} - e$. The opposition's problem given these best responses are for $s = 1$ to maximize $\int_0^{\widehat{\theta}} (R\sqrt{\frac{e\theta}{R+p}} - be)f(1)d\theta \Leftrightarrow e(s = 1) = \frac{R^2\widehat{\theta}}{9b^2(R+p)}$ and for $s = 0$: $\frac{\partial U_D}{\partial t} = \int_{\widehat{\theta}}^1 [\frac{1}{2}\sqrt{\frac{R\theta}{e}} - b]f(0)d\theta \Leftrightarrow e(0) =$

$\frac{R(1-\widehat{\theta}^{\frac{3}{2}})^2}{(9b^2)(1-\widehat{\theta})^2}$. From the response functions of the government and opposition, we can derive the contest success function when $s = 1$, $\pi(t, e; s = 1) = 1 - \frac{R\sqrt{\widehat{\theta}\theta}}{3b(R+p)}$. And when $s = 0$, $\pi(t, e; s = 1) = 1 - (\frac{\sqrt{\theta}}{3b})[\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}]$. Governments will prefer to sign when $U_G(t, e, s = 1; \theta) \geq U_G(t, e, s = 0; \theta)$, when $(R + p)\left[1 - \frac{R\sqrt{\widehat{\theta}\theta}}{3b(R+p)}\right] - \theta\left(\frac{R}{3b}\sqrt{\frac{\widehat{\theta}}{\theta}} - \frac{R^2\widehat{\theta}}{9b^2(R+p)}\right) - p \geq$
$R\left[1 - \frac{\sqrt{\theta}}{3b}\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right)\right] - \theta\left[\frac{R}{3b}\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right)\sqrt{\frac{1}{\theta}} - \frac{R}{9b^2}\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right)\right]$,
when $\frac{6b}{\sqrt{\theta}}\left[R\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right) - \sqrt{\widehat{\theta}}\right] \geq \left[R\left(\frac{1-\widehat{\theta}^{\frac{3}{2}}}{1-\widehat{\theta}}\right)\right]^2 - \frac{1}{(R+p)}\widehat{\theta}$
The LHS is monotonic and decreasing in θ over the unit interval, and the right hand side is constant, so if $\widehat{\theta}$ is interior, it is the low cost (low θ) types that sign. To show $\widehat{\theta}$ is interior, it is necessary to show that there exists a value of $\theta \in (0, 1)$ such that the above expression holds at equality when $\theta = \widehat{\theta}$. A fully analytic solution is elusive, but an interior solutions exist for a variety of parameterizations in simulations. For instance, at $(R, p, b) = \left(1, 100, \frac{1}{3}\right)$, $\widehat{\theta} = 0.497 \in (0, 1)$.

Proposition 2. *As the level of post-tenure punishments grows increasingly large, signatory governments are willing to devote more effort to repression and the opposition devotes less effort to removing the government on witnessing treaty accession. Survival times of signatory governments will thus increase with p.*

Proof: From the response functions above, it follows that $t(e(1); \theta) = \frac{R\sqrt{\tilde{\theta}}}{3b\sqrt{\theta}} - \frac{R^2\tilde{\theta}}{9b^2(R+p)}$ and $e(s = 1) = \frac{R^2\tilde{\theta}}{9b^2(R+p)}$. Clearly the former is increasing and the latter decreasing in p. Moreover, survival is given by $\pi(t, e; s = 1) = 1 - \frac{R\sqrt{\tilde{\theta}\theta}}{3b(R+p)}$ which is increasing in p.

B.2 Varying Values to Holding Office

Assume that governments vary in the value the place on retaining office, rather than in the costs they suffer from engaging in repression. Thus, the signing of the CAT will act as a signal that a government places a high value on office, rather than a signal of the government's low marginal cost to repression. Let θ be constant, and common knowledge. Denote the value the

government attaches to office as $R_G \sim U[0,1]$. The government observes the realization of this variable. The opposition is only aware of the distribution from which it is drawn. Denote the value the opposition attaches to office as R_D. This value is a constant and is common knowledge. As in the above, allow for post-tenure punishments of value p. The government's utility function is thus given by $U_G(t,e,s;R_G) = \pi(t,e)[R_G + sp] - [sk + (1-s)]\theta t - sp$, while the opposition's is given by $U_D(t,e,s) = [1-\pi(t,e)]R_D - be$.

Proposition 3. *There exists $\tilde{R_G} \in (0,1)$ such that for $k < \frac{1-\tilde{R_G}}{\left(\sqrt{1+p}-\sqrt{\tilde{R_G}+p}\right)\sqrt{\tilde{R_G}}}$, a semi-separating equilibrium exists where all high-value governments $R_G \geq \tilde{R_G}$ sign the treaty; while all low-value governments $R_G < \tilde{R_G}$ do not.*

Proof: For $s=1$, the government's response function is $t(e(1);R_G) = \sqrt{\frac{e(1)(R_G+p)}{k\theta}} - e(1)$; for $s=0$, the government's response function is $t(e(0);R_G) = \sqrt{\frac{e(0)R_G}{\theta}} - e(0)$. The opposition's problem when $s=1$ is given by $EU_D = \int_{\tilde{R_G}}^{1}[R_D\sqrt{\frac{e(1)k\theta}{R_G+p}} - be]f(1)dR_G$ where $f(1)$ represents the posterior distribution over $[\tilde{R_G},1]$ given that $s=1$. Solving for this expression yields:

$$EU_D = R_D\sqrt{e(1)k\theta}\int_{\tilde{R_G}}^{1}\frac{1}{\sqrt{R_G+p}}f(1)dR_G - be(1)$$
$$\Leftrightarrow e(1) = \frac{R_D^2(\sqrt{1+p}-\sqrt{\tilde{R_G}+p})^2k\theta}{(1-\tilde{R_G})^2b^2}$$

Similarly, when $s=0$, the opposition's problem is given by $EU_D = \int_0^{\tilde{R_G}}[R_D\sqrt{\frac{e\theta}{R_G}} - be]f(0)dR_G$ where $f(0)$ represents the posterior distribution over $[0,\tilde{R_G}]$ given that $s=0$. Solving this expression yields:

$$EU_D = R_D\sqrt{e(0)\theta}\int_{0}^{\tilde{R_G}}\frac{1}{\sqrt{R_G}}f(0)dR_G - be(0)$$
$$\Leftrightarrow e(0) = \frac{R_D^2\theta}{b^2\tilde{R_G}}$$

A government will be willing to sign the treaty when $U_G(t(e(1);R_G),e(1);R_G) \geq U_G(t(e(0);R_G),e(0);R_G)$. This inequality then yields:

$$\left(\sqrt{R_G} - \frac{R_D}{b}k\theta\frac{1}{1-\tilde{R_G}}\left(\sqrt{1+p} - \sqrt{\tilde{R_G}+p}\right)\right)^2 - \left(\sqrt{R_G} - \frac{R_D\theta}{b}\sqrt{\frac{1}{\tilde{R_G}}}\right)^2 \geq \left[2\frac{R_D}{b}k\theta\frac{1}{1-\tilde{R_G}}\left(\sqrt{1+p} - \sqrt{\tilde{R_G}+p}\right)\right]\left[\sqrt{R_G+p} - \sqrt{R_G}\right]$$

Note that the RHS is monotonic and decreasing in R_G. The LHS will be monotonic and increasing in R_G if:

$$\frac{R_D\left(\sqrt{1+p} - \sqrt{\tilde{R_G}+p}\right)k\theta}{(1-\tilde{R_G})b} < \frac{R_D\theta}{b\sqrt{\tilde{R_G}}} \Leftrightarrow k < \frac{1-\tilde{R_G}}{\left(\sqrt{1+p} - \sqrt{\tilde{R_G}+p}\right)\sqrt{\tilde{R_G}}}$$

Note that if $\tilde{R_G}$ is interior, this implies that it will the high-value types that sign the treaty. Again a fully analytic solution is elusive, but an interior solutions exist for a variety of parameterizations in simulations. For instance, at $(p,k,\theta,b,R_D) = \left(50,\frac{1}{2},\frac{1}{2},\frac{1}{2},1\right)$, $\tilde{R_G} = 0.66576 \in (0,1)$.