On the perils of commitment to punishment when criminals are strategic[*]

Shaun Larcom[†] Mare Sarr[‡]

May 31, 2017

Abstract

For some crimes the perpetrator can be detected costlessly but can only be apprehended at significant cost, or not at all (for some period of time). In an effort to deter strategic behavior in the period between detection and apprehension, authorities may wish to commit themselves to punishing the criminal once apprehended, regardless of the perpetrator's behavior or threats. However we show that such efforts at commitment to ex post punishment may induce worse behavior and that it selects potential criminals of a worse type. We show that when law enforcement authorities cannot commit themselves perfectly, it is dangerous for them to try to commit as it may invoke a strategic response that can worsen the situation. When law enforcement authorities do increase their commitment to punish such offenders, it is likely to lead to less but more gruesome crimes.

JEL-classification: F55, K14, O12

Key words: Marginal Deterrence, Commitment, Kidnapping, International Criminal Court, Amnesty, Impunity, Signaling.

[*] We are indebted to Tim Willems for all his assistance in the production of this manuscript and constructive discussions on earlier versions of the paper. We thank Haim Abraham, Alp Atakan, Francesco Caselli, Alex de Cornière, Avinash Dixit, Mikhail Drugov, Georgy Egorov, Peter Eso, Leonardo Felli, Ian Jewitt, Jean-Paul L'Huillier, Antonio Mele, Meg Meyer, Ines Moreno de Barreda, Ricardo Nunes, Torsten Persson, Rick van der Ploeg, Eric Posner, Kevin Sheedy, Ragnar Torvik, Sweder van Wijnbergen, and audiences at the ESSET Conference in Gerzensee, the IEA World Congress in Jordan, the SIRE Political Economy Workshop in Glasgow, CERGE-EI in Prague, and the Universities of Bristol, Oxford, St Andrews, and Pretoria for useful comments. We also thank Mark Freeman, Markus Gehring, and Roger O'Keefe for their comments relating to International Criminal Law.

[†] Department of Land Economy, University of Cambridge E-mail: stl25@cam.ac.uk.

[‡] School of Economics, University of Cape Town, Private Bag, Rondebosch 7701, South Africa. Email: mare.sarr@uct.ac.za

1 Introduction

There is a class of crime where standard deterrence models do not fit well.[1] For some crimes, the perpetrator can be costlessly detected but can only be apprehended at significant cost, or not at all (for some period of time). Furthermore, during the period between detection and apprehension the perpetrator may be able to inflict significant harm. This class of crime includes kidnapping, ransom demands, and war crimes committed by despots and warlords. Indeed, it applies to any circumstance where the wrongdoer is able to inflict significant harm on a regulator in the period between detection and apprehension.

The concept of marginal deterrence highlights the need to ensure that perpetrators continue to face incremental punishments once they have begun a criminal activity such as kidnapping (see Stigler 1970, Friedman and Sjostrom 1993, Mookherjee and Png 1994, Shavell 1992, and Detotto, et al. 2014 and 2015).[2] However, sometimes, given the perpetrator's ability to inflict such a high degree of harm during the period in-between detection and apprehension, authorities may be willing to give in (in terms of allowing an escape or offering an amnesty or asylum) for an end to the harmful behavior (or threats of harmful behavior). In such cases, authorities are faced with a fundamental trade-off between deterring such activity ex ante and ending the harmful behavior as painlessly and quickly as possible ex post.[3] One seemingly sensible solution, and consistent with the broad policy prescriptions of Kydland and Prescott (1997), is to commit to punishing such behavior, regardless of the ability of the perpetrator to inflict harm in the period between detection and apprehension. However, we show that in such circumstances, commitments to punishing perpetrators may actually generate worse criminal activity and a worse type of perpetrator committing them. This centers on the idea that despite an authority's best intentions to commit to punishing a perpetrator, they might ultimately find themselves unable as the costs of doing so are simply too high. Knowing this, and acting strategically, perpetrators may commit more and worse crimes in an effort to break the resolve to punish them. In doing so, we build on the work of Schwarz and Sonin (2008) by incorporating the concept of brinkmanship into law enforcement in a two period game.

Our model is motivated by the observation of actual events. Consider kidnapping. Although kidnappers typically threaten to kill the victim unless a ransom is paid, this is

[1] See Polinsky and Shavell (2007) for an overview of the public enforcement literature.

[2] See Vannini, et al. (2015) for an overview of this literature.

[3] There is an extensive literature exploring the trade-off between justice and peace. See see Sutter (1995), Scharf (1999), Osiel (2000), Goldsmith (2003), Snyder and Vinjamuri (2006), and Escribà-Folch (2013), Escribà-Folch and Wright, 2015), Ritter and Wolford (2012).

not always their true intention.[4] Consequently, authorities face an inference problem on the type of the kidnappers, while kidnappers have an incentive to signal that they are of the worst possible type (as that maximizes the probability that their demands will be met). One of the most fitting examples is that of John Paul Getty III.[5] This grandson of an oil tycoon was kidnapped in 1973 in Italy. Initially, Getty's grandfather refused to pay the ransom, referring to the risk of moral hazard: "I have 14 other grandchildren and if I pay one penny now, then I'll have 14 kidnapped grandchildren." The kidnappers recognized the problem and broke the stalemate by sending a package containing Getty's ear to a newspaper. The accompanying note read: "This is Paul's first ear. If within ten days the family still believes that this is a joke (...) then the other ear will arrive. In other words, he will arrive in little bits." After this, Getty Sr. quickly changed his mind (realizing that being time-inconsistent was probably the best option left) and paid $2.9 million for the return of his grandson. Following a wave of kidnappings in the 1970s and 1980s Italy implemented a law in 1991 that froze the family assets of hostages (increasing commitment to not paying a ransom). While the number of kidnappings went down those that did take place were more gruesome and lasted longer (Detotto et al., 2015). For example, in 1997 Italian authorities only gave in, and allowed the family to pay a ransom, after the kidnapped Giuseppe Soffiantini had both of his ears cut off. Because of the new law, receiving the first ear was apparently no longer enough to break the ex-ante commitment to not paying anything.

Other examples relate to the administration of international justice. Following numerous amnesty and asylum deals in the post World War II period, the permanent International Criminal Court (ICC) was established in 2002 to try persons for acts of genocide, war crimes, and crimes against humanity.[6] Countries party to the treaty are obliged to cooperate with the ICC and are in principle no longer allowed to grant amnesty or asylum to ICC-indicted individuals. In essence, signatories have committed themselves to punishing perpetrators. However, in this new institutional environment some perpetrators have tested the resolve of authorities. For instance, soon after it emerged that the Congolese government was going to enforce the ICC indictment of Bosco Ntaganda (an infamous warlord known as 'the Terminator'), his rebel group attacked the city of

[4] Although those subject to demands often avoid taking any risk and pay a ransom, still about 18 percent of all victims are released alive by the kidnappers without any form of payment. In only 6 percent of all cases, the victim is actually killed (Australian Senate, 2011: 16).

[5] See http://www.washingtonpost.com/wp-dyn/content/article/2011/02/07/AR2011020706089.html

[6] The practice of "trading justice for peace" was once encouraged by the United Nations as well (Scharf, 1999). Perhaps the most famous example is Idi Amin, one of the world's most infamous dictators whose regime is believed to have led to about 300,000 deaths in Uganda, spent his post-dictatorship years comfortably in a Saudi Arabian hotel - dying there in 2003, without being held to account for his crimes.

Goma. According to a spokesman, they were not interested in control of the city, but they wanted "the Congolese government [to] sit down at the negotiating table".[7] While such a tactic may seem cavalier, and did ultimately fail, it is not without logic. While the Rome Statute (that provides the legal framework underpinning the ICC) generally binds states from offering amnesties or asylum, it contains explicit exemptions, opening up the possibility for negotiation.[8] Analogous forces seem to have been at work in Zimbabwe, were members of Robert Mugabe's government have been "trying to force the political opposition into granting them amnesty for their past crimes by abducting, detaining and torturing opposition officials and activists"[9] - a striking statement which, as we will see later on, is exactly in line with our model's prediction.

Our analysis is built on the premise that although an authority's loss associated with giving in to a perpetrator increases with greater efforts at commitment (whether this be the enactment of new binding laws against negotiating with kidnappers or the installation of the ICC), it continues to be finite. Consequently, there still exists a critical level of wrongdoing beyond which the authority chooses to take the loss associated with giving in, rather than the pay-off resulting from sticking to its earlier threats to punish the wrongdoer. We show that wrongdoers anticipating this may choose to commit more crimes than they would actually like to from a static perspective, purely to worsen the situation and "unlock" the impunity option. In doing so, they are essentially forcing the authority to re-optimize. By behaving worse, wrongdoers can thus make their effective punishment function non-monotonic. As a result of this effect, increasing the costs of giving in may induce some criminals to commit more and more gruesome crimes, as that becomes necessary to make the amnesty-option available to them. Importantly, it must be stressed that this channel exists over and above any adverse consequences from a wrongdoer digging in for survival if he or she believes that his exit route is blocked.

We proceed by first introducing our model in Section II. Its implications are analyzed in Section III and used to derive policy prescriptions. Finally, Section IV concludes.

[7] See http://www.bbc.co.uk/news/world-africa-18821962.

[8] Specifically, see Articles 16, 17 and 53.

[9] See "Mugabe Aides Said to Use Violence to Get Amnesty" in The New York Times of April 9, 2009 (http://www.nytimes.com/2009/04/10/world/africa/10zimbabwe.html).

2 The Model

This section presents our model. We start by describing the environment and solving the model under perfect commitment (by which we gain understanding of our model), after which we turn to a more realistic setup in which commitment is imperfect. The timing in our model is as follows:

1. At the beginning of period 1, all potential criminals learn their type θ and decide whether they want to attempt to commit a crime.

2. Given his or her type θ, a criminal commits a crime x_1 during period 1.

3. At the beginning of period 2, the law enforcement authority forms an expectation of his or her type, based upon first-period behavior x_1. It subsequently decides whether to bargain with the criminal or not.

4. If a bargain is offered and accepted, $x_2 = 0$ and criminal obtains the value of the bargain (e.g. free passage to a safe haven). If no deal is made, the criminal commits further crimes x_2^* in period 2.

5. If no deal is made at Stage 4 the criminal is punished for any crimes that have been committed in the past.

As the model is most easily explained backwards, we end by describing the entry decision for potential criminals at the start of the game.

2.1 Model Environment and Solution under Perfect Commitment

Our model consists of two periods, $t = 1, 2$, and its players are the law enforcement authority (henceforth 'the authority') and a pool of potential criminals. The goal of the authority is to minimize the total amount of crime that a criminal will commit over both periods. One could also incorporate a time-discount factor if one considers that appropriate. Using x_t to denote the intensity of crime committed by a criminal in period t , the authority aims to minimize the expected value of:

$$\mathcal{L}^A \equiv x_1 + x_2 \tag{1}$$

It does so by threatening to punish criminals at the end of period 2 for any crimes that they have committed during periods 1 and 2. Importantly our model refers to that class of crimes where a perpetrator can be costlessly detected but that it takes time to apprehend him or her (if at all). For instance, a kidnapper in hiding or dictator or warlord with a standing army.[10] Consequently, the criminal has an intermediate phase available during which he is able to play a game with the authority. The length of a period is indeterminate, so readers may think of the model's time horizon as they consider appropriate.

Suppose, initially, that the authority can credibly commit to ex-post punishment and that it adopts a punishment function given by $h(x_1 + x_2) = \phi\,[x_1 + x_2]$, where the value of $\phi > 0$ is known to all agents in the model. This parameter combines the actual punishment conditional on the criminal being tried and convicted, with the probability that a criminal will be caught and brought to court - so ϕ should be seen as the expected punishment parameter.

We assume that it is not feasible to set ϕ prohibitively high due to notions of proportionality or laws against torture. We assume that the criminal derives benefits from committing the crime x_t. These take the form of monetary rewards.

Engaging in criminal activity is costly however, both in terms of opportunity cost and direct costs. We assume that the criminal's cost function of committing x_t crimes is given by $c(x_t) = \gamma x_t$. The value of $\gamma > 0$ is known to all agents in the model.

The criminal's objective is to maximize his or her own lifetime utility, which - after assuming log-utility for concreteness - is given by:

$$\max_{x_1,x_2}\ \theta \sum_{t=1}^{2} \log\left(x_t\right) - \gamma \sum_{t=1}^{2} x_t - \phi \sum_{t=1}^{2} x_t \tag{2}$$

Here, $\theta \geq 0$ is a time-invariant parameter capturing the criminal's type: a criminal with a high θ suffers no psychological penalty from committing crime relative to punishment, while a low θ implies high psychological costs from committing crime. In this sense, one could interpret θ as the intrinsic malevolence of the criminal. A crucial part of the model is that the value of θ is private information to the criminal. The authority can only form beliefs about it by observing the criminal's behavior in period 1. We assume that θ is distributed according to some known cumulative distribution function $F\left(\theta\right)$ in the pool of criminals and associated probability density function $f(\theta)$.

From solving the optimization problem (2), we can see that the first-order condition

[10] It for example took two wars and many sanctions - spanning a period of about 13 years - to remove Saddam Hussein from power in Iraq. There are also cases (like Fidel Castro in Cuba and Robert Mugabe in Zimbabwe) where international attempts have been unsuccessful in bringing about regime change. We will return to this assumption at the end of Section IV.

(henceforth referred to as "FOC") implies that:

$$x_t^* = \frac{\theta}{\gamma + \phi} \quad \text{for } t = 1, 2 \tag{3}$$

So at the beginning of period 2 (after observing x_1) the authority is able to make an inference about the criminal's type θ by using (3), leading to:

$$\mathbb{E}_2^A \{\theta | x_1\} = [\gamma + \phi] \, x_1, \tag{4}$$

where γ, ϕ, and x_1 are all known at this stage.[11]

2.2 Partial Commitment

So far, we have maintained the strong assumption that the authority is ex ante able to commit itself to punishing the criminal ex post according to some pre-specified punishment function $h(\cdot)$. In reality, however, such a perfect form of commitment is unlikely to be possible. Although the authority is able to increase the costs of allowing impunity (for example by creating laws or signing treaties that bind itself), it is unlikely to be able to increase these costs all the way up to infinity (which is necessary for perfect commitment). Indeed, in both examples in the introduction (Italian kidnapping laws and the Rome Statute) there are escape clauses for exceptional circumstances.

Therefore, let us suppose that the authority incurs a loss equal to a *finite* $\Lambda \geq 0$ if it ignores the punishment function $h(\cdot)$ at the beginning of period 2 and allows impunity (or only a minor punishment), in return for the criminal ceasing to inflict harm.[12] Since capture only takes place at the *end* of period 2, our model captures the idea a deal is able end the harm earlier.

Λ can encompass many different costs: some are institution-related (such as the loss of credibility for the authority or loss of prestige for law enforcement officials), while others would occur in any environment (such as the moral hazard for future criminals). We will assume that:

Assumption 1 *Λ is an increasing function of ϕ, i.e. $\Lambda = \Lambda(\phi)$ with $d\Lambda(\phi)/d\phi > 0$*

[11] We take the criminal's cost parameter γ to be public knowledge, but one could also assume that this information is private to the criminal (γ then takes over the role that θ plays in the current setup).

[12] Alternatively, one can interpret this as the law enforcement authority adhering to a rule that has an explicit escape clause. As shown by Lohmann (1992), the outcome under such a rule typically dominates that obtained under full commitment, so in that sense such a policy can also be optimal. Interestingly, the Rome Statute of the ICC contains explicit escape clauses - see especially its Articles 16, 17 and 53.

and $\Lambda(0) \geq 0$.

Our assumption that $d\Lambda(\phi)/d\phi > 0$ captures the intuitive notion that if the authority announces ex ante that it is going to prosecute these type of criminals with higher probability (thereby increasing expected punishment ϕ), then the (reputational) loss associated with subsequently *not* doing so will be higher.[13]

The solution-concept is Perfect Bayesian Equilibrium. Denoting the authority's action at the beginning of period 2 by $s \in \{deal, punish\}$, the equilibrium can be defined as:

Definition *A pure strategy Perfect Bayesian Equilibrium is a strategy profile* (x_t^{eq}, s^{eq}) *and a system of beliefs* $\mu(\theta|x_1)$ *such that:*

1. *After observing* x_1*, and given beliefs* $\mu(\theta|x_1)$*, the authority chooses an action* $s^{eq}(x_1)$ *that minimizes its expected loss function* $\mathbb{E}_2^A\{\mathcal{L}|x_1\}$.

2. *The beliefs* $\mu(\theta|x_1)$ *held by the authority about the criminal's type* θ *are obtained through Bayes' rule, where possible.*

3. *Taking the authority's best response* $s^{eq}(x_1)$ *into account, a criminal of type* θ *decides to commit crimes* $x_t^{eq}(\theta)$ *that maximize his utility.*

We denote the value of the impunity offer to the criminal by V^I. It consists of the value that the criminal attaches to not being tried (or receiving only a minor punishment), as well as of any potential additional benefits. In order to induce the criminal to accept the offer, the authority will have to set $V^I > \theta \log(x_2^*) - \gamma x_2^* - \phi[x_1 + x_2^*]$. The RHS of this inequality represents the value of the criminal's outside option. Note that the exact location of this "threat point" is private information to the criminal, because θ is private information. Due to the resulting informational advantage of the criminal over the authority, the criminal is able to exploit this edge when negotiating on V^I (see e.g. Sobel and Takahashi (1983) on how a party that is better informed is also able to achieve a better bargaining outcome). In particular, we assume that the criminal is able to capture

[13] As we show in Appendix B, the results that are to follow are robust to using the more general function $\Lambda(\phi, \phi x_1)$ with $\partial\Lambda(\phi, \phi x_1)/\partial\phi > 0$ and $\partial\Lambda(\phi, \phi x_1)/\partial(\phi x_1) > 0$. The second argument ϕx_1, which is punishment for first-period crimes according to the pre-specified punishment function ("punishment according to the law"), can now be thought of as representing a loss resulting from "not doing justice". Assuming that $\partial\Lambda(\phi, \phi x_1)/\partial(\phi x_1) > 0$ makes it costlier to allow impunity to those who acted worse in period 1. This allows for the idea that there might be more public outcry if amnesty is granted to a criminal who behaved very badly in period 1 and should have received a large punishment according to the strict letter of the law (i.e.: a criminal with high ϕx_1).

rents $\Delta > 0$ in the deal negotiations, such as a nice place to spend his post-criminal years. Consequently, V^I takes the following "outside-option plus surplus"-form:

$$V^I(\theta, \phi) = \theta \log(x_2^*) - \gamma x_2^* - \phi [x_1 + x_2^*] + \Delta \tag{5}$$

Because the authority's promise of impunity post-deal might not be fully credible either, V^I can be seen as an *expected* value to the criminal (and note that the model's solution is fully determined by expected values where applicable). In equation (5), the size of the bargaining-surplus Δ may depend upon various factors, such as the criminal's negotiation skills relative to those of the authority and the urgency of the situation. We leave these components un-modeled and furthermore set $\partial\Delta/\partial\mathbb{E}_2^A\{\theta|x_1\} = 0$. This implies that any strategic behavior that a criminal is going to display, is not going to arise with the aim of increasing negotiation-rents Δ - only with the aim of unlocking the impunity deal in the first place.

After observing x_1, the authority's expected loss function becomes:

$$\mathbb{E}_2^A\{\mathcal{L}|x_1\} = \begin{cases} x_1 + \Lambda(\phi) & \text{if an impunity deal is made} \\ x_1 + \mathbb{E}_2^A\{x_2|x_1\} & \text{otherwise} \end{cases} \tag{1'}$$

This modified loss function expresses the idea that if the criminal escapes via an impunity deal, x_2 is expected to collapse to zero. But by making such a deal, the authority simultaneously incurs a loss of $\Lambda(\phi)$, due to a loss of credibility, or due to moral hazard for future criminals.

Remembering that the authority is only able to infer the incumbent criminal's type from his first-period behavior, its beginning of period 2 expectation of the crimes that the criminal will commit during period 2 is subsequently given by:

$$\mathbb{E}_2^A\{x_2|x_1\} = \frac{\mathbb{E}_2^A\{\theta|x_1\}}{\gamma + \phi} \tag{6}$$

One can then see from the modified loss function (1') that the impunity deal will become available to the criminal as soon as the authority believes that:

$$\mathbb{E}_2^A\{x_2|x_1\} = \frac{\mathbb{E}_2^A\{\theta|x_1\}}{\gamma + \phi} \geq \Lambda(\phi) \Leftrightarrow \mathbb{E}_2^A\{\theta|x_1\} \geq (\gamma + \phi)\Lambda(\phi) \tag{7}$$

When this condition holds, the criminal is believed to be of such a bad type θ, that the loss the authority expects to incur if it sticks to its earlier intention to punish the criminal, is larger than the loss associated with allowing impunity. In those cases, the authority prefers to end the harm. Here, the *effective* punishment function becomes non-monotonic.

At this stage, one can define the critical type $\widehat{\theta}$. This type is defined such that when the criminal adheres to his FOC (3), and sets $x_1 = \widehat{x}_1 \equiv \widehat{\theta}/(\gamma+\phi)$, the authority's rational belief about θ is such that it becomes indifferent between impunity and punishment. $\widehat{\theta}$ therefore satisfies:

$$\mathbb{E}_2^A \left\{ \theta \middle| x_1 = \frac{\widehat{\theta}}{\gamma + \phi} \right\} = (\gamma + \phi)\, \Lambda\, (\phi) \tag{8}$$

Henceforth, we will refer to those criminals with $\theta \geq \widehat{\theta}$, as the "bad" ones: by adhering to their FOC, the authority is better off by offering them impunity in return for them escaping to another jurisdiction.

Most importantly, for a criminal of type $\theta < \widehat{\theta}$, unlocking the impunity deal calls for setting $x_1 > x_1^*$ (the latter choice being given by the FOC (3)). In those cases, a reasoning, "Beckerian" criminal who recognizes the finiteness of the authority's loss associated with allowing impunity, may choose to make an "investment" in period 1. The criminal "invests" by committing more crimes than he would actually like to from a static perspective - purely for the sake of unlocking the impunity-option, so that he can in the end walk away with no (or only a minor) punishment.

In this case, the criminal uses period 1 to try and signal that he is of the bad type (i.e.: that he is of a type $\theta \geq \widehat{\theta}$) thereby threatening that he will commit many crimes in period 2 as well. So many, that the authority is actually better off by making an impunity deal at the beginning of period 2.[14] The criminal is trading off the loss of not being at the static optimum in the first period, with the dynamic gain of being able to enjoy an exit of the game via an impunity deal. With reference to the behavior of Bosco Ntaganda, we will call this the "Terminator effect".

Note that these investments (or: signaling costs) are only going to be worthwhile for those criminals whose true preference parameter θ is sufficiently close to $\widehat{\theta}$, so it only pays for the high θ types to try and mimic the period 1 behavior of "bad" criminals. To see this, start by noting that a criminal whose $\theta < \widehat{\theta}$ can adopt two strategies:

[14] This is very much like the strategy employed by Bosco Ntaganda (nicknamed "The Terminator") who attacked the city of Goma in 2008 not because he was interested in its control, but only because he "wanted to make the government sit at the negotiating table" (cf. footnote 5). Robert Mugabe is said to have employed a similar strategy (recall footnote 6). In the 2012 documentary *Peace vs. Justice,* ICC-investigator Matthew Brubacher compares this behavior of dictators and warlords to blackmailing. He states: "It's essentially blackmailing. They kill as massively, as intensively, and as brutally as possible until the international community basically puts up its hand and says: 'OK, let's just go for peace. We'll give you money, we'll give you food, we'll give you whatever you want - just stop killing people.' "

$\mathcal{F}$. Follow the FOC (3), after which he expects punishment at the end of period 2.

$\mathcal{D}$. Deviate from the FOC (3) in period 1, to unlock the amnesty offer at the beginning of period 2 (which results in a final pay-off equal to V^I).

What is the utility that a type θ-criminal derives from these options? Let us first analyze the value from following strategy $\mathcal{F}$, henceforth indicated by $U^{\mathcal{F}}$. Since the criminal chooses to set $x_1 = x_2 = \theta / [\gamma + \phi]$ by FOC (3), it follows that:

$$U^{\mathcal{F}} = 2\theta \log\left(\frac{\theta}{\gamma + \phi}\right) - 2\theta \tag{9}$$

Given that we focus on criminals with $\theta < \widehat{\theta}$, setting x_1 according to the FOC (3) does not unlock the impunity-option (because $x_1^* < \widehat{x}_1$ for those types, with $\widehat{x}_1$ being the critical level of first-period crimes that unlocks the impunity-option).

The cheapest way for such a criminal to unlock the impunity-option nevertheless, would be to mimic a type $\widehat{\theta}$-criminal by setting $x_1 = \widehat{x}_1$ (this is strategy $\mathcal{D}$).[15] Following this strategy would yield a criminal of type $\theta < \widehat{\theta}$ lifetime utility:

$$U^{\mathcal{D}} = \theta \log(\widehat{x}_1) - \gamma \widehat{x}_1 + V^I(\theta, \phi) \tag{10}$$

Hence, a criminal of type $\theta < \widehat{\theta}$ will invest in committing crimes (and pool with the "bad" types) in period 1 iff his θ is such that $U^{\mathcal{D}} \geq U^{\mathcal{F}}$, i.e. iff:

$$y(\theta, \phi) \equiv \theta \log(\widehat{x}_1) - \gamma \widehat{x}_1 + V^I(\theta, \phi) - 2\theta \log\left(\frac{\theta}{\gamma + \phi}\right) + 2\theta \geq 0 \tag{11}$$

When $y(\theta, \phi) = 0$, it implicitly defines the critical type $\overline{\theta}$ beyond which a criminal becomes willing to invest in first-period crime for the sole purpose of unlocking the amnesty-option in period 2, such that he can exit via the amnesty route, enjoying $V^I(\overline{\theta}, \phi)$. We then know from combining equations (5) and (11) that $\overline{\theta}$ should satisfy:

$$y(\overline{\theta}, \phi) = \overline{\theta}\left[\log\left(\frac{\widehat{\theta}}{\overline{\theta}}\right) + 1\right] - \widehat{\theta} + \Delta = 0 \tag{12}$$

[15] After all, since $\widehat{x}_1 > x_1^*$, the marginal benefits from committing crimes are lower than its marginal costs at this point, as a result of which it will never be optimal for a criminal to go beyond that level of crime (setting $x_1 = \widehat{x}_1$ suffices to unlock the amnesty-option, so setting $x_1 > \widehat{x}_1$ only brings additional net costs).

Assuming that $\Delta < \widehat{\theta}$,[16] there exists a $\overline{\theta} < \widehat{\theta}$ for which $y(\overline{\theta}, \phi) = 0$ (because $y(\cdot)$ is a continuous function with $y(0, \phi) < 0$, $y(\widehat{\theta}, \phi) > 0$, and $\partial y(\theta, \phi)/\partial\theta|_{\theta=\overline{\theta}} > 0$). Consequently, all types $\theta \in [\overline{\theta}, \widehat{\theta})$ will find it optimal to mimic the $\widehat{\theta}$-type by setting $x_1 = \widehat{x}_1$. The reason is that in the face of a commitment problem for authorities, it pays to be *very* bad, rather than just a little bad.

At this stage, we are also able to specify how the authority forms its expectation about the criminal's type θ in Perfect Bayesian Equilibrium after observing the criminal's choice of x_1. It is given by:[17]

$$\mathbb{E}_2^A \{\theta | x_1\} = \begin{cases} \mathbb{E}\left\{\theta | \overline{\theta} \leq \theta \leq \widehat{\theta}\right\} = \int_{\overline{\theta}}^{\widehat{\theta}} \frac{\theta f(\theta) d\theta}{F(\widehat{\theta}) - F(\overline{\theta})} & \text{when } x_1 = \widehat{x}_1 \\ [\gamma + \phi]\, x_1 & \text{otherwise} \end{cases} \tag{13}$$

The intuition is that when the authority observes $\widehat{x}_1 \equiv \widehat{\theta}/[\gamma + \phi]$ (that level of crime which is just enough to unlock the impunity-option), the naive expectation would be to assume that the criminal is exactly of type $\widehat{\theta}$ (this would follow from blindly applying equation (4)). In equilibrium, however, the authority realizes that $\widehat{x}_1$ could also be set by a criminal of type $\theta \in [\overline{\theta}, \widehat{\theta})$, who is trying to mimic a criminal of type $\widehat{\theta}$ so as to unlock the impunity-option (the Terminator effect). Consequently, the rational expectation is $\mathbb{E}^A \left\{\theta | \overline{\theta} \leq \theta \leq \widehat{\theta}\right\} < \widehat{\theta}$ implying that the authority "discounts" the criminal's first-period behavior somewhat as it realizes that he might just be mimicking a bad criminal, without actually being one himself.[18]

To close the model, we still need to specify how the authority would respond if an out-of-equilibrium action (call that $\widetilde{x}$) were to be observed - otherwise there is a plethora of equilibria. We impose that $\mu(\theta | \widetilde{x}) = [\gamma + \phi]\, \widetilde{x} \equiv \widetilde{\theta}$. This implies that any out-of-equilibrium action is interpreted as coming from a criminal-type whose FOC (3) prescribes that action. Next to this belief seeming reasonable if one assumes that the authority allows for the possibility that some criminals may act non-strategically (dogmatically adhering to their FOC instead), it also survives the D1 refinement-arguments in Cho and Kreps (1987).[19]

[16] Otherwise $y(\theta, \phi) > 0\ \forall\theta$ and we end up in a situation where *all* types with $\theta > 0$ are going to mimic. This would actually strengthen our results, but strikes us as being rather extreme.

[17] We thank Avinash Dixit for pointing out an error in equation (13) in an earlier draft.

[18] Remember that, by equation (8), the authority becomes indifferent between impunity and punishment at type $\dddot{\theta} \equiv (\gamma + \phi)\, \Lambda(\phi)$ if the criminal's type were publicly observable. But since θ is hidden in our model, the authority discounts any first-period behavior as a result of which a criminal has to mimic a type $\widehat{\theta} > \dddot{\theta}$ to unlock the impunity-option.

[19] See Ramey (1996) who extends the D1-criterion to a continuum of types. While semi-pooling equilibria are often ruled out along these lines, it is sustained in our framework. The reason is the fact that

At this point, we know that $\overline{\theta}$ should solve (12), while $\widehat{\theta}$ solves equation (8) (with $\mathbb{E}_2^A\{\theta|x_1\}$ being formed as in (13)). Assuming for concreteness that $\theta \sim U(0,\theta_{\max}]$, we have $\mathbb{E}_2^A\{\theta|\widehat{x}_1\} = (\overline{\theta}+\widehat{\theta})/2$ and condition (8) implies:

$$\widehat{\theta} = 2\left(\gamma+\phi\right)\Lambda\left(\phi\right) - \overline{\theta} \tag{14}$$

Substituting this into (12) gives the following implicit equation characterizing $\overline{\theta}$:

$$y(\overline{\theta},\phi) = \overline{\theta}\log\left(\frac{2\left(\gamma+\phi\right)\Lambda\left(\phi\right)-\overline{\theta}}{\overline{\theta}}\right) + 2\overline{\theta} - 2\left(\gamma+\phi\right)\Lambda\left(\phi\right) + \Delta = 0 \tag{15}$$

This completes the solution for this part of the model.

2.3 Entry

What remains is a description of entry into criminal activity. At the beginning of period 1, there are two pools of potential criminals: pool $\mathcal{M}$ (where all potential criminals are malevolent, i.e. of type $\theta > 0$) and pool $\mathcal{H}$ (where all potential criminals are "harmless", i.e. of type $\theta = 0$). We normalize the measure of the malevolent pool to 1, while the harmless pool has measure b. Entering the criminal activity requires the payment of a fixed (and sunk) entry cost $\mathcal{E} > 0$, after which the agent will initiate the criminal activity, which has success-probability p.

We assume that harmless types are never willing to begin such criminal activity. Agents in pool $\mathcal{M}$ on the other hand, are malevolent and only interested in maximizing their own private pay-off. They are distributed according to cumulative distribution function $F(\theta)$, with support $(0,\theta_{\max}]$. Upon observing their private draw of θ, these malevolent agents decide whether to initiate the criminal activity.

Now consider a criminal with $\theta < \overline{\theta}$. By definition of $\overline{\theta}$, he does not want to unlock the impunity-option by mimicking a bad criminal, because doing so is too costly for him given his type. Instead, he will follow his FOC, which would bring him lifetime utility as in (9). Consequently, a type θ agent in the potential pool of malevolent criminals would

the action of the receiver in our game (the authority) is binary: it offers immunity, or it sticks to ex-post punishment. This implies that $\widehat{\theta}$-types have no incentive to try and separate themselves from lower, mimicking types.

only want to initiate the criminal activity if his θ is such that:

$$z(\theta,\phi) \equiv p\left(\underline{\theta}\right)\left[2\theta \log\left(\frac{\theta}{\gamma+\phi}\right) - 2\theta\right] - \mathcal{E} \geq 0 \tag{16}$$

When $z(\theta,\phi) = 0$, it defines the critical level $\underline{\theta}$ below which potential malevolent criminals do not find it worthwhile to initiate the crime.[20] The function $p\left(\underline{\theta}\right)$ (with $dp(\underline{\theta})/d\underline{\theta} > 0$) captures the idea that any given individual initiating a crime has a higher probability of success, if the number of initiators is relatively low (which is the case when the entry threshold $\underline{\theta}$ is high).[21]

We furthermore impose that at least some malevolent criminals are willing to initiate a crime (which is the whole premise of this paper). This requires $\gamma + \phi < \underline{\theta}$, otherwise $z(\theta,\phi) < 0\ \forall\theta$. Note that $\underline{\theta}$ is unique, as $z(\cdot)$ is a continuous function with $z(\theta,\phi)|_{\theta<\underline{\theta}} < 0$, $z(\theta,\phi)|_{\theta>\underline{\theta}} > 0$, and $\partial z(\theta,\phi)/\partial\theta > 0$.

Potential malevolent criminals of type $\theta < \underline{\theta}$ are deterred by the severe punishment for crimes. As a result it is not worthwhile for them to pay the entrance fee $\mathcal{E}$ to engage in a criminal activity. Consequently, only a fraction $[1 - F(\underline{\theta})]$ out of the pool of potential malevolent criminals will decide to initiate such a crime.

2.4 Summary

In summary, our model leads to four groups of malevolent criminals:

- The "petty criminals" with $0 < \theta < \underline{\theta}$. They choose not to initiate a crime because of the entry fee $\mathcal{E}$, as their low θ implies that they are not able to derive enough utility from committing crimes to make up for that fee.

- The "ordinary criminals" with $\underline{\theta} \leq \theta < \overline{\theta}$. This group chooses to commit a crime, but for them it is too costly to change their behavior in order to unlock the impunity-option. Consequently, they just stick to their FOCs and optimally choose to undergo the punishment at the end of period 2. Stated in terms of the signaling literature: these criminals are not able to mimic the behavior of the bad types, for whom $\theta \geq \widehat{\theta}$.

- The "mimicking criminals" with $\overline{\theta} \leq \theta < \widehat{\theta}$. This is an interesting group, as they choose to modify their first-period behavior by "investing" in committing more or

[20] We assume $\mathcal{E}$ to be low enough such that $\underline{\theta} < \overline{\theta}$, otherwise all entering malevolent criminals will again choose to mimic the bad ones. As noted before, such a specification would strengthen the results that are to follow, but strikes us as being rather extreme.

[21] See Friedman and Sjostrom (1993) who model the success of criminal activity on the pool of potential offenders.

more gruesome first-period crimes, for the sole purpose of unlocking the impunity-option at the beginning of period 2. In signaling terms, they are able to mimic and form a pooling equilibrium with the bad criminals. These types are actively trying to break the authority's earlier commitment to ex-post punishment.

- The "bad criminals" with $\theta \geq \widehat{\theta}$. For them, following the FOCs already suffices to unlock the impunity option– no additional investments are necessary. Note that criminals in this group have no incentive to try and form a separating equilibrium (as this is costly to them, without delivering any benefits).

All of these thresholds are unique and satisfy the ordering $\underline{\theta} < \overline{\theta} < \widehat{\theta}$ - as discussed in Sections 2.2 and 2.3.

3 Comparative Statics

We are now able to analyze what would happen if the authority raises the expected punishment parameter ϕ, thereby raising the probability of prosecution. To answer this question, it is crucial to analyze the comparative statics of the various thresholds with respect to ϕ, which is done in Propositions 1-4. Analytical results can only be obtained when θ is assumed to follow a uniform distribution. Extensive numerical analysis using other distributions on the positive domain for θ (such as the log-normal, the gamma, and the Pareto) however suggests that our results apply more generally. It is moreover possible to solve the model *without* making any distributional assumptions on θ if one supposes that the authority is boundedly rational and naively applies (4) (rather than (13)) when forming $\mathbb{E}_2^A\left\{\theta|x_1\right\}$. Reassuringly, our findings are robust to that specification as well.

Proposition 1.

(i) Provided that some malevolent types are willing to initiate a crime (which is the case when $\underline{\theta} > \gamma+\phi$*), an increase in expected punishment raises the entry-threshold. That is:* $\partial\underline{\theta}/\partial\phi > 0$.

(ii) An increase in expected punishment reduces the likelihood of initiating a crime while selecting more malevolent criminals. That is: $\partial \Pr\left[\mathbf{1}_{\mathcal{H}} = 1\right]/\partial\phi > 0$ *and* $\partial\mathbb{E}\left\{\theta|\theta \geq \underline{\theta}\right\}/\partial\phi > 0$

All proofs can be found in Appendix A.

Proposition 1 states that an increase in ϕ raises the threshold beyond which agents in the pool of potential malevolent criminals become willing to initiate a crime (imposing that at least some malevolent criminals are willing to try, i.e. $\underline{\theta} > \gamma + \phi$).[22] Consequently, the greater the authority's commitment to prosecuting criminal activities (higher ϕ), the lower the fraction $[1 - F(\underline{\theta})]$ of potential malevolent criminals that will decide to initiate a crime. This is the deterrence effect (captured by equation 17 in Appendix A). As a result, the probability that a crime is initiated decreases with the punishment parameter ϕ, which is intuitive.

However, conditional on a crime being initiated, a greater commitment to prosecuting crimes also selects criminals who, on average, will be of a worse type. The reason is that the deterrence effect only drives out potential petty criminals (with a relatively low θ). Thus, the benign deterrence effect is accompanied by a malign adverse selection effect. Which of these two effects dominates depends on the distribution of θ.

Similarly, we can be establish how increased commitment to punishing offenders affect thresholds $\overline{\theta}$ and $\widehat{\theta}$.

Proposition 2. *An increase in expected punishment raises the threshold beyond which criminals become willing to invest in committing crimes. That is:* $\partial\overline{\theta}/\partial\phi > 0$*.*

Remember that all criminals of type $\theta > \overline{\theta}$ will choose to form a pooling equilibrium with the $\widehat{\theta}$-types. With respect to the location of this level, it holds that:

Proposition 3. *An increase in expected punishment raises the threshold beyond which the authority chooses to enter into an impunity deal. That is:* $\partial\widehat{\theta}/\partial\phi > 0$*.*

This proposition is intuitive given that the costs of granting an impunity deal ex post are increasing in ϕ (remember that $d\Lambda(\phi)/d\phi > 0$). It implies that when ϕ is higher, criminals of type $\theta < \widehat{\theta}$ will have to put in a greater mimicking effort (commit more gruesome acts) if they want to unlock the amnesty-option.

[22] Since the last term in the denominator of $\partial\underline{\theta}/\partial\phi$ is positive (see the expression in Appendix A), this condition is actually stricter than necessary. But given that this paper starts from the observation that some malevolent criminals are willing to attempt to commit a crime, we maintain it nevertheless as it maximizes both clarity as well as descriptive realism without serious loss of generality.

With respect to the size of the mimicking group, which is increasing in $(\widehat{\theta} - \overline{\theta})$, one can show that:

PROPOSITION 4. *The size of the mimicking group is increasing in expected punishment. That is:* $\partial(\widehat{\theta} - \overline{\theta})/\partial\phi > 0$.

This proposition indicates that as ϕ increases, more criminals become subject to the Terminator effect and choose to worsen their behavior to unlock the impunity-option. The reason is that when expected punishment ϕ is higher, exit via the impunity-escape route becomes relatively more valuable. Consequently, more become willing to pool with the bad types at $\widehat{\theta}$.

Finally, there is also a "disciplining effect" for those criminals with $\theta \in \big[\underline{\theta}, \overline{\theta}\big)$ or $\theta \in [\widehat{\theta}, \theta_{\max}]$. They just choose to adhere to their FOCs (given by (3)). As shown in the following proposition, the prospect of higher punishment at the end of period 2 induces them to commit less crime:

PROPOSITION 5. *Criminals with* $\theta \in \big[\underline{\theta}, \overline{\theta}\big) \cup [\widehat{\theta}, \theta_{\max}]$ *are disciplined by increases in expected punishment. That is:* $\partial x_t^*/\partial\phi < 0$.

Combining these results (summarized in Table 1), tells us that increasing ϕ (and hence $\Lambda(\phi)$) will lead to greater dispersion in outcomes. On the one hand, there will be fewer crimes in the long run (by the deterrence effect captured by equation (17) in Appendix A) and some types of criminals (those with $\theta \in \big[\underline{\theta}, \overline{\theta}\big)$ or $\theta \in [\widehat{\theta}, \theta_{\max}]$) will be disciplined (this effect is present in the short run; Proposition 5). On the other hand, however, there are two malign effects that make the realization of more extreme, less favorable outcomes more likely. First, there is the (long run) adverse selection effect which worsens the quality of the active criminal pool (see equation (18) in Appendix A). Second, by the Terminator effect (which operates in the short run), criminals in the mimicking group will choose to commit more crimes following an increase in ϕ, as this is now necessary to unlock the impunity-option (Proposition 3). In addition, the size of this mimicking group (determined by $(\widehat{\theta} - \overline{\theta})$) increases with ϕ (Proposition 4) - so increased commitment to ex-post punishment leads to a worse type of criminal who will choose to *act worse* for strategic reasons. Whether the expected net effect is benign, cannot be established unambiguously. In any case, parties involved should ask themselves whether they consider the inevitable increase in dispersion in outcomes to be desirable or not.

	Benign	*Malign*
Short run	Disciplining effect for $\theta \in \big[\underline{\theta}, \overline{\theta}\big) \cup [\widehat{\theta}, \theta_{\max}]$	Terminator effect for $\theta \in [\overline{\theta}, \widehat{\theta})$
Long run	Deterrence effect	Adverse selection effect

Table 1: Effects brought about by an increase in ϕ.

With respect to Table 1, it is interesting to note how the nature of the long run effects is unconditional: the deterrence effect is unconditionally benign, while the adverse selection effect is unconditionally malign. The short run effect of increasing ϕ, on the other hand, is conditional: when the criminal is of type $\theta \in \big[\underline{\theta}, \overline{\theta}\big)$ or $\theta \in [\widehat{\theta}, \theta_{\max}]$, the effect is good, whereas it is bad if the criminal is of type $\theta \in [\overline{\theta}, \widehat{\theta})$. This contributes to the aforementioned polarization in behavior following an increase in ϕ.

It more generally shows that, when authorities lack a perfect commitment mechanism, increased commitment to ex-post punishment can paradoxically lead to more crimes being committed (if the Terminator and adverse selection effect outweigh the disciplining and deterrence effect). This can actually happen in the short run (if the Terminator effect dominates the disciplining effect) and forms a striking contrast with standard "Beckerian" models of criminal activity (where increases in expected punishment tends to reduce crime).

4 Conclusion

This paper has analyzed a criminal enforcement problem for a special class of crime. This class of crime concerns cases where the criminal can be easily detected but not easily apprehended (at least for some period time time). In those cases, the criminal has the ability to inflict significant harm in the interim period between detection and apprehension. Criminals may therefore actively engage in strategic behavior to try to force authorities to ignore earlier commitments to severe punishment. In such an environment, it is dangerous for regulators to *try* to commit, particularly when they lack a perfect commitment technology.

Our model identifies two benign effects associated with higher expected punishment: first of all, there will be a long run deterrence effect - leading to fewer malevolent criminals in the future. Secondly, some malevolent criminals will be disciplined in the short run. Simultaneously, however, more malevolent offenders who will choose to commit *more* atrocities when faced with a higher expected punishment (to try and force authorities to grant an impunity deal). Dispersion in outcomes will therefore go up, so parties involved should ask themselves whether they consider such polarization to be a desirable outcome.

In summary, while commitment should reduce the instances of these crimes, those that do occur are likely to be worse and more gruesome.

Our results are largely positive (as opposed to normative). Nonetheless we consider them to be important, as finding the right institutional response to these types of crimes has the ability to greatly increase social welfare. For one, while the types of crimes analyzed above are relatively rare, they can strike terror into a populace. Second, in an international law context, the stakes can be high, and a design-flaw may cost millions of lives.

5 Appendix A

Proposition 1 *(i) When $\underline{\theta} > \gamma + \phi$, $\partial\underline{\theta}/\partial\phi > 0$.*

Proof. The threshold $\underline{\theta}$ is implicitly defined by $z(\underline{\theta}, \phi) = 0$ (with z following the specification of equation (16)). Applying the implicit function theorem yields:

$$\frac{\partial\underline{\theta}}{\partial\phi} = \frac{\frac{\underline{\theta}}{\gamma+\phi}}{\log\left(\frac{\underline{\theta}}{\gamma+\phi}\right) + \frac{\mathcal{E}}{p(\underline{\theta})^2}\frac{dp(\underline{\theta})}{d\underline{\theta}}}$$

Since $\mathcal{E}$ >0, condition (16) shows that malevolent criminals with $\theta > \underline{\theta}$ will only be willing to run for criminal activity when $\gamma + \phi < \underline{\theta}$. Recalling that $0 < p\left(\underline{\theta}\right) < 1$ and $dp\left(\underline{\theta}\right)/d\underline{\theta} > 0$, this immediately implies that $\partial\underline{\theta}/\partial\phi > 0$.

(ii) $\partial\Pr\left[\mathbf{1}_{\mathcal{H}} = 1\right]/\partial\phi > 0$ and $\partial\mathbb{E}\left\{\theta|\theta \geq \underline{\theta}\right\}/\partial\phi > 0$

Let $\mathbf{1}_{\mathcal{H}}$ be an indicator function that takes the value 1 if no one commits a crime. The ex-ante probability of this event occurring is:

$$\Pr\left[\mathbf{1}_{\mathcal{H}} = 1\right] = \frac{b}{b + 1 - F(\underline{\theta})},$$

with:

$$\frac{\partial\Pr\left[\mathbf{1}_{\mathcal{H}} = 1\right]}{\partial\phi} = \frac{b \cdot \partial F(\underline{\theta})/\partial\underline{\theta} \cdot \partial\underline{\theta}/\partial\phi}{\left[b + 1 - F(\underline{\theta})\right]^2} > 0, \text{ since } \partial\underline{\theta}/\partial\phi > 0 \tag{17}$$

In addition, since $\mathbb{E}\left\{\theta|\mathbf{1}_{\mathcal{H}} = 0\right\} = \mathbb{E}\left\{\theta|\theta \geq \underline{\theta}\right\}$, it holds that:

$$\frac{\partial\mathbb{E}\left\{\theta|\theta \geq \underline{\theta}\right\}}{\partial\phi} > 0, \tag{18}$$

again because $\partial\underline{\theta}/\partial\phi > 0$. ■

Proposition 2 $\partial\overline{\theta}/\partial\phi > 0$.

Proof. The threshold $\overline{\theta}$ is implicitly defined by $y(\overline{\theta}, \phi) = 0$. Differentiating y (specified in equation (15)) with respect to its two arguments, and using equation (14) gives:

$$\begin{aligned}\frac{\partial y\left(\overline{\theta}, \phi\right)}{\partial\phi} &= 2\left[\Lambda\left(\phi\right) + \left(\gamma + \phi\right)\frac{d\Lambda\left(\phi\right)}{d\phi}\right]\left[\frac{\overline{\theta}}{2\left(\gamma + \phi\right)\Lambda\left(\phi\right) - \overline{\theta}} - 1\right] \\ &= 2\left[\Lambda\left(\phi\right) + \left(\gamma + \phi\right)\frac{d\Lambda\left(\phi\right)}{d\phi}\right]\left[\frac{\overline{\theta}}{\widehat{\theta}} - 1\right] < 0,\end{aligned}$$

since $\overline{\theta} < \widehat{\theta}$. By the same arguments,

$$\frac{\partial y\left(\overline{\theta}, \phi\right)}{\partial \overline{\theta}} = \log\left(\frac{2\left(\gamma+\phi\right)\Lambda\left(\phi\right)-\overline{\theta}}{\overline{\theta}}\right) + \frac{2\left[\left(\gamma+\phi\right)\Lambda\left(\phi\right)-\overline{\theta}\right]}{2\left(\gamma+\phi\right)\Lambda\left(\phi\right)-\overline{\theta}} = \log\left(\frac{\widehat{\theta}}{\overline{\theta}}\right) + \frac{\widehat{\theta}-\overline{\theta}}{\widehat{\theta}} > 0$$

Consequently, the implicit function theorem implies:

$$\frac{\partial \overline{\theta}}{\partial \phi} = -\frac{\partial y/\partial \phi}{\partial y/\partial \overline{\theta}} = -\frac{2\left[\Lambda\left(\phi\right)+\left(\gamma+\phi\right)\frac{d\Lambda(\phi)}{d\phi}\right]\left[\frac{\overline{\theta}}{\widehat{\theta}}-1\right]}{\log\left(\frac{\widehat{\theta}}{\overline{\theta}}\right)+\frac{\widehat{\theta}-\overline{\theta}}{\widehat{\theta}}} > 0$$

■

Proposition 3 $\partial\widehat{\theta}/\partial\phi > 0$.

Proof. Differentiating (14), and realizing that $\overline{\theta} < \widehat{\theta}$, gives:

$$\begin{aligned}\frac{\partial \widehat{\theta}}{\partial \phi} &= 2\left[\Lambda\left(\phi\right)+\left(\gamma+\phi\right)\frac{d\Lambda\left(\phi\right)}{d\phi}\right] - \frac{\partial \overline{\theta}}{\partial \phi} \\ &= 2\left[\Lambda\left(\phi\right)+\left(\gamma+\phi\right)\frac{d\Lambda\left(\phi\right)}{d\phi}\right]\left[\frac{\log\left(\frac{\widehat{\theta}}{\overline{\theta}}\right)}{\log\left(\frac{\widehat{\theta}}{\overline{\theta}}\right)+\frac{\widehat{\theta}-\overline{\theta}}{\widehat{\theta}}}\right] > 0\end{aligned}$$

■

Proposition 4 $\partial(\widehat{\theta}-\overline{\theta})/\partial\phi > 0$.

Proof. Combining the derivatives obtained in Propositions 2 and 3 yields:

$$\frac{\partial(\widehat{\theta}-\overline{\theta})}{\partial \phi} = \frac{2\left[\Lambda\left(\phi\right)+\left(\gamma+\phi\right)\frac{d\Lambda(\phi)}{d\phi}\right]\left[\log\left(\frac{\widehat{\theta}}{\overline{\theta}}\right)+\frac{\overline{\theta}}{\widehat{\theta}}-1\right]}{\log\left(\frac{\widehat{\theta}}{\overline{\theta}}\right)+\frac{\widehat{\theta}-\overline{\theta}}{\widehat{\theta}}}$$

Note from this expression that $sgn\left(\partial(\widehat{\theta}-\overline{\theta})/\partial\phi\right) = sgn\left(\log\left(\widehat{\theta}/\overline{\theta}\right)+\overline{\theta}/\widehat{\theta}-1\right)$. To determine the latter, define $q \equiv \widehat{\theta}/\overline{\theta} > 1$ and $f(q) \equiv \log(q) + 1/q - 1$. When $f(q) > 0$, we have that $\partial(\widehat{\theta}-\overline{\theta})/\partial\phi > 0$. One can show that this indeed is the case for $q > 1$ by using the Lambert function $W(z)$, for which it holds that $W(z)\exp\left(W(z)\right) = z$. To see that $f(q) > 0$ when $q > 1$, rewrite:

$$\log(q)+\frac{1}{q}-1 > 0 \Leftrightarrow q > \exp\left(1-\frac{1}{q}\right) = \exp(1)\exp\left(-\frac{1}{q}\right) \Leftrightarrow -\frac{1}{\exp(1)} < -\frac{1}{q}\exp\left(-\frac{1}{q}\right)$$

Using the Lambert W function gives:

$$W\left(-\frac{1}{\exp(1)}\right) < W\left(-\frac{1}{q}\exp\left(-\frac{1}{q}\right)\right)$$

By the defining property of the Lambert function we know that $W\left(-1/q\exp\left(-1/q\right)\right) = -1/q$ and $W\left(-1/\exp(1)\right) = -1$. It thus follows that $\log\left(q\right) + 1/q - 1 > 0$ for:

$$-1 < -\frac{1}{q} \Leftrightarrow q > 1,$$

which is the case since $q \equiv \widehat{\theta}/\overline{\theta}$ and $\overline{\theta} < \widehat{\theta}$. ■

Proposition 5 $\partial x_t^*/\partial\phi < 0$.

Proof. All criminals with $\theta \in \left[\underline{\theta}, \overline{\theta}\right) \cup [\widehat{\theta}, \theta_{\max}]$, set x_t according to FOC (3). The proof then follows immediately from differentiating that expression. ■

6 Appendix B

In this Appendix we show that all propositions stated in Section III are robust to using the more general $\Lambda = \Lambda(\phi, \phi x_1)$, with $\Lambda_1 \equiv \partial\Lambda\left(\phi, \phi x_1\right)/\partial\phi > 0$ and $\Lambda_2 \equiv \partial\Lambda\left(\phi, \phi x_1\right)/\partial\left(\phi x_1\right) > 0$. Having $\Lambda_2 > 0$ captures the idea that it is costlier for the authority to allow impunity to criminals who behave worse in period 1 and should have received a large punishment according to the strict letter of the law. Propositions 1 and 5 are not affected by this modification, so we only have to check Propositions 2, 3, and 4.

With $\Lambda = \Lambda(\phi, \phi x_1)$ we have that:

$$\frac{\partial y\left(\overline{\theta}, \phi\right)}{\partial\phi} = 4\left[\Lambda\left(\phi, \phi\widehat{x}_1\right) + (\gamma+\phi)\Lambda_1\left(\phi, \phi\widehat{x}_1\right) + \frac{\gamma\widehat{\theta}}{\gamma+\phi}\Lambda_2\left(\phi, \phi\widehat{x}_1\right)\right]\left[\frac{\overline{\theta} - (\gamma+\phi)\Lambda\left(\phi, \phi\widehat{x}_1\right)}{2(\gamma+\phi)\Lambda\left(\phi, \phi\widehat{x}_1\right) - \overline{\theta}}\right]$$

Now recall that $\widehat{\theta} = 2(\gamma+\phi)\Lambda\left(\phi, \phi\widehat{x}_1\right) - \overline{\theta}$, which implies $\frac{\widehat{\theta}-\overline{\theta}}{2} = (\gamma+\phi)\Lambda\left(\phi, \phi\widehat{x}_1\right) - \overline{\theta} > 0$. As a result:

$$\frac{\partial y\left(\overline{\theta}, \phi\right)}{\partial\phi} = 4\left[\Lambda\left(\phi, \phi\widehat{x}_1\right) + (\gamma+\phi)\Lambda_1\left(\phi, \phi\widehat{x}_1\right) + \frac{\gamma\widehat{\theta}}{\gamma+\phi}\Lambda_2\left(\phi, \phi\widehat{x}_1\right)\right]\left[-\frac{\frac{\widehat{\theta}-\overline{\theta}}{2}}{\widehat{\theta}}\right] = 2\Psi\left[\frac{\overline{\theta}}{\widehat{\theta}} - 1\right] < 0,$$

where $\Psi \equiv \Lambda\left(\phi, \phi \widehat{x}_1\right) + (\gamma + \phi)\Lambda_1\left(\phi, \phi \widehat{x}_1\right) + \frac{\gamma \widehat{\theta}}{\gamma + \phi}\Lambda_2\left(\phi, \phi \widehat{x}_1\right) > 0$ and $\overline{\theta} < \widehat{\theta}$.

As before:

$$\frac{\partial y\left(\overline{\theta}, \phi\right)}{\partial \overline{\theta}} = \log\left[\frac{\widehat{\theta}}{\overline{\theta}}\right] + \frac{\widehat{\theta} - \overline{\theta}}{\widehat{\theta}} > 0$$

Consequently, by the implicit function theorem:

$$\frac{\partial \overline{\theta}}{\partial \phi} = \frac{2\Psi\left[\frac{\widehat{\theta} - \overline{\theta}}{\widehat{\theta}}\right]}{\log\left[\frac{\widehat{\theta}}{\overline{\theta}}\right] + \frac{\widehat{\theta} - \overline{\theta}}{\widehat{\theta}}} > 0,$$

which proves Proposition 2 for the case in which $\Lambda = \Lambda(\phi, \phi x_1)$.

Using straightforward calculus, one can subsequently show that the derivative central to Proposition 3 is now going to be given by:

$$\frac{\partial \widehat{\theta}}{\partial \phi} = \frac{2\Psi \log\left[\frac{\widehat{\theta}}{\overline{\theta}}\right]}{\log\left[\frac{\widehat{\theta}}{\overline{\theta}}\right] + \frac{\widehat{\theta} - \overline{\theta}}{\widehat{\theta}}} > 0,$$

which confirms Proposition 3.

From this, it immediately follows that:

$$\frac{\partial\left(\widehat{\theta} - \overline{\theta}\right)}{\partial \phi} = \frac{2\Psi\left(\log\left[\frac{\widehat{\theta}}{\overline{\theta}}\right] + \frac{\overline{\theta}}{\widehat{\theta}} - 1\right)}{\log\left[\frac{\widehat{\theta}}{\overline{\theta}}\right] + \frac{\widehat{\theta} - \overline{\theta}}{\widehat{\theta}}}$$

Since $\log\left[\frac{\widehat{\theta}}{\overline{\theta}}\right] + \frac{\overline{\theta}}{\widehat{\theta}} - 1 > 0$ for $\frac{\widehat{\theta}}{\overline{\theta}} > 1$, this proves Proposition 4.

7 References

Australian Senate (2011). Held Hostage: Government's Response to Kidnapping of Australian Citizens Overseas. Canberra: Senate Printing Unit.

Detotto, C., McCannon, B.C. and Vannini, M., 2015. Evidence of marginal deterrence: Kidnapping and murder in Italy. International Review of Law and Economics, 41, 63-67.

Cho, I.K. and Kreps, D.M. (1987). Signaling Games and Stable Equilibria. Quarterly Journal of Economics, 102 (2), 179-221.

D'ancona, M. (2013). In It Together: The Inside Story of the Coalition Government. New York: Viking Press.

Detotto, C., McCannon, B. C., & Vannini, M. (2014). Understanding ransom kidnappings and their duration. The BE Journal of Economic Analysis & Policy, 14(3), 849-871.

Escriba-Folch, A. (2013). Accountable for What? Regime Types, Performance, and the Fate of Outgoing Dictators, 1946-2004. Democratization, 20 (1), 160-185.

Escriba-Folch, A. and Wright, J. (2015). Human Rights Prosecutions and Autocratic Survival. International Organization, 69(2), 343-373.

Freeman, M. (2009). Necessary Evils: Amnesties and the Search for Justice. Cambridge: Cambridge University Press.

Friedman, D. and Sjostrom, W. (1993). Hanged for a sheep: The economics of marginal deterrence. The Journal of Legal Studies, 22(2), 345-366.

Goldsmith, J. (2003). The Self-defeating International Criminal Court. The University of Chicago Law Review, 70(1), 89-104.

Kydland, F.E. and Prescott, E.C. (1977). Rules Rather than Discretion: The Inconsistency of Optimal Plans. Journal of Political Economy, 85 (3), 473-492.

Lohmann, S. (1992). Optimal Commitment in Monetary Policy: Credibility versus Flexibility. American Economic Review, 82 (1), 273-286.

Mookherjee, D. and Png, I.P. (1994). Marginal deterrence in enforcement of law. Journal of Political Economy, 102(5), 1039-1066.

Osiel, M. (2000). Why Prosecute? Critics of Punishment for Mass Atrocity. Human Rights Quarterly, 22(1), 118-147.

Polinsky, A.M. and Shavell, S. (2007). The theory of public enforcement of law. Handbook of Law and Economics, 1, 403-454.

Persson, M. and Siven, C.H. (2007). The Becker Paradox and Type I versus Type II Errors in the Economics of Crime. International Economic Review, 48 (1), 211-233.

Ramey, G. (1996). D1 Signaling Equilibria with Multiple Signals and a Continuum of Types. Journal of Economic Theory, 69 (2), 508-531.

Ritter, E.H. and Wolford, S. (2012). Bargaining and the Effectiveness of International Criminal Regimes. Journal of Theoretical Politics, 24 (2), 149-171.

Scharf, M.P. (1999). The Amnesty Exception to the Jurisdiction of the International Criminal Court. Cornell International Law Journal, 32, 507-527.

Scharf, M.P. (2006). From the eXile Files: An Essay on Trading Justice for Peace. Washington and Lee Law Review, 63 (1), 339-376.

Schwarz, M. and Sonin, K. (2008). A theory of brinkmanship, conflicts, and commitments. Journal of Law, Economics, and Organization, 24(1), 163-183.

Shavell, S. (1992). A note on marginal deterrence. International review of Law and Economics, 12(3), 345-355.

Snyder, J. and Vinjamuri, L. (2006). Trials and Errors: Principle and Pragmatism in Strategies of International Justice. International Security, 28(3), 5-44.

Sobel, J. and Takahashi, I. (1983). A Multistage Model of Bargaining. Review of Economic Studies, 50 (3), 411-426.

Stigler, G.J. (1970). The optimum enforcement of laws. Journal of Political Economy, 78(3), 526-536.

Sutter, D. (1995). Settling Old Scores Potholes along the Transition from Authoritarian Rule. Journal of Conflict Resolution, 39(1), 110-128.

Vannini, M., Detotto, C. and McCannon, B. (2015). Ransom Kidnapping. Encyclopedia of Law and Economics. 1-12.