

What shall we do with the bad dictator?

Time-consistency in the presence of strategic agents*

Shaun Larcom[†]

Mare Sarr[‡]

Tim Willems[§]

October 2014

Abstract

Recently, the international community has increased its commitment to prosecute malevolent dictators by establishing the International Criminal Court, thereby raising its loss of being time-inconsistent (granting amnesties ex post). This deters dictators from committing war crimes ex ante. Simultaneously, however, such commitment selects dictators of a worse type. Moreover, when a dictator behaves so badly that the costs of keeping him in place are considered to be greater than those of being time-inconsistent, the international community will still grant amnesty. Consequently, increased commitment to ex-post punishment may actually induce dictators to *worsen* their behavior, purely to "unlock" the amnesty option and force the international community into time-inconsistency. This leads to a general lesson: when regulators lack a perfect commitment technology, it is dangerous for them to *try to* commit as that may invoke a strategic response from regulatees which worsens the original problem.

JEL-classification: F55, K14, O12

Key words: Dictatorship, Time-consistency, International Criminal Court, Amnesty, Institutions

*We thank Haim Abraham, Alp Atakan, Francesco Caselli, Alex de Cornière, Avinash Dixit, Mikhail Drugov, Georgy Egorov, Peter Eso, Leonardo Felli, Ian Jewitt, Antonio Mele, Meg Meyer, Ines Moreno de Barreda, Ricardo Nunes, 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, and at the Universities of Oxford 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 and Centre for Development, Environment and Policy, SOAS, University of London. E-mail: stl25@cam.ac.uk.

[‡]School of Economics, University of Cape Town. E-mail: mare.sarr@uct.ac.za.

[§]Nuffield College, Department of Economics, University of Oxford, and Centre for Macroeconomics. E-mail: tim.willems@economics.ox.ac.uk.

1 Introduction

If you kill one person, you go to jail; if you kill 20, you go to an institution for the insane; if you kill 20,000, you get political asylum.

Reed Brody, counsel for Human Rights Watch

When it comes to the question of how to punish malicious rulers, there is a conflict between the ex ante and the ex post perspective. Ex ante, authorities want to threaten potential malevolent dictators with severe penalties to deter them from committing atrocities. But while these crimes are being committed, there may be strong pressures to put the sense of justice aside and grant the dictator amnesty in return for his abdication, to end further suffering for civilians in the quickest possible way. In this paper, we develop a model that is able to study this time-consistency problem.

Our analysis is motivated by the observation that the international community has recently increased its commitment to prosecute malevolent dictators: on May 30 2012, the Special Court for Sierra Leone sentenced Charles Taylor, former president of Liberia, to 50 years imprisonment for war crimes and crimes against humanity. Taylor is the first former head of state to be convicted by an international criminal tribunal since Karl Dönitz (the German admiral who briefly succeeded Adolf Hitler upon his death). In response to Taylor's conviction, the prosecutor of the Court concluded that this "historic judgment reinforces the new reality, that heads of state will be held to account for war crimes"; Amnesty International hailed the verdict as "sending an important message to high-ranking state officials".

Recent events indeed suggest that dictators are facing a new set of incentives. Next to the ex-post establishment of several ad-hoc tribunals (such as those for Rwanda, former Yugoslavia, and Sierra Leone), the international community has installed a *permanent* International Criminal Court (henceforth: "ICC"). It was established in 2002 to try persons for acts of genocide, war crimes, and crimes against humanity. 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.

All this constitutes an important regime change: in the past, malicious rulers were often granted amnesty or asylum to hasten their departure and alleviate suffering. This practice of "trading justice for peace" was not only popular with dictators,¹ but was often

¹The list of malicious dictators who have accepted amnesty or asylum offers is long. In recent decades amnesties were offered (and accepted) as part of peace arrangements in Angola, Argentina, Brazil, Cam-

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

Similarly, Charles Taylor stepped down as president of Liberia in 2003 when rebel troops were surrounding the city of Monrovia, in return for asylum in Nigeria. Scharf (2006) considers that this action, which brought the fighting to an immediate halt, may have saved the lives of tens of thousands of civilians. In addition, Taylor's abdication has improved the quality of life in Liberia: under the current president, Ellen Johnson Sirleaf - a Nobel Peace Laureate - Liberia's Human Development Index is showing a steady increase after more than two decades of decline. Although the Nigerian government in the end broke its promise (rumoured to be in exchange for debt relief) and handed Taylor over to the Special Court,² this case does illustrate the potential of improving conditions by promising impunity to dictators in return for their abdication.

Some legal scholars, such as Scharf (1999), have therefore criticized the new incentives - arguing that an ICC may bring "justice at the expense of peace". While earlier work has already pointed out that it may be harder to remove dictators if they face high penalties upon their exit (*cf.* Escribà-Folch (2013)), our model uncovers a new cost associated with increased commitment to ex-post punishment. It starts by noting that although the international community's loss associated with being time-inconsistent has increased with the installation of the ICC, it continues to be finite. Consequently, there still exists a critical level of atrocities beyond which the international community (or any government, which could offer asylum or local amnesty) chooses to take the loss associated with being time-inconsistent, rather than the pay-off that results from sticking to its earlier threats to prosecute the dictator. Consider the following statement by David Cameron, prime minister of the United Kingdom (which is party to the Rome Statute of the ICC), in response to the crimes committed by Bashar al-Assad in Syria around November 2012:

bodia, Chile, El Salvador, Guatemala, Honduras, Ivory Coast, Nicaragua, Peru, Sierra Leone, South Africa, Togo, and Uruguay, while asylum was for example given to the former Ethiopian leader Mengistu Haile Mariam in Zimbabwe, former Cuban president Fulgencio Batista in Portugal, and former Philippine leader Ferdinand Marcos in Hawaii. More examples will follow in the remainder of this paper.

²This shows that there is also a time-consistency problem with promises *not* to prosecute. In this sense, an asylum or amnesty backed by the international community (like the Westphalia Peace Treaty) seems to entail more commitment than a local amnesty offer. Nevertheless, dictators are still eager to accept such offers (recall footnote 1, while there are also numerous examples of dictators who have accepted amnesty offers "post-Taylor": see p. 4) - perhaps because it at least enables them to delay prosecution, with a good possibility of not being prosecuted at all. Consequently, the *expected* value of an offer can still be high enough such that the dictator is willing to accept it (especially given the fact that examples of the international community breaking its promise are rare). Freeman (2009: 29) echoes this view.

Anything, anything, to get that man out of the country and to have a safe transition in Syria. [...] If he wants to leave [...] that could be arranged. [...] Clearly we would like Assad to face justice for what he has done, but our priority, given the situation in that country, has to be an end to violence and a transition. And that cannot take place while Assad remains in place.³

Similarly, Qatar considered granting Assad asylum as "the region needs to stabilize".⁴

Next to the case of Assad, there are many more examples showing that the international community's commitment to ex-post punishment can still be broken in the post-ICC regime (legally, this is also possible as the Rome Statute of the ICC contains several "escape clauses", see footnote 13). During the Libyan Civil War, the British Secret Service MI6 reportedly arranged asylum for the ICC-indicted dictator Muammar Gaddafi in Equatorial Guinea (D'Ancona, 2013). That country does not recognize the ICC and was willing to host Gaddafi (who, in turn, had allegedly accepted the offer). But before they had a chance to execute the plan, Gaddafi was lynched by his own population in October 2011. Earlier (in 2004), the former Haitian president Jean-Bertrand Aristide accepted an asylum offer from South Africa, while Saudi Arabia struck an asylum deal with the former Tunisian president Ben Ali in 2011. Similarly, a local (US/EU/UNSC-backed) amnesty-abdication deal was made with Abdullah Saleh of Yemen in 2012. It thus seems to be the case that dictators are still willing to accept amnesty-abdication deals, while the international community is still willing to offer them (only less easily).

We show that dictators anticipating this may choose to commit more war crimes than they would actually like to from a static perspective, purely to worsen the situation and "unlock" the amnesty (or asylum) option - their only possible way out. They are essentially forcing the international community to become time-inconsistent. By behaving worse, dictators can thus make the *effective* punishment function non-monotonic - which is the essence of Reed Brody's quotation at the beginning of this paper. Importantly, it must be stressed that this channel exists over and above any adverse consequences from a dictator digging in for survival if he believes that his exit route is blocked. To the best of our knowledge, the phenomenon of a dictator worsening his behavior with the aim of re-writing his own punishment function, is new to the literature.

As a result of this effect, increasing the costs of time-inconsistency may induce some dictators to commit *more* war crimes, as that becomes necessary to make the amnesty-option available to them. This suggests that increased commitment to ex-post punishment

³See "Assad's safe exit 'could be arranged'" by *Reuters*, November 6 2012.

⁴See the "wiki-leaked" email from a daughter of the Qatari Emir to Assad's wife at <http://www.guardian.co.uk/world/2012/mar/14/bashar-al-assad-syria9>.

will lead to greater dispersion in outcomes: while there will be fewer malevolent dictators (by a deterrence effect), those who do end up in power, will behave worse.

Recent events seem consistent with this second prediction (more time is needed to judge the first one): the LRA-troops of Joseph Kony *worsened* their behavior after the ICC indicted him in 2005. Prior to this, there were peace talks and the LRA was rather inactive. But following the indictment, the LRA re-built troop numbers, while Kony vowed not to sign the 2008 peace treaty unless charges against him were dropped. Later that year, the LRA went on a major offensive and this unstable situation continues today.⁵

A similar development took place in Congo: soon after it emerged that the Congolese government was going to enforce the ICC indictment of Bosco Ntaganda, his rebel group attacked the city of Goma. According to a spokesman, they were not interested in control of the city, but all they wanted "is that the Congolese government sit down at the negotiating table".⁶ Both Kony and Ntaganda may have considered that their only possible way out was via an amnesty or asylum, but since an ICC indictment increases the international community's loss associated with granting one, additional violence is necessary to enable an exit via this route.

Analogous forces seem to have been at work in Zimbabwe. According to members of Robert Mugabe's government, his "top lieutenants are trying to force the political opposition into granting them amnesty for their past crimes by abducting, detaining and torturing opposition officials and activists"⁷ - a striking statement which, as we will see later on, is exactly in line with our model's prediction.

At the same time, proponents of criminal courts emphasize their ex-ante effects. When the ICC was established, Kofi Annan for example expressed his hope that the ICC would "deter future war criminals". In this respect, there is evidence that the amnesty given to Turkish officials responsible for the massacre of over one million Armenians during World War I led Adolf Hitler to conclude that he could pursue his genocidal policies with impunity (Scharf, 1999: 514): in a 1939 speech to his General Staff (who were concerned about accountability for acts of genocide), Hitler rhetorically asked: "Who after all is today speaking about the destruction of the Armenians?". This shows that the practice of granting amnesty may indeed bring moral hazard.

In this paper, we construct a model with which we analyze this trade-off. The remainder of this paper is structured as follows: first, Section 2 discusses the related literature. In Section 3, we introduce a model that captures many of the aforementioned elements.

⁵See e.g. <http://www.bbc.co.uk/news/world-africa-17299084>.

⁶See <http://www.bbc.co.uk/news/world-africa-18821962>.

⁷See "Mugabe Aides Said to Use Violence to Get Amnesty" in *The New York Times* of April 9, 2009.

The implications of our model are analyzed in Section 4 and used to derive policy prescriptions. Section 5 subsequently discusses the applicability of our model to other settings, as well as directions for possible future work. Finally, Section 6 concludes.

2 Related literature

The paper builds upon several literatures. First, it draws on the work on time-consistency, pioneered by Kydland and Prescott (1977). Following that paper, most applications have focused on the time-consistency of monetary policy (see Barro and Gordon (1983)) or tax policy (*cf.* Fischer (1980)). In this paper we apply similar ideas to the treatment of dictators who commit war crimes. Related work by Anderlini, Felli and Riboni (forthcoming) focuses on time-consistency problems in court rulings in general. They point out that "stare decisis" (the common law principle which requires judges to respect precedents established by earlier decisions) can alleviate the time-consistency problem. The present paper, in contrast, develops and analyzes a model that adds a new dimension to the standard time-consistency problem, namely that of strategic behavior from regulatees (as they may actively try to force regulators to give up on their earlier commitments and become time-inconsistent instead). Hereby, our paper also contributes to our understanding of the use of blackmailing and threats. As we will argue in Section 5, this extension may be of independent interest.

Second, we build upon deterrence theory and the theory of the public enforcement of law, as developed after Becker (1968). In particular, we will follow this literature by assuming that war criminals are reasoning creatures and respond to incentives (Schelling (1966) also takes this approach, as do Acemoglu and Robinson (2006) and Esteban, Morelli and Rohner (2012) more recently). Adolf Hitler's aforementioned rhetorical question on the Armenian genocide, suggests that they indeed do so. As a special case of the now-familiar "Beckerian criminals", we thus analyze "Beckerian dictators" who are considering the magnitude and number of war crimes to commit. In sharp contrast to earlier papers in the crime-literature, we will show that the level of criminal activity may actually *go up* in response to an increase in expected punishment.

Our paper also extends the literature on the economics of conflict (see Garfinkel and Skaperdas (2007) for an overview). In that literature, parties use resources in order to increase their probability of winning a violent conflict against a rival (with the winner obtaining a certain "prize"). In our setup however, dictators not only have an incentive to commit war crimes against others, but also to engage in *additional* violence as that

may "unlock" a previously unavailable amnesty escape route for them.

Finally, this paper relates to recent theoretical work on conflict and institutions. Examples include Aslaksen and Torvik (2006), who analyze how resource rents affect the political choice between conflict and democracy; Acemoglu and Robinson (2006, 2008), who develop a framework for analyzing the creation and consolidation of democracy; Besley and Persson (2011ab), who investigate incentives for states to invest in either state capacity or violence; and Guimaraes and Sheedy (2012), who analyze institutions resulting from the process of payoff maximization by an elite group that holds power, subject to the threat of removal by a group of rebels.

In the present paper, we focus on the impact of future punishment on the conduct of rulers. In addressing this issue, we abstract from the question whether these rulers were put in place by democratic elections or not (see Acemoglu and Robinson (2006, 2008) and Acemoglu, Robinson and Torvik (2013) for models that can explain whether a country will develop into a stable democracy). Instead, we focus on the *behavior* of rulers and investigate how legislative institutions affect their incentives to "behave". Given that many democratically elected leaders subsequently turned into malevolent dictators (Adolf Hitler, Robert Mugabe, Charles Taylor), this question seems interesting in its own right.

3 Model

This section presents our model. We start by describing the environment and solving the model under perfect commitment, after which we turn to the crucial issue of time-(in)consistency. As the model is most easily explained backwards, we end with the entry decision for potential dictators at the start of the game.

3.1 Model environment and solution under perfect commitment

Our model consists of two periods, $t = 1, 2$, and its players are the international community and a pool of potential dictators (one of whom is going to enter office in the country under consideration at the beginning of $t = 1$). More broadly, one could also see this paper's "dictator" as any leader who is able to commit war crimes against civilians (for example due to the backing of an army; cf. Joseph Kony and other "warlords").

The goal of the international community is to minimize the total amount of war crimes that a dictator will commit, for example because it directly cares about repressed civilians or because it benefits from stability in a particular region. One could also incorporate a time-discount factor if one considers that appropriate. Using x_t to denote the intensity

of war crimes committed by a dictator in period t (which can take the form of murder, torture, using child soldiers, looting, rape, etc.), the international community thus aims to minimize the expected value of:

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

It does so by threatening to punish dictators at the end of period 2 for any war crimes that they have committed during periods 1 and 2. As it takes time to oust a dictator and bring him to court, we assume that the international community is not able to remove and punish the dictator involuntarily before the end of period 2 (if at all).⁸ Consequently, an incumbent dictator has an intermediate phase available during which he is able to play a game with the international community. 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 international community 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. One can think of ϕ as combining the *actual punishment* conditional on the dictator being tried and convicted (call that ψ), with the *probability* that a dictator will be caught and brought to court (call that π) - so ϕ can be seen as an expected value.⁹ Note that π can be smaller than 1, so we allow for the possibility that the international community is not able to catch the dictator before he leaves the stage. For the main part of our argument, it suffices to work with the product of these two parameters (the expected punishment), defined as:

$$\phi \equiv \pi\psi \tag{2}$$

We assume that it is not feasible to set ϕ prohibitively high. As in all Beckerian models of criminal activity, letting $\phi \rightarrow \infty$ forms the first-best in which no crimes are

⁸It 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 5.

⁹Alternatively, $h(\cdot)$ can be interpreted as the local punishment function; π then represents the probability that the local population ousts the dictator. One could complicate the analysis by assuming that the probability of removal and prosecution π is a function of x_t . In that case, the dictator is able to affect the expected duration of his tenure by his choice for x_t (which for example includes repression). Whether a higher choice for x_t increases or decreases tenure is not obvious: while repressing civilians presumably decreases the success-probability of a local revolution, it also increases the probability of a foreign intervention. We therefore leave this extension for future work, as it would introduce many issues that are not central to the time-consistency problem studied in this paper. Also see the discussion in Section 5 on why such an extension is not of immediate relevance to the main message of this paper.

committed to begin with, but this solution is generally considered to be rather Utopian and practically infeasible (cf. Persson and Siven (2007); arguably, the maximum possible punishment will also be bounded if the dictator attaches only a finite value to his life).

Since many of the aforementioned leaders seem to have enjoyed direct utility from committing war crimes (think about the gratification that Adolf Hitler derived from killings Jews, Saddam Hussein from killing Kurds, Joseph Kony from abusing sex slaves, and so forth), we assume that our dictator derives direct benefits from implementing the policy x_t .¹⁰ Alternatively, one could see this as a way of describing a dictator who derives security-services from exercising repression x_t , for example because he enjoys rent extraction which would lead to a revolution by an angered crowd in the absence of repression (rent extraction and repression tend to go hand-in-hand as the former creates a need for the latter; see Larcom, Sarr and Willems (2014) who develop a model along these lines).

Engaging in war crimes is costly to a dictator (for example because he needs to finance an army) and we assume that the dictator's cost function of committing x_t war crimes is given by $c(x_t) = \gamma x_t$. The value of $\gamma > 0$ is known to all agents in the model.

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

$$\max_{x_1, x_2} \theta \sum_{t=1}^2 \log(x_t) - \gamma \sum_{t=1}^2 x_t - \phi \sum_{t=1}^2 x_t \quad (3)$$

Here, $\theta \geq 0$ is a time-invariant parameter capturing the dictator's type: a dictator with a high θ derives a lot of satisfaction from committing war crimes relative to punishment (think of Adolf Hitler), while this holds less so for a dictator with a low θ (think of Lee Kuan Yew). In this sense, one could interpret θ as the intrinsic malevolence of the dictator, or as relating to the dictator's degree of impatience (as he derives utility from committing war crimes in periods 1 and 2, while punishment can only follow at the end of period 2).¹¹ A crucial part of the model is that the value of θ is private information to the dictator. The international community can only form beliefs about it by observing the dictator's behavior in period 1. We assume that θ is distributed according to some known c.d.f. $F(\theta)$ in the pool of potential dictators (but more on this below).

¹⁰For example, one of Hitler's co-organizers of the Holocaust, Adolf Eichmann, has stated that "I will leap into my grave laughing because the feeling that I have five million human beings on my conscience is a source of extraordinary satisfaction for me." Also see Coolidge and Segal (2009) for psychological evidence that many dictators and war criminals indeed suffered from a sadistic personality disorder. Idi Amin's brutal behavior is often explained along similar lines (Nayenga, 1979: 131).

¹¹For this reason, we abstract from discounting in equation (3) for notational simplicity, as a discount factor would fulfil a similar role as θ without delivering any non-trivial insights.

From solving the optimization problem (3), we can see that the first-order condition (henceforth referred to as "FOC") implies that:

$$x_t^* = \frac{\theta}{\gamma + \phi} \text{ for } t = 1, 2 \quad (4)$$

So at the beginning of period 2 (after observing x_1) the international community is able to make an inference about the dictator's type θ by using (4), leading to:

$$\mathbb{E}_2^{IC} \{\theta|x_1\} = [\gamma + \phi] x_1, \quad (5)$$

where γ , ϕ , and x_1 are all known at this stage.¹²

3.2 Partial commitment and time-inconsistency

So far, we have maintained the strong assumption that the international community is ex ante able to commit itself to punishing the dictator 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 international community is able to increase the costs of being time-inconsistent (for example by installing criminal courts and indicting individuals, as granting amnesties to indicted individuals brings a significant reputational loss), it is unlikely to be able to increase these costs all the way up to infinity (which is necessary for perfect commitment).

Instead, let us suppose that the international community incurs a loss equal to a *finite* $\Lambda \geq 0$ if it is time-inconsistent.¹³ In our model, time-inconsistency implies forgetting about the punishment function $h(\cdot)$ at the beginning of period 2 and offering the dictator amnesty (or asylum or only a minor punishment, but for the sake of brevity we will just talk about amnesty here), in return for his abdication. Since involuntary removal of the dictator can only take place at the *end* of period 2 (recall footnote 8), our model captures the idea that an amnesty-abdication deal is able to shorten the dictator's tenure.

Λ can encompass many different costs: some are institution-related (such as the loss

¹²We take the dictator's cost parameter γ to be public knowledge, but one could also assume that this information is private to the dictator without affecting our results (γ then just takes over the role that θ plays in the current setup).

¹³Alternatively, one can interpret this as the international community adhering to a rule that has an explicit escape clause. As shown by Flood and Isard (1989), Persson and Tabellini (1990), and 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.

of credibility for the international community if it previously established institutions, like criminal courts, that were supposed to try these dictators), while others would occur in any environment (such as the moral hazard for future dictators). We will assume that:

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

Our assumption that $d\Lambda(\phi)/d\phi > 0$ captures the intuitive notion that if the international community announces ex ante that it is going to prosecute misbehaving dictators with higher probability (i.e.: increasing π), then the (reputational) loss associated with subsequently *not* doing so will be higher (remember that $\phi \equiv \pi\psi$ by (2)). Note that this is consistent with the idea that by installing a permanent ICC (increasing π), the international community has raised Λ and thereby its commitment to the ex-post punishment of malicious rulers: if it were to strike an amnesty-abdication deal with a suspected war criminal nevertheless, it would inflict a significant reputational loss upon the ICC.

As we show in the Appendix, 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 grant amnesties to dictators who acted worse in period 1. This allows for the idea that there might be more public outcry if amnesty is granted to a dictator 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 dictator with high ϕx_1).

Given a pre-specified punishment parameter ϕ , the timing in our model is as follows:

1. At the beginning of period 1, all potential dictators learn their type θ and decide whether they want to run for office. One of the running candidates ends up as the country's dictator (this entry-step will be explained in Section 3.3 below).
2. Given his type θ , the dictator-in-office commits war crimes x_1 during period 1.
3. At the beginning of period 2, the international community forms an expectation of the dictator's type, based upon his first-period behavior x_1 . It subsequently decides whether to make the dictator an amnesty-abdication offer or not.
4. If an amnesty is offered and accepted, $x_2 = 0$ and the dictator obtains the value of the amnesty offer. If no amnesty-abdication deal is made, the dictator commits war crimes x_2^* in period 2.

5. (Only if no amnesty-abdication deal is made at Stage 4:) If the dictator is caught before the end of period 2 (which happens with probability π), he is punished for any crimes that he has committed in the past. Otherwise, he is able to leave the stage unpunished.

The solution-concept is Perfect Bayesian Equilibrium. Denoting the international community's action at the beginning of period 2 by $s \in \{\textit{offer amnesty}, \textit{punish}\}$, 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 international community chooses an action $s^{eq}(x_1)$ that minimizes its expected loss function $\mathbb{E}_2^{IC} \{\mathcal{L}|x_1\}$.*
2. *The beliefs $\mu(\theta|x_1)$ held by the international community about the dictator's type θ are obtained through Bayes' rule, where possible.*
3. *Taking the international community's best response $s^{eq}(x_1)$ into account, a dictator of type θ decides to commit war crimes $x_t^{eq}(\theta)$ that maximizes his utility.*

We denote the value of the amnesty offer to the dictator by V_A . It consists of the value that the dictator attaches to not being tried (or receiving only a minor punishment), as well as of any potential additional benefits (recall Idi Amin, who managed to negotiate spending his last years in a nice hotel). In order to induce the dictator to accept the offer and step down, the international community will have to set $V_A > \theta \log(x_2^*) - \gamma x_2^* - \phi[x_1 + x_2^*]$. The RHS of this inequality represents the value of the dictator's outside option of staying in power. Note that the exact location of this "threat point" is private information to the dictator, because θ is. Due to the resulting informational advantage of the dictator over the international community, the dictator is able to exploit this edge when negotiating on V_A (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 dictator is able to capture rents $\Delta > 0$ in the amnesty-abdication negotiations, such as a nice place to spend his post-dictatorship years, or access to his foreign assets. Consequently, V_A takes the following "outside-option plus surplus"-form:

$$V_A(\theta, \phi) = \theta \log(x_2^*) - \gamma x_2^* - \phi[x_1 + x_2^*] + \Delta \tag{6}$$

Because the international community's promise of impunity post-abdication might not be fully credible either (cf. footnote 2), one may see V_A as an *expected* value to the dictator (and note that the model's solution is fully determined by expected values where applicable). As evidenced by the long list of dictators who have accepted amnesty/asylum offers in the past (see footnote 1), the commitment problems associated with these deals do not seem to be prohibitively large - probably because there are only very few examples of amnesty/asylum offers being revoked ex post.¹⁴ As a result, the international community is apparently able to offer enough "expected value" to induce dictators to step down. The recent examples of Aristide, Ben Ali, and Saleh (who all accepted amnesty/asylum offers made after 2002) confirm that this continues to be the case in the post-ICC regime.

In equation (6), the size of the bargaining-surplus Δ may depend upon various factors, such as the dictator's negotiation skills relative to those of the international community and the urgency of the situation for civilians. We leave these components unmodelled and set $\partial\Delta/\partial\mathbb{E}_2^{IC}\{\theta|x_1\} = 0$. This implies that any strategic behavior that a dictator is going to display, is not going to arise with the aim of increasing negotiation-rents Δ - only with the aim of unlocking the amnesty-option in the first place. Judging from the examples of Kony, Mugabe and Ntaganda (given in the Introduction), this seems to be the more important effect in reality.¹⁵

When a malevolent dictator abdicates, the international community believes that it is able to replace him by a leader of a better type; if it did not hold this belief, then it would have no incentive to remove the incumbent (along these lines, it has for example been argued that a fear of "who comes next" is one of the reasons for why Assad is still in power in Syria). We assume that the international community holds the optimistic belief that, following an amnesty-abdication deal, it can select and install a benevolent dictator with $\theta' = 0$, such that it is able to "trade justice for peace" - letting x_2 collapse to zero. The

¹⁴For example: in South Africa, the ANC-government has continued to respect the local amnesties granted to those responsible for crimes under apartheid (despite the fact that many of the accused are still living there and could easily be arrested). Why governments are seemingly more able to stick to amnesty-promises than to prosecution-promises is an interesting question, but it is of subordinate importance for the issue under consideration here. All that matters for our analysis is that the international community is able to convince dictators to step down in return for a *promise* of amnesty. The recent amnesty-abdication deals mentioned in the main text are evidence that the international community is indeed still able to do this. This suggests that dictators either believe that the international community will honor such promises, or that they greatly value the expected time in freedom that is on offer (even if they suspect the amnesty might be revoked at a later date). Whether the amnesty-promise is subsequently respected or not, does not matter for the mechanism underlying our paper.

¹⁵Assuming that $\partial\Delta/\partial\mathbb{E}_2^{IC}\{\theta|x_1\} > 0$ (capturing the idea that dictators will enter amnesty-abdication negotiations with more bargaining power if the international community believes them to be of a worse type), would only give a dictator an additional reason to engage in first-period violence - over and above the channel that will be explained below.

best example of this actually occurring in practice, is probably that of Liberia where peace was restored by replacing Charles Taylor with Nobel Peace Laureate Ellen Johnson Sirleaf. Whether this optimistic belief is justified or not, does not matter for the model's solution (again because the latter is fully determined by expected values);¹⁶ qualitatively similar results would be obtained in the more general case where the international community believes that it is able to replace the abdicating dictator by any better type $\theta' \in [0, \theta]$. Hence, after observing x_1 , the international community's expected loss function becomes:

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

This modified loss function expresses the idea that if the malevolent dictator is removed via an amnesty, x_2 is expected to collapse to zero (capturing the notion that an amnesty-abdication deal is the quickest way to end violence). But by making such a deal, the international community simultaneously incurs a loss of $\Lambda(\phi)$, due to a loss of credibility, or due to moral hazard for future dictators.

As before, the international community is only able to infer the incumbent dictator's type from his first-period behavior. Its beginning of period 2 expectation of the crimes that the dictator will commit during period 2 is subsequently given by:

$$\mathbb{E}_2^{IC} \{x_2|x_1\} = \frac{\mathbb{E}_2^{IC} \{\theta|x_1\}}{\gamma + \phi} \quad (7)$$

One can then see from the modified loss function (1') that the amnesty-option will become available to the dictator as soon as the international community believes that:

$$\mathbb{E}_2^{IC} \{x_2|x_1\} = \frac{\mathbb{E}_2^{IC} \{\theta|x_1\}}{\gamma + \phi} \geq \Lambda(\phi) \Leftrightarrow \mathbb{E}_2^{IC} \{\theta|x_1\} \geq (\gamma + \phi) \Lambda(\phi) \quad (8)$$

After all, when this condition holds, the dictator is believed to be of such a bad type θ , that the loss that the international community expects to incur if it sticks to its earlier intention to try the dictator, is larger than the loss associated with being time-inconsistent. In those cases, the international community (or an individual member of it, for example the country which has the greatest incentive to bring stability) prefers

¹⁶Moreover, any war crimes that might be committed by a successor can be captured through Λ . The model thus assumes that the amnesty offer is able to "trade justice for peace" in the short run, but that there may be future costs associated with such a deal - for example because of the moral hazard induced by the time-inconsistency of the international community.

to end violence immediately by making an amnesty-abdication deal. Here, the *effective* punishment function becomes non-monotonic. This is consistent with the pattern behind Reed Brody's remark at the beginning of this paper (of which David Cameron's statement in the Introduction is a clear recent example).

At this stage, one can define the critical type $\hat{\theta}$, whose FOC (4) dictates setting $x_1 = \hat{x}_1 \equiv \hat{\theta}/(\gamma + \phi)$, such that the international community is indifferent between amnesty and punishment. It satisfies:

$$\mathbb{E}_2^{IC} \left\{ \theta \left| x_1 = \frac{\hat{\theta}}{\gamma + \phi} \right. \right\} = (\gamma + \phi) \Lambda(\phi) \quad (9)$$

Henceforth, we will refer to those dictators with $\theta \geq \hat{\theta}$, as the "evil" ones: by adhering to their FOC, they already commit so many war crimes that the international community is better off by offering them amnesty in return for their abdication.

For a dictator of type $\theta < \hat{\theta}$, unlocking the amnesty-option calls for setting $x_1 > x_1^*$ (the latter being the statically optimal choice given by the FOC (4)). In those cases, a reasoning, "Beckerian" dictator who recognizes the finiteness of the international community's loss associated with being time-inconsistent, may choose to make an "investment" in period 1. The dictator "invests" by committing more war crimes than he would actually like to from a static perspective - purely for the sake of unlocking the amnesty-option, so that he can in the end walk away with no (or only a minor) punishment.

In this case, the dictator uses period 1 to try and signal that he is of the evil type (i.e.: that he is of a type $\theta \geq \hat{\theta}$) thereby threatening that he will commit many war crimes in period 2 as well. So many, that the international community is actually better off by making an amnesty-abdication deal at the beginning of period 2.¹⁷ The dictator 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 amnesty. With reference to the behavior of Bosco Ntaganda (see footnote 17), we will call this the "Terminator effect" (although "Mugabe effect" would be another fitting name, as his "top lieutenants are trying to force

¹⁷This 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 6). Relatedly, Schelling (1966: v) writes that "the power to hurt (...) is a kind of bargaining power, not easy to use but used often". ICC-investigator Matthew Brubacher compares this behavior of dictators and warlords to "blackmailing" in the 2012 documentary *Peace vs. Justice*. 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.'"

the political opposition into granting them amnesty for their past crimes by abducting, detaining and torturing opposition officials and activists"; recall footnote 7).

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

- \mathcal{F} . Follow the FOCs given by (4), after which he expects punishment at the end of period 2. Since π can be smaller than 1, we also allow for the possibility that the dictator manages to escape from punishment (but the *expected* penalty $\phi \equiv \pi\psi$ is all that matters here).
- \mathcal{D} . Deviate from the FOC (4) in period 1, to unlock the amnesty offer at the beginning of period 2 (which results in a final pay-off equal to V_A).

What is the utility that a type θ -dictator derives from these options? Let us first analyze the value from following strategy \mathcal{F} , henceforth indicated by $U^{\mathcal{F}}$. Since the dictator chooses to set $x_1 = x_2 = \theta/[\gamma + \phi]$ by FOCs (4), it follows that:

$$U^{\mathcal{F}} = 2\theta \log \left(\frac{\theta}{\gamma + \phi} \right) - 2\theta \quad (10)$$

Given that we focus on dictators with $\theta < \hat{\theta}$, setting x_1 according to the FOC (4) does not unlock the amnesty-option (because $x_1^* < \hat{x}_1$ for those types, with \hat{x}_1 being the critical level of first-period crimes that unlocks the amnesty-option).

The cheapest way for such a dictator to unlock the amnesty-option nevertheless, would be to mimic a type $\hat{\theta}$ -dictator by setting $x_1 = \hat{x}_1$ (this is strategy \mathcal{D}). After all, since $\hat{x}_1 > x_1^*$, the marginal benefits from committing war crimes are lower than its marginal costs at this point, as a result of which it will never be optimal for a dictator to go beyond that level of crime (setting $x_1 = \hat{x}_1$ suffices to unlock the amnesty-option, so setting $x_1 > \hat{x}_1$ only brings additional net costs). Following this strategy would yield a dictator of type $\theta < \hat{\theta}$ lifetime utility:

$$U^{\mathcal{D}} = \theta \log (\hat{x}_1) - \gamma \hat{x}_1 + V_A(\theta, \phi) \quad (11)$$

Hence, a dictator of type $\theta < \hat{\theta}$ will "invest" in committing war crimes (and pool with

the "evil" types) in period 1 iff his θ is such that $U^D \geq U^F$, i.e. iff:

$$y(\theta, \phi) \equiv \theta \log(\hat{x}_1) - \gamma \hat{x}_1 + V_A(\theta, \phi) - 2\theta \log\left(\frac{\theta}{\gamma + \phi}\right) + 2\theta \geq 0 \quad (12)$$

When $y(\theta, \phi) = 0$, it implicitly defines the critical type $\bar{\theta}$ beyond which a dictator 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 $V_A(\bar{\theta}, \phi)$. We then know from combining equations (6) and (12) that $\bar{\theta}$ should satisfy:

$$y(\bar{\theta}, \phi) = \bar{\theta} \left[\log\left(\frac{\hat{\theta}}{\bar{\theta}}\right) + 1 \right] - \hat{\theta} + \Delta = 0 \quad (13)$$

Assuming that $\Delta < \hat{\theta}$,¹⁸ there exists a $\bar{\theta} < \hat{\theta}$ for which $y(\bar{\theta}, \phi) = 0$ (because $y(\cdot)$ is then a continuous function with $y(0, \phi) < 0$, $y(\hat{\theta}, \phi) > 0$, and $\partial y(\theta, \phi)/\partial \theta|_{\theta=\bar{\theta}} > 0$). Consequently, all types $\theta \in [\bar{\theta}, \hat{\theta})$ will find it optimal to mimic the $\hat{\theta}$ -type by setting $x_1 = \hat{x}_1$. The reason is that in the face of a time-consistency 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 international community forms its expectation about the dictator's type θ in Perfect Bayesian Equilibrium after observing the dictator's choice of x_1 . It is given by:¹⁹

$$\mathbb{E}_2^{IC} \{\theta | x_1\} = \begin{cases} \mathbb{E} \left\{ \theta | \bar{\theta} \leq \theta \leq \hat{\theta} \right\} = \int_{\bar{\theta}}^{\hat{\theta}} \frac{\theta f(\theta) d\theta}{F(\hat{\theta}) - F(\bar{\theta})} & \text{when } x_1 = \hat{x}_1 \\ [\gamma + \phi] x_1 & \text{otherwise} \end{cases} \quad (14)$$

The intuition is that when the international community observes $\hat{x}_1 \equiv \hat{\theta}/[\gamma + \phi]$ (that level of crime which is just enough to unlock the amnesty-option), the naive expectation would be to assume that the responsible dictator is exactly of type $\hat{\theta}$ (this would follow from blindly applying equation (5)). In equilibrium, however, the international community realizes that \hat{x}_1 could also be set by a dictator of type $\theta \in [\bar{\theta}, \hat{\theta})$, who is trying to mimic a dictator of type $\hat{\theta}$ so as to unlock the amnesty-option (the "Terminator effect"). Consequently, the rational expectation is $\mathbb{E} \left\{ \theta | \bar{\theta} \leq \theta \leq \hat{\theta} \right\} < \hat{\theta}$ implying that the international community "discounts" the dictator's first-period behavior somewhat as it realizes that he might just be mimicking an evil dictator, without actually being one himself.

¹⁸Otherwise $y(\theta, \phi) > 0 \forall \theta$ 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.

¹⁹We thank Avinash Dixit for pointing out an error in equation (14) in an earlier draft.

To close the model, we still need to specify how the international community would respond if an out-of-equilibrium action (call that \tilde{x}) were to be observed - otherwise there is a plethora of equilibria. We impose that $\mu(\theta|\tilde{x}) = [\gamma + \phi]\tilde{x} \equiv \tilde{\theta}$. This implies that any out-of-equilibrium action is interpreted as coming from a dictator-type whose FOC (4) prescribes that action. Next to this belief seeming reasonable if one assumes that the international community allows for the possibility that some dictators may act non-strategically (dogmatically adhering to their FOC instead), it also survives the D1 refinement-arguments in Cho and Kreps (1987).²⁰

At this point, we know that $\bar{\theta}$ should solve (13), while $\hat{\theta}$ solves equation (9) (with $\mathbb{E}_2^{IC}\{\theta|x_1\}$ being formed as in (14)). Assuming for concreteness that $\theta \sim U(0, \theta_{\max}]$, we have $\mathbb{E}_2^{IC}\{\theta|\hat{x}_1\} = (\bar{\theta} + \hat{\theta})/2$ and condition (9) implies:

$$\hat{\theta} = 2(\gamma + \phi)\Lambda(\phi) - \bar{\theta} \quad (15)$$

Substituting this into (13) gives the following implicit equation characterizing $\bar{\theta}$:

$$y(\bar{\theta}, \phi) = \bar{\theta} \log \left(\frac{2(\gamma + \phi)\Lambda(\phi) - \bar{\theta}}{\bar{\theta}} \right) + 2\bar{\theta} - 2(\gamma + \phi)\Lambda(\phi) + \Delta = 0 \quad (16)$$

This completes the solution for this part of the model.

3.3 Dictator entry

What remains is a description of dictator entry. At the beginning of period 1, there are two pools of potential dictators: pool \mathcal{M} (where all dictators are malevolent, i.e. of type $\theta > 0$) and pool \mathcal{B} (where all dictators are benevolent, i.e. of type $\theta = 0$). We normalize the measure of the malevolent pool to 1, while the benevolent pool has measure b .²¹ Running for dictatorship requires the payment of a fixed (and sunk) entry cost $\mathcal{E} > 0$, after which the agent will be successful, and end up as the country's dictator, with probability p (which will depend negatively upon the measure of competing entrants).

²⁰See 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 the action of the receiver in our game (the international community) is binary: it offers amnesty, or it sticks to ex-post punishment. This implies that $\hat{\theta}$ -types have no incentive to try and separate themselves from lower, mimicking types.

²¹We introduce the mass point at $\theta = 0$ to be able to capture the deterrence effect in equation (18) below, but this could be done in other ways. One could for example also assume that the country becomes a stable democracy with a probability that is decreasing in the mass of malevolent dictators that is competing for office.

We assume that benevolent dictators are always willing to run for office. Hence, these dictators have a strong sense of duty and "ask not what their country can do for them, but they ask what they can do for their country".

Agents in pool \mathcal{M} on the other hand, are malevolent and only interested in maximizing their own private pay-off. The distribution of θ in pool \mathcal{M} has c.d.f. $F(\theta)$, with support $(0, \theta_{\max}]$. We use $f(\theta)$ to denote the associated p.d.f. Agents in this pool first get to observe their private draw of θ , after which they have to decide whether they find it worthwhile to run for dictatorship (which has success-probability p).

Now consider a dictator with $\theta < \bar{\theta}$. By definition of $\bar{\theta}$, he does not want to unlock the amnesty-option by mimicking an evil dictator, as doing so is too costly for him given his type. Instead, he will follow his FOCs, which would bring him lifetime utility as in (10). Consequently, a type θ agent in the potential pool of malevolent dictators would only want to run for dictatorship if his θ is such that:

$$z(\theta, \phi) \equiv p(\underline{\theta}) \left[2\theta \log \left(\frac{\theta}{\gamma + \phi} \right) - 2\theta \right] - \mathcal{E} \geq 0 \quad (17)$$

When $z(\theta, \phi) = 0$, it defines the critical level $\underline{\theta}$ below which potential malevolent dictators do not find it worthwhile to run for office (where we assume \mathcal{E} to be low enough such that $\underline{\theta} < \bar{\theta}$, otherwise all entering malevolent dictators will again choose to mimic the evil ones. As noted before, such a specification would strengthen the results that are to follow, but strikes us as being rather extreme). The function $p(\underline{\theta})$ (with $dp(\underline{\theta})/d\underline{\theta} > 0$) captures the idea that any given individual running for dictatorship has a higher probability of ending up in office, if the number of competitors is relatively low (which is the case when the entry threshold $\underline{\theta}$ is high).

We furthermore impose that at least some malevolent dictators are willing to run for office (as this seems to be the case in reality, while it is also 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 dictators of type $\theta < \underline{\theta}$ are deterred by the severe punishment for war crimes, as a result of which it is not worthwhile for them to pay the entrance fee \mathcal{E} to try and become a dictator (given that their low value for θ implies that they can derive only little utility from committing war crimes). Consequently, only a fraction $[1 - F(\underline{\theta})]$ out of the pool of potential malevolent dictators will decide to run for dictatorship.

After defining $\mathbf{1}_{\mathcal{B}}$ as an indicator function that takes the value 1 if a benevolent dictator ends up in office in the country under consideration (and zero otherwise), one can express

the ex-ante probability of this occurring as:

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

with:

$$\frac{\partial \Pr[\mathbf{1}_{\mathcal{B}} = 1]}{\partial \phi} = \frac{b \cdot \partial F(\underline{\theta}) / \partial \underline{\theta} \cdot \partial \underline{\theta} / \partial \phi}{[b + 1 - F(\underline{\theta})]^2} > 0, \quad (18)$$

since $\partial \underline{\theta} / \partial \phi > 0$ (see Proposition 1 below). This shows that there is a deterrence effect: the more serious the international community appears to be ex ante about prosecuting malevolent dictators, the less willing malevolent dictators are going to be to run for office. Benevolent dictators on the other hand are unaffected by this threat, as they are not planning to commit war crimes given that they can't derive any utility from doing so. Consequently, the probability that a country ends up with a malevolent ruler decreases with the punishment parameter ϕ , which is intuitive.

However, conditional on ending up with a malevolent dictator, a higher choice for ϕ also implies that that dictator will on average be of a worse type since $\mathbb{E}\{\theta | \mathbf{1}_{\mathcal{B}} = 0\} = \mathbb{E}\{\theta | \theta \geq \underline{\theta}\}$, for which it holds that:

$$\frac{\partial \mathbb{E}\{\theta | \theta \geq \underline{\theta}\}}{\partial \phi} > 0, \quad (19)$$

again because $\partial \underline{\theta} / \partial \phi > 0$. The reason is that the deterrence effect only drives out potential malevolent dictators who are actually not that bad (the ones with a relatively low θ). Consequently, the benign deterrence effect is accompanied by a malign adverse selection effect - more on which in Section 4. Which of these two effects dominates, cannot be established unambiguously as this depends on the distribution of θ .

3.4 Summary

Eventually, our model can be summarized by noting that it leads to four groups of malevolent dictators:

- The "not-so-bad dictators" with $0 < \theta < \underline{\theta}$. They choose not to enter because of the entry fee \mathcal{E} , as their low θ implies that they are not able to derive enough utility from committing war crimes to make up for that fee.
- The "bad dictators" with $\underline{\theta} \leq \theta < \bar{\theta}$. This group chooses to enter, but for them it is too costly to change their behavior in order to unlock the amnesty-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 dictators are not able to mimic the behavior of the evil types, for whom $\theta \geq \hat{\theta}$.

- The "mimicking dictators" with $\bar{\theta} \leq \theta < \hat{\theta}$. This is an interesting group, as they choose to modify their first-period behavior by "investing" in committing first-period war crimes, for the sole purpose of unlocking the amnesty-option at the beginning of period 2 (the "Terminator effect"). In signaling terms, they are able to mimic and form a pooling equilibrium with the evil dictators. These types want to actively force the international community to become time-inconsistent.
- The "evil dictators" with $\theta \geq \hat{\theta}$. For them, following the FOCs already suffices to unlock the amnesty-option – no additional investments are necessary. Note that dictators 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, while they also satisfy the ordering $\underline{\theta} < \bar{\theta} < \hat{\theta}$ - as discussed in Sections 3.2 and 3.3.

4 Analysis

We are now able to analyze what would happen if the international community raises the expected punishment parameter ϕ - as it has done in reality by establishing the ICC, thereby raising the probability of prosecution π (while subsequently using the ICC to indict someone, increases π (and hence ϕ) for that particular individual even further).

To answer this question, it is crucial to analyze the comparative statics of the various thresholds with 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 international community is boundedly rational and naively applies (5) (rather than (14)) when forming $\mathbb{E}_2^{IC} \{\theta|x_1\}$. Reassuringly, our findings are robust to that specification as well.

Proposition 1 *Provided that some potential malevolent dictators are willing to run for dictatorship, $\partial \underline{\theta} / \partial \phi > 0$.*

Proof. The threshold $\underline{\theta}$ is implicitly defined by $z(\underline{\theta}, \phi) = 0$ (with z following the specification of equation (17)). 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 (17) shows that malevolent dictators with $\theta > \underline{\theta}$ will only be willing to run for dictatorship when $\gamma + \phi < \underline{\theta}$. Recalling that $0 < p(\underline{\theta}) < 1$ and $dp(\underline{\theta})/d\underline{\theta} > 0$, this immediately implies that $\partial \underline{\theta}/\partial \phi > 0$. ■

Proposition 1 states that an increase in ϕ raises the threshold beyond which agents in the pool of potential malevolent dictators become willing to run for dictatorship (imposing that at least some malevolent dictators are running for office, i.e. $\underline{\theta} > \gamma + \phi$, which is the whole premise of this paper and seems to be the case in reality).²² Consequently, the higher the international community's choice for ϕ , the lower the fraction $[1 - F(\underline{\theta})]$ out of the pool of potential malevolent dictators that will decide to run for dictatorship. This is the deterrence effect captured by equation (18). Simultaneously, however, Proposition 1 also implies that conditional on a malevolent dictator taking office, that dictator will on average be of a worse type (recall the adverse selection effect underlying equation (19)).

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

Proof. The threshold $\bar{\theta}$ is implicitly defined by $y(\bar{\theta}, \phi) = 0$. Differentiating y (specified in equation (16)) with respect to its two arguments, and using equation (15) gives:

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

since $\bar{\theta} < \hat{\bar{\theta}}$. By the same arguments,

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

²²Since the last term in the denominator of $\partial \underline{\theta}/\partial \phi$ is positive, this condition is actually stricter than necessary. But given that this paper starts from the observation that some malevolent dictators are willing to run for office, we maintain it nevertheless as it maximizes both clarity as well as descriptive realism without serious loss of generality.

Consequently, the implicit function theorem implies:

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

■

This proposition tells us that an increase in ϕ , raises the threshold beyond which dictators become willing to "invest" in committing war crimes and form a pooling equilibrium with the evil types ($\theta > \hat{\theta}$).

With respect to the threshold level for these evil types, it holds that:

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

Proof. Differentiating (15) gives:

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

since $\bar{\theta} < \hat{\theta}$. ■

This states that the location of the threshold type $\hat{\theta}$ beyond which it becomes optimal for the international community to offer the dictator amnesty, is increasing in ϕ - which is intuitive given that the costs of being time-inconsistent are increasing in ϕ (by Assumption 1, $d\Lambda(\phi)/d\phi > 0$). It implies that when ϕ is higher, dictators of type $\theta < \hat{\theta}$ will have to put in a greater mimicking effort (and commit more crimes) if they want to unlock the amnesty-option.

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

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

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

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

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

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

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 $W(-1/q \exp(-1/q)) = -1/q$, while it also holds that $W(-1/\exp(1)) = -1$. It thus follows that $\log(q) + 1/q - 1 > 0$ for:

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

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

This means that the size of the mimicking group, which is subject to the "Terminator effect" and chooses to worsen its behavior to unlock the amnesty-option, is increasing in ϕ . The reason is that when prospective punishment ϕ is higher, exit via the amnesty escape route becomes relatively more valuable. Consequently, more dictator-types become willing to pool with the evil types at $\widehat{\theta}$, so as to unlock the amnesty offer.

The various threshold effects captured by Propositions 1-4 can be represented as in Figure 1 (which assumes that $f(\theta)$ is the uniform p.d.f.).

In this figure, the dashed black line represents the true (uniform) distribution of θ , while the solid red line represents θ 's "effective" distribution (after entry and mimicking incentives have been taken into account). Note how the "Terminator effect" produces a hole in the distribution on the $[\bar{\theta}, \widehat{\theta})$ -interval (plus an atom at $\widehat{\theta}$). The various arrows illustrate the effects brought about by an increase in ϕ (where the numbers refer to the

Propositions that they represent). The relative sizes of the arrows at 2 and 3 indicate that $\partial\hat{\theta}/\partial\phi > \partial\bar{\theta}/\partial\phi$, which implies that the mimicking group grows in size when ϕ increases (this explains why arrow 4 points up).

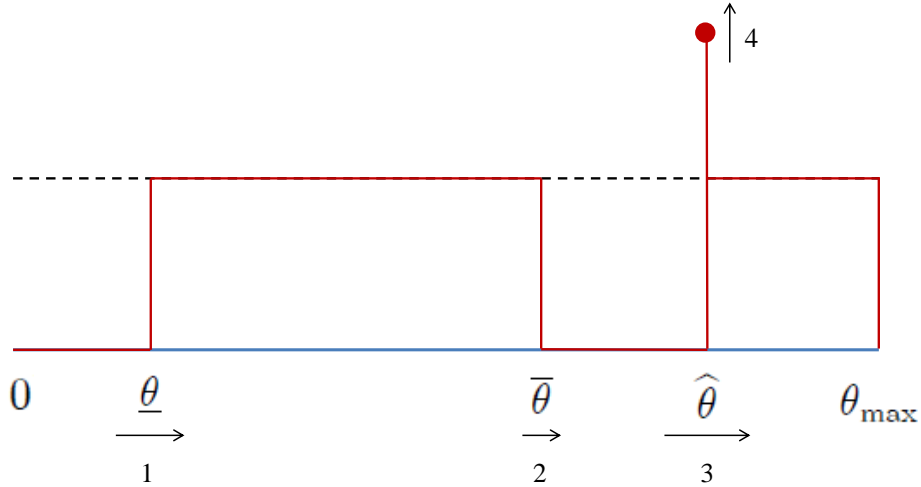


Figure 1: Graphical illustration of the setup when θ follows a uniform distribution

Finally, there is also a "disciplining effect" for those dictators with $\theta \in [\underline{\theta}, \bar{\theta})$ or $\theta \in [\hat{\theta}, \theta_{\max}]$. They just choose to adhere to their FOCs (given by (4)). As shown in the following proposition, the prospect of higher punishment at the end of period 2 induces them to commit less war crimes:

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

Proof. Follows immediately from differentiation of (4). ■

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 malevolent dictators in the long run (by the deterrence effect captured by (18)) and malevolent dictators with $\theta \in [\underline{\theta}, \bar{\theta})$ or $\theta \in [\hat{\theta}, \theta_{\max}]$ will be disciplined (this effect is already present in the short run; Proposition 5). Simultaneously, 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 dictator pool (equation (19)). Second, by the "Terminator effect" (which operates in the short run already), dictators in the mimicking group will choose to commit more crimes following an increase in ϕ , as this is now necessary to unlock the amnesty-option (Proposition 3). In addition, the size of this mimicking group (determined by $(\hat{\theta} - \bar{\theta})$) increases with ϕ (Proposition 4) - so increased commitment to ex-post punishment leads to

more dictators 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 question whether they consider the inevitable increase in dispersion in outcomes to be desirable or not.

	<i>Benign</i>	<i>Malign</i>
<i>Short run</i>	Disciplining effect for $\theta \in [\underline{\theta}, \bar{\theta}) \cup [\widehat{\theta}, \theta_{\max}]$	Terminator effect for $\theta \in [\bar{\theta}, \widehat{\theta})$
<i>Long run</i>	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 incumbent dictator is of type $\theta \in [\underline{\theta}, \bar{\theta})$ or $\theta \in [\widehat{\theta}, \theta_{\max}]$, the effect is good, whereas it is bad if the incumbent is of type $\theta \in [\bar{\theta}, \widehat{\theta})$. This contributes to the aforementioned polarization in the behavior of dictators following an increase in ϕ .

It more generally shows that, when authorities face a time-consistency problem, 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 already (if the "Terminator effect" dominates the disciplining effect) and forms a striking contrast with standard models of criminal activity (where increases in expected punishment tend to reduce crime). The examples of Joseph Kony and Bosco Ntaganda, who both worsened their behavior after the ICC increased its commitment to prosecute them, seem to corroborate the existence of the counterintuitive dynamics predicted by our model. Likewise, Robert Mugabe and his officials seem to fear an ICC prosecution,²³ as a result of which they have resorted to the Terminator strategy as well (recall footnote 7).

This implies that consideration should be given to a more gradual introduction of the new punishment regime. Currently, the ICC has jurisdiction over dictators who were already in power before the ICC came into being in 2002 (and over whom the ex ante deterrence effect thus did not operate). Whether an increase in ϕ will improve the short run behavior of these dictators, fully depends on their type: if they are of type $\theta \in [\underline{\theta}, \bar{\theta})$ or $\theta \in [\widehat{\theta}, \theta_{\max}]$ they will improve their behavior by the disciplining effect, but if they are of

²³See <http://www.theglobeandmail.com/globe-debate/when-justice-stands-in-the-way-of-a-dictators-departure/article623912/>.

type $\theta \in [\bar{\theta}, \hat{\theta})$, the "Terminator effect" will cause them to commit more crime in response to increased commitment of the international community to ex-post punishment.

If one believes that dictators with $\theta \in [\bar{\theta}, \hat{\theta})$ are relatively more numerous,²⁴ it would have been better to give the ICC only jurisdiction over dictators that chose to enter *after* its establishment.²⁵ In that case, the malign "Terminator effect" would at least co-exist with the benign deterrence effect, making it more likely that the long run net effect will be favorable (but remember that this is still not guaranteed, as the ICC also brings an unconditionally malign adverse selection effect).

The former Swedish Foreign Minister Carl Bildt seems to be aware of this. Although Sweden is party to the ICC's Rome Statute, Bildt was the only Foreign Minister of an EU Member State who, in January 2013, refused to support an ICC-indictment for the Syrian president Bashar al-Assad (who had already entered the dictator game before the ICC was established in 2002). Bildt claimed that such a move would be "counterproductive", as "it would put Assad in a headlock and make him less flexible, because we'd be telling him: 'your only option is to fight to the death'".²⁶ Our paper adds to this that Assad also has an option to "fight his way to amnesty" (recall David Cameron's quote in the Introduction and Idi Amin's example, among many others) - which makes it even more likely that violence in Syria would surge if the ICC were to indict Assad.

5 Discussion and directions for future work

This paper has analyzed a new type of time-consistency problem, in which agents can actively engage in strategic behavior to try to force principals to become time-inconsistent. To the best of our knowledge, this "Terminator effect" has not been identified before and carries a general lesson: when regulators lack a perfect commitment technology, it is dangerous for them to *try to* commit as that may invoke a strategic response from regulatees which worsens the original problem.

²⁴Note in this respect that the size of this "Terminator group" is increasing in ϕ , while the group's existence is guaranteed by the properties of (13). Existence of the $[\underline{\theta}, \bar{\theta})$ -group is somewhat more fragile, as it requires both \mathcal{E} being low enough and $\Delta < \hat{\theta}$ (recall Sections 3.3 and 3.2, respectively). If these conditions are not met, all entering malevolent dictators will become part of the "Terminator group" and will hence choose to mimic evil dictators - thereby strengthening our results.

²⁵The ICC's jurisdiction extends only to crimes *committed after* its inception (see Article 11 of the Rome Statute), but our model suggests that this is not yet enough. What matters is the legal regime which was in place at the *entry date* of the dictator, as that determines what type of dictator ended up in office. When this is a dictator of type $\theta \in [\bar{\theta}, \hat{\theta})$, he will be subject to the "Terminator effect" and he will worsen his behavior in response to an increase in ϕ .

²⁶See <http://www.thelocal.se/45642/20130116>.

This insight, which qualifies the original recommendations of Kydland and Prescott (1977), may have applications beyond the dictator setup. Examples include hostage takers who begin their actions by committing atrocities, to signal that they are serious with the aim of opening up negotiations. For instance, in the 1970 "Dawson's Field hijackings", the "Popular Front for the Liberation of Palestine" hijacked and blew up three airplanes. To show that they were to be reckoned with, the terrorists kept the hostages on the planes until only a couple of minutes before they exploded. Shocked by this signal, the international community gave in and released several Palestinian militants from prison. Similarly, a member of the Sicilian Mafia recently revealed that they made plans in 1992 to blow up the Leaning Tower of Pisa "to shock the authorities into making concessions".²⁷

Corresponding considerations apply to kidnapping. Although all kidnapers typically threaten to kill the victim unless a ransom is paid, this is not always their true intention.²⁸ Consequently, authorities face an inference problem on the type of the kidnapers, while kidnapers 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. This grandson of an oil tycoon was kidnapped in 1973. 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 kidnapers 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.

Interestingly, the comparative statics predicted by our model are also observed in reality: after Italy implemented a law in 1991 that froze the family assets of hostages (increasing commitment to not pay a ransom), the number of kidnappings went down but those that did take place were more gruesome and lasted longer (Detotto, McCannon and Vannini, 2012). 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 force time-inconsistency.

²⁷Cf. http://www.thetimes.co.uk/tto/news/world/europe/article3984464.ece?CMP=OTH-gnws-standard-2014_01_23.

²⁸Although 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 kidnapers *without* any form of payment. In only 6 percent of all cases, the victim is actually killed (Australian Senate, 2011: 16).

Turning back to the behavior of dictators, Charles Taylor also employed a variant of the Terminator strategy in Liberia's 1997 presidential election. During the campaign, Taylor ran the threatening slogan "He killed my ma, he killed my pa, but I will vote for him". Taylor subsequently won the election by taking 75 percent of the vote - a success that is attributed to the belief among war-weary voters that he would have pushed his country back into war if he lost.²⁹

Other applications of this concept may lie in financial markets: on "Black Wednesday" in 1992, George Soros' short positions managed to break the Bank of England's commitment to stay within the ERM. Alluding to this example, Richard Fisher (president of the Federal Reserve Bank of Dallas) recently warned that the "feral hogs" of financial markets were trying to force the Fed to relinquish its plans to slow bond purchases by overreacting to the mere announcement of such "tapering". In the *Financial Times* of June 25th 2013, Fisher stated that "markets tend to test things. We haven't forgotten what happened to the Bank of England on Black Wednesday. I don't think anyone can break the Fed. But I do believe that big money does organise itself somewhat like feral hogs. If they detect a weakness or a bad scent, they'll go after it."

Once one starts thinking about time-consistency and dictator punishment, many more questions arise. This paper only takes a first step and leaves several issues for future work. Most extensions that have been made to Barro and Gordon's (1983) work on time-consistency and monetary policy, are relevant to the present setup as well. An interesting extension might be one along the lines of Backus and Driffill (1985). They consider a game in which the public is not aware of the monetary policy maker's type (tough or weak on inflation?). A tough policy maker always chooses zero inflation, while a weak one might inflate - e.g. when there is unemployment. By setting a positive inflation rate, a policy maker would therefore reveal himself as "weak", after which he will suffer from an inflationary bias. Consequently, it might be optimal for a weak central banker to mimic a tough one for a while, so as to bring expected inflation down.³⁰ In the context of this paper, it seems likely that dictators are currently uncertain about the "type" ϕ of relevant new institutions (like the ICC) that have not established a firm reputation yet. Hence, strategic issues underlying this learning process might be worth studying as well.³¹

One may also want to consider a variant of the model in which the dictator has

²⁹See <http://www.theguardian.com/world/2003/aug/04/westafrica.qanda>.

³⁰Note that in the Backus-Driffill model, mimicking takes place on the side of the regulator - not on the side of the regulatee as in our paper.

³¹In this respect, it is noteworthy that the ICC's first sentence (against Thomas Lubanga, who was convicted to 14 years in prison) was surprisingly lenient to many. Within a year following this verdict, Bosco Ntaganda (Lubanga's co-accused) handed himself in.

more control over the expected duration of his tenure (for example via the amount of repression that he exercises). Although this would change the *level* of war crimes chosen by the dictator, intuition suggests that it would not affect the existence of the various effects identified in (and central to) this paper. After all, the existence of these effects stems from the notion that the tenure of malevolent dictators can be shortened through amnesty-abdication deals - a mechanism that would still be present in a setup where the dictator has more control over the duration of his time in office. As such an extension would also introduce additional issues that are orthogonal to the time-consistency problem that we have focused on in this paper (recall footnote 9), we leave it for future work.

Relatedly, the current paper has primarily focused on that class of dictators which derives direct utility from engaging in war crimes. While this class contains many important despots from the past,³² there may also exist a group of dictators that primarily engages in war crimes as a means to stay in power (see Larcom, Sarr and Willems (2014)). The reasoning set out in the previous paragraph suggests that the effects central to this paper will continue to be present in that setup, but this of course remains to be verified through a formal analysis.

Finally, our model can also be extended by giving the international community more control over the speed with which a malevolent dictator is removed and prosecuted (which currently takes two periods of indeterminate, exogenous length and succeeds with probability π). Again, such an extension would not affect the existence of the effects identified in (and central to) this paper, as they derive from the more general idea that an amnesty-abdication deal is the quickest way to bring the atrocities to an end.

6 Conclusion

The punishment strategy for malevolent dictators is complicated by a time-consistency problem: whereas authorities typically want to threaten rulers with a harsh penalty for committing war crimes *ex ante*, there are strong incentives to forget about these threats while these crimes are being committed as an amnesty-abdication deal has the potential to end further suffering in the quickest possible way.

Recently, the international community has broken with its amnesty-ridden past by increasing its commitment to prosecute malevolent dictators (for example through the

³²Think about the likes of Adolf Hitler and Saddam Hussein who seem to have obtained direct gratification from implementing policies that were aimed at exterminating particular groups (Jews, Kurds, et cetera). Also recall footnote 10 and see Coolidge and Segal (2009) for psychological evidence that many dictators suffer from a sadistic personality disorder.

installation of a permanent International Criminal Court which has already indicted 30 high-profile individuals). In this paper, we have constructed a model that enables an analysis of the consequences of this policy change - taking into account that regulatees in this setup may act strategically.

Our model identifies two benign effects associated with increased commitment to the ex-post punishment of malicious dictators: first of all, there will be a long run deterrence effect - leading to fewer malevolent rulers in the future. Secondly, some malevolent rulers will be disciplined in the short run already.

Simultaneously, however, there are also types who will choose to commit *more* war crimes when faced with a higher prospective punishment. Dispersion in outcomes will therefore go up, so parties involved should ask themselves whether they consider such polarization to be desirable or not.

The reason for the malign effect is twofold: first of all, higher prospective punishment selects malevolent dictators of a worse type, as only the "not-so-bad ones" are deterred by ex ante threats. Secondly, as long as the international community's loss associated with time-inconsistency remains *finite*, there will always exist a critical level of bloodshed beyond which time-inconsistency becomes optimal. In those circumstances, the international community would rather incur the loss associated with being time-inconsistent (i.e.: making an amnesty-abdication deal), than the pay-off that it expects to receive from sticking to its earlier promise to try the dictator. This gives rise to a "mimicking dictator group" which consciously chooses to worsen its behavior in period 1. Through their brutal first-period behavior, these dictators pretend to be of the "evil" type, to whom the international community should offer amnesty (to end suffering for the population). Hence, dictators in the mimicking group try to "unlock" the amnesty-option by actively forcing the international community into time-inconsistency. Since both the size of this group as well as the malevolence of its behavior increase with the international community's commitment to ex-post punishment, this brings more war crimes. In sharp contrast to standard Beckerian crime models, this shows that criminal activity may actually *go up* in response to an increase in expected punishment.

Interestingly, the mimicking behavior predicted by our model is indeed observed in reality: when Bosco Ntaganda's rebel troops attacked the city of Goma in response to Ntaganda's ICC-indictment, they issued a statement saying that they were not interested in control of the city, but all they wanted "is that the Congolese government sit down at the negotiating table". Similarly, members of Robert Mugabe's government have stated that "top lieutenants are trying to force the political opposition into granting them amnesty for

their past crimes by abducting, detaining and torturing opposition officials and activists".

So in the presence of a time-consistency problem for authorities, increased commitment to ex-post punishment of crimes can paradoxically lead to more crimes being committed (as it then pays to be *very* bad, rather than just a little bad). This shows that in the absence of a *perfect* commitment technology, which prevents regulators from being fully time-consistent, it is a risky strategy for them to *try* to be tough: civilians may be killed by dictators, ears may be chopped off by kidnappers, or the Tower of Pisa may be blown up by the Mafia (recall the examples given in Section 5) - purely with the aim of forcing authorities into time-inconsistency. This qualifies the original policy implications of Kydland and Prescott (1977) and the subsequent literature on "rules rather than discretion". This paper instead suggests that in the presence of strategic regulatees, "discretion" may very well dominate "rules" - a finding that turns the original Kydland-Prescott result on its head.

Whether the benign forces are able to outweigh the malign adverse selection and "Terminator" effects, is generally ambiguous and remains a question that is in great need for additional research. We therefore hope that the first steps taken in this paper, will be followed by additional strides so as to get a better understanding of this important issue.

7 Appendix

In this Appendix we show that all propositions stated in Section 4 are robust to using the more general $\Lambda = \Lambda(\phi, \phi x_1)$, with $\Lambda_1 \equiv \partial \Lambda(\phi, \phi x_1) / \partial \phi > 0$ and $\Lambda_2 \equiv \partial \Lambda(\phi, \phi x_1) / \partial (\phi x_1) > 0$. Having $\Lambda_2 > 0$ captures the idea that it is costlier for the international community to grant amnesties to dictators who behaved 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:

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

Now recall that $\hat{\theta} = 2(\gamma + \phi) \Lambda(\phi, \phi \hat{x}_1) - \bar{\theta}$, which implies $\frac{\hat{\theta} - \bar{\theta}}{2} = (\gamma + \phi) \Lambda(\phi, \phi \hat{x}_1) - \bar{\theta} > 0$.

As a result:

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

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

As before:

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

Consequently, by the implicit function theorem:

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

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:

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

which confirms Proposition 3.

From this, it immediately follows that:

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

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

8 References

Acemoglu, D. and J.A. Robinson (2006), *Economic Origins of Dictatorship and Democracy*, Cambridge: Cambridge University Press.

Acemoglu, D. and J.A. Robinson (2008), "Persistence of Power, Elites, and Institutions", *American Economic Review*, 98 (1), pp. 267-293.

Acemoglu, D., J.A. Robinson and R. Torvik (2013), "Why Do Voters Dismantle Checks and Balances?", *Review of Economic Studies*, 80 (3), pp. 845-875.

Anderlini, L., L. Felli and A. Riboni (forthcoming), "Why Stare Decisis?", *Review of Economic Dynamics*.

- Aslaksen, S. and R. Torvik (2006), "A Theory of Civil Conflict and Democracy in Rentier States", *Scandinavian Journal of Economics*, 108 (4), pp. 571-585.
- Australian Senate (2011), *Held Hostage: Government's Response to Kidnapping of Australian Citizens Overseas*, Canberra: Senate Printing Unit.
- Backus, D. and J. Driffill (1985), "Inflation and Reputation", *American Economic Review*, 75 (3), pp. 530-538.
- Barro, R.J. and D.B. Gordon (1983), "Rules, Discretion and Reputation in a Model of Monetary Policy", *Journal of Monetary Economics*, 12 (1), pp. 101-121.
- Becker, G.S. (1968), "Crime and Punishment: An Economic Approach", *Journal of Political Economy*, 76 (2), pp. 169-217.
- Besley, T. and T. Persson (2011a), "The Logic of Political Violence", *Quarterly Journal of Economics*, 126 (3), pp. 1411-1445.
- Besley, T. and T. Persson (2011b), "Fragile States and Development Policy", *Journal of the European Economic Association*, 9 (3), pp. 371-398.
- Cho, I.K. and D.M. Kreps (1987), "Signaling Games and Stable Equilibria", *Quarterly Journal of Economics*, 102 (2), pp. 179-221.
- Coolidge, F.L. and D.L. Segal (2009), "Is Kim Jong-il like Saddam Hussein and Adolf Hitler? A Personality Disorder Evaluation", *Behavioral Sciences of Terrorism and Political Aggression*, 1 (3), pp. 195-202.
- D'Ancona, M. (2013), *In It Together: The Inside Story of the Coalition Government*, New York: Viking Press.
- Detotto, C., B.C. McCannon and M. Vannini (2012), "Understanding Ransom Kidnapping and its Duration", CRENoS Working Paper No. 19.
- Escribà-Folch, A. (2013), "Accountable for what? Regime Types, Performance, and the Fate of Outgoing Dictators, 1946-2004", *Democratization*, 20 (1), pp. 160-185.
- Esteban, J., M. Morelli and D. Rohner (2012), "Strategic Mass Killings", mimeo, Columbia University.
- Fischer, S. (1980), "Dynamic Inconsistency, Cooperation and the Benevolent Dissembling Government", *Journal of Economic Dynamics and Control*, 2, pp. 93-107.
- Flood, R.P. and P. Isard (1989), "Monetary Policy Strategies", *IMF Staff Papers*, 36 (3), pp. 612-632.
- Freeman, M. (2009), *Necessary Evils: Amnesties and the Search for Justice*, Cambridge: Cambridge University Press.
- Garfinkel, M.R. and S. Skaperdas (2007), "Economics of Conflict: An Overview", in: T. Sandler and K. Hartley (eds.), *Handbook of Defense Economics*, Amsterdam: Elsevier.

Guimaraes, B. and K.D. Sheedy (2012), "A Model of Equilibrium Institutions", mimeo, London School of Economics.

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

Larcom, S., M. Sarr and T. Willems (2014), "Dictators Walking the Mogadishu Line: How Men Become Monsters and Monsters Become Men", mimeo, University of Oxford.

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

Nayenga, P.F.B. (1979), "Myths and Realities of Idi Amin Dada's Uganda", *African Studies Review*, 22 (2), pp. 127-138.

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

Persson, T. and G. Tabellini (1990), *Macroeconomic Policy, Credibility, and Politics*, London: Harwood Academic Publishers.

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

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

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

Schelling, T.C. (1966), *Arms and Influence*, New Haven: Yale University Press.

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