What shall we do with the bad dictator?*

Shaun Larcom† Mare Sarr‡ Tim Willems§

June 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 atrocities 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, Avinash Dixit, Mikhail Drugov, Georgy Egorov, Markus Gehring, Ian Jewitt, Antonio Mele, Meg Meyer, Ines Moreno de Barreda, Ricardo Nunes, Roger O'Keefe, Rick van der Ploeg, Eric Posner, Kevin Sheedy, Ragnar Torvik, Sweder van Wijnbergen, and audiences at the 17th IEA World Congrress and at the Universities of Oxford and Pretoria for useful comments.

†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 malevolent 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 and on March 12 2012, the former Congolese military commander Thomas Lubanga became the first person to be convicted by the ICC (he was sentenced to 14 years imprisonment). 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 (the Westphalia Peace Treaty of 1648 seems the first major case in international law of this

sort). This practice of "trading justice for peace" was not only popular with dictators,[1] but was often 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 on the top two floors of a Saudi Arabian hotel - dying there in 2003, without ever being held to account for his crimes. In fact, Amin was even paid a "handsome monthly sum" (Amin's own words) by the Saudis, with tacit support of the Reagan Administration, under the condition that he would stay away from politics. This specific asylum was however not successful in improving the quality of governance in Uganda as Amin was replaced by Milton Obote, who is held responsible for even more deaths than Amin.[2]

There are more successful examples: Charles Taylor stepped down as president of Liberia in 2003 when rebel troops were surrounding the city of Monrovia, in return for a guarantee of 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 (leading to the aforementioned conviction),[3] this case does illustrate the potential of improving conditions by promising impunity to dictators in return for their abdication.

Various legal scholars, such as Scharf (1999: 507), have therefore criticized the new incentives - arguing that an ICC may bring "justice at the expense of peace". Our model

[1] The 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, Cambodia, 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.

[2] Obote's regime ended with an asylum as well (he fled to Tanzania and later to Zambia). This asylum was more successful in improving local living conditions as Obote was succeeded by Yoweri Museveni, who has brought relative stability and economic growth to Uganda. Recently, however, the situation has deteriorated again (due to Joseph Kony's "Lord's Resistance Army"), more on which below.

[3] This 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.

suggests that there indeed lies some truth in this. 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 atrocities committed by Bashar al-Assad in Syria around November 2012:

> 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.[4]

Similarly, Qatar has considered granting Assad asylum because "the region needs to stabilize".[5]

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 under the new post-ICC regime (legally, this is also perfectly possible given that the Rome Statute of the ICC contains several explicit "escape clauses", see footnote 15). D'Ancona (2013) for example presents evidence that during the Libyan Civil War, the British Secret Service MI6 was arranging asylum for the ICC-indicted dictator Muammar Gaddafi in Equatorial Guinea. That country does not recognize the ICC and was willing to host Gaddafi (who was in turn willing to go there). 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 the former Tunisian president Ben Ali accepted an asylum offer from Saudi Arabia in 2011. Similarly, in 2012 a local (US/EU/UNSC-backed) amnesty-abdication deal was made with Abdullah Saleh of Yemen.

We show that dictators anticipating this may choose to commit more atrocities 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

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

[5] See 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.

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.

Increasing the costs of time-inconsistency may therefore induce some dictators to commit *more* atrocities, as this is now necessary to make the amnesty option available to them. This suggests that increased commitment to ex post punishment 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 commit more atrocities.

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 ongoing 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 the charges against him were dropped. Later that year, the LRA went on a major offensive - carrying out a massacre on Christmas Day. Over the following three weeks, they killed at least 800 people and abducted hundreds more - and this unstable situation continues to today.[6] 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".[7] 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. After committing many more atrocities, Ntaganda however came under pressure from a rival rebel group as a result of which he chose to hand himself in on 18 March 2013. Ntaganda thus ultimately failed to break the international community's intention to prosecute him;[8] Kony on the other hand, is still at large.

Similar forces seem to be at work in Zimbabwe. According to members of the Mugabe-government, "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"[9] - a striking statement that is exactly in line with our model's prediction.

At the same time, proponents of criminal courts emphasize their ex ante effects. When

[6] See e.g. http://www.bbc.co.uk/news/world-africa-17299084.

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

[8] In addition, Ntaganda's surrender may be related to the aforementioned conviction of Thomas Lubanga (which was surprisingly lenient to many), more on which in footnote 30 below.

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

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 made Adolf Hitler 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. 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 present paper builds upon several literatures. It first of all 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)). This paper, in contrast, applies similar ideas to the treatment of dictators who are committing atrocities while in power. Moreover, this paper develops and analyzes a model that adds a new dimension to the standard problem, namely that of strategic behavior from regulatees (as they may actively try to force regulators to forget about their earlier commitments and become time-inconsistent instead). Hereby, our paper also contributes to the literature on blackmailing and threats (see Shavell (1993) and Schwarz and Sonin (2008) for earlier contributions). 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 Egorov and Sonin (2005) 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.

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 atrocities 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 institutions. Examples include Acemoglu and Robinson (2006, 2008), who develop a framework for analyzing the creation and consolidation of democracy; Besley and Persson (2010, 2011), 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.

The present paper focuses 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) 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 (cf. 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, 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 Environment

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 atrocities 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 atrocities 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 atrocities 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 atrocities 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).[10] 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. Our results also generalize to a setup in which the periods are of unequal lengths.

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).[11] 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

[10] It for example took two wars and many sanctions - spanning a period of about 13 years - to remove Saddam Hussein from power in Iraq, while 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.

[11] Alternatively, $h(x_1 + x_2)$ can be interpreted as the punishment function of the local population; π then represents the probability that the local population ousts the dictator. One could also 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. Whether a higher choice for x_t in- or decreases expected tenure is not obvious: while oppressing civilians presumably decreases the probability of a local revolution, it simultaneously increases the probability of a foreign military intervention. We therefore leave an analysis of 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.

with the product of these two parameters, defined as:

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

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

A dictator in power can commit atrocities in periods 1 and 2 from which he derives utility. One can think of this leader as a sadistic dictator who takes pleasure in maltreating his subordinates/opponents (as in Wintrobe (1998)),[12] or as a reduced form way of modeling a dictator who derives utility from extracting rents, but subsequently needs to exercise repression x_t so as to prevent the (angered) public from ousting him (rent extraction and repression tend to go hand-in-hand as the former creates a need for the latter; also see footnote 12 and Larcom, Sarr and Willems (2014) who develop a model along these lines).

Exercising repression 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 atrocities 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 then given by:

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

Here, $\theta \geq 0$ is a time-invariant parameter capturing the dictator's type: a dictator with a high θ takes a lot of pleasure from committing atrocities relative to the punishment, while this holds less so for a dictator with a low θ. 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 pleasure from committing atrocities in periods 1 and 2, while

[12] As the "Stanford Prison Experiment" has shown, even "normal" people can turn into sadists once they find themselves with power over others (Zimbardo, 2007). Also recall the 2003-2004 experiences in the Iraqi Abu Ghraib prison (where human rights of detainees were violated by members of the US army without an apparent reason). Moreover, psychological research has established that many notorious dictators suffered from a sadistic personality disorder (see e.g. Coolidge and Segal (2009)).

punishment can only follow at the end of period 2).[13] 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).

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

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

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

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

where γ, ϕ, and x_1 are all known at this stage.[14]

3.2 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.[15] In our model, time-inconsistency implies forgetting

[13] For 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.

[14] We 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).

[15] Alternatively, 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: the United Nations Security Council can ask the Court to defer a prosecution for one year (which could be requested annually) (Article 16); the Court could be persuaded that a state has

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, 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 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 will be getting more serious about prosecuting malevolent dictators (i.e.: increasing π, the probability with which a misbehaving dictator will be tried), or if it states that it will punish such dictators more severely (i.e.: increasing ψ), then the (reputational) loss associated with subsequently *not* doing so will be higher (remember that $\phi \equiv \pi\psi$; cf. equation (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.

The value of the amnesty offer to the dictator is 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 that he could spend his last years in a Saudi Arabian hotel suite while receiving "a handsome monthly sum" from the Saudi government). The international community will ensure that it is indeed optimal for the dictator to accept the offer and step down by setting $V_A > \theta \log(x_2^*) - \gamma x_2^* - \phi[x_1 + x_2^*]$.[16] Since the dictator has an informational advantage over the international community at this stage (he knows his true value of θ, while the

already carried out a prosecution (even if only a light punishment was attached) (Article 17); and finally, charges could also be dropped if the prosecutor believes that a prosecution would "not serve the interests of justice" (Article 53).

[16] The RHS of this inequality represents the utility to a dictator who stays in office after having chosen x_1 in period 1, and sets x_2 optimally according to FOC (4). Note that there is no point for the international community to make an offer that does not satisfy this condition. As evidenced by the long list of dictators who have accepted amnesty or asylum offers in the past (see footnote 1), the international community seems able to meet this condition in reality. The recent examples discussed in the Introduction (of Aristide, Ben Ali, Gaddafi, and Saleh, who all accepted amnesty/asylum offers made after 2002) show that this continues to hold in the post-ICC regime.

international community does not) he 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 equals:

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

Because this promise of the international community might not be fully credible either (recall footnote 3), one may interpret V_A as an *expected* value to the dictator (and note that the model's solution is fully determined by expected values where applicable).

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). We assume that the international community holds the rather optimistic belief that it can install a benevolent dictator with $\theta' = 0$, such that it is able to "trade justice for peace" - letting x_2 collapse to zero. Whether this belief is justified or not, does not matter for the model's solution (again because the latter is fully determined by expected values); qualitatively similar results would be obtained in the more general case where the international community believes to be able to replace the abdicating dictator by any better type $\theta' \in [0, \theta)$ (which is to be expected by "regression to the mean" when θ is relatively high).[17] Hence, after observing x_1, the international community's expected loss function becomes:

$$\mathbb{E}_2^{IC}\{\mathcal{L}|x_1\} = \begin{cases} x_1 + \mathbb{E}_2^{IC}\{x_2|x_1\} & \text{if } \mathbf{1}_A = 0 \\ x_1 + \Lambda(\phi) & \text{if } \mathbf{1}_A = 1 \end{cases}, \tag{1'}$$

where $\mathbf{1}_A$ is an indicator function that takes the value 1 if the incumbent dictator

[17] Moreover, any atrocities 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 (as in Liberia where Charles Taylor was replaced by Ellen Johnson Sirleaf), 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. This seems to have happened in Haiti (Scharf, 2006): in the early 1990s, Haiti was ruled by a malicious military regime headed by Raoul Cédras and Philippe Biamby. Around 1994, the United Nations mediated negotiations in which the military leaders agreed to step down and permit the return of the democratically elected president Jean-Bertrand Aristide in exchange for a full amnesty for the members of the regime. Initially, this deal had the desired effect: Aristide returned, reinstalled a civilian government, and human rights abuses ended immediately with hardly any bloodshed. However, by the early 2000s Aristide's government had turned into a corrupt one that was abusing human rights, which led to a wave of violent protests and riots - eventually making Aristide accept an asylum offer from South Africa in 2004.

abdicates at the beginning of period 2 in return for amnesty (or only a minor punishment), and zero otherwise. 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 atrocities that the dictator will commit during period 2 is subsequently given by:

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

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:

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

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 promise to try the dictator, is larger than the one associated with being time-inconsistent. In those cases, the international community (or an individual member of it, in casu the country which is most altruistic or has the greatest incentive to bring stability) prefers to end violence immediately by making an amnesty-abdication deal and 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). Note that condition (8) directly implies that the greater the international community's ex ante commitment to punishing malevolent dictators (i.e.: the higher its choice for ϕ and hence $\Lambda(\phi)$), the more atrocities a dictator has to commit before the amnesty option is unlocked ex post.

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

$$\mathbb{E}_2^{IC}\left\{\theta \middle| x_1 = \frac{\widehat{\theta}}{\gamma+\phi}\right\} = (\gamma+\phi)\Lambda(\phi) \tag{9}$$

Henceforth, we will refer to those dictators with $\theta \geq \widehat{\theta}$, as the "evil" ones: by adhering

to their FOC, they already commit so many atrocities that the international community is better off by offering them amnesty in return for their abdication.

For a dictator of type $\theta < \widehat{\theta}$, 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 atrocities 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 his $\theta \geq \widehat{\theta}$) and that he will commit many atrocities in the second period as well.[18] So many, that the international community is actually better off by making an amnesty-abdication deal at the beginning of period 2. 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 Ntanganda (see footnote 18), we will call this the "Terminator effect" (although "Mugabe effect" would be another fitting name, as 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" - exactly in line with our model's prediction; recall footnote 9).

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 $\widehat{\theta}$, so it only pays for the high θ types to try and mimic the period 1 behavior of evil dictators. To see this formally, start by noting that a dictator whose $\theta < \widehat{\theta}$ 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).

[18] This is very much like the strategy employed by Bosco Ntaganda (nicknamed "The Terminator") who attacked the city of Goma in 2008 not because he was interested in its control, but only because he "wanted to make the government sit at the negotiation table" (cf. footnote 7). 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.'" As we will argue in Section 5, this strategy is similar to that often displayed by hostage takers or kidnappers. To signal that they are serious, they tend to start their actions in an atrocious way so as to "break" authorities and open up negotiations.

$\mathcal{D}$. 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:

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

Since we focus on dictators with $\theta < \widehat{\theta}$, setting x_1 according to the FOC (4) does not unlock the amnesty option (because $x_1^* < \widehat{x}_1$ for those types, with $\widehat{x}_1$ being the critical level of first-period atrocities that unlocks the amnesty option). The cheapest way for such a dictator to unlock the amnesty option nevertheless, would be to mimic a type $\widehat{\theta}$-dictator by setting $x_1 = \widehat{x}_1$. After all, since $\widehat{x}_1 > x_1^*$, the marginal benefits from committing atrocities 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 atrocities (setting $x_1 = \widehat{x}_1$ suffices to unlock the amnesty option, so setting $x_1 > \widehat{x}_1$ only brings additional net costs). Following this strategy would yield a dictator of type $\theta < \widehat{\theta}$ lifetime utility:

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

Hence, a dictator will "invest" in committing atrocities (and form a pooling equilibrium with the "evil" types) in period 1 iff his θ is such that $U(\mathcal{D}) \geq U(\mathcal{F})$, i.e. iff:

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

When $y(\theta, \phi) = 0$, it implicitly defines the critical type-level $\overline{\theta}$ beyond which a dictator becomes willing to invest in first-period atrocities for the sole purpose of unlocking the amnesty option in period 2, such that he can exit via the amnesty route, enjoying $V_A(\overline{\theta}, \phi)$. We then know from combining equations (6) and (12) that $\overline{\theta}$ should satisfy:

$$y(\overline{\theta}, \phi) = \overline{\theta}\left[\log\left(\frac{\widehat{\theta}}{\overline{\theta}}\right) + 1\right] - \widehat{\theta} + \Delta = 0 \tag{13}$$

Assuming that $\Delta < \widehat{\theta}$ (otherwise $y(\theta, \phi) > 0 \ \forall \theta$ and we end up in a situation where *all* types with $\theta > 0$ are mimicking, which would actually strengthen our results but strikes us as being rather extreme) there exists a $\overline{\theta} < \widehat{\theta}$ for which $y(\overline{\theta}, \phi) = 0$ (because

$y(\cdot)$ is then a continuous function with $y(0,\phi) < 0$, $y(\widehat{\theta},\phi) > 0$, and $\partial y(\theta,\phi)/\partial\theta|_{\theta=\overline{\theta}} > 0$). Consequently, all types $\theta \in \left[\overline{\theta},\widehat{\theta}\right)$ will find it optimal to mimic the $\widehat{\theta}$-type by setting $x_1 = \widehat{x}_1$. The reason is that in the face of a 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]

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

The intuition is that when the international community observes $\widehat{x}_1 \equiv \widehat{\theta}/[\gamma+\phi]$ (that level of atrocities which is just enough to unlock the amnesty option), the naive expectation would be to assume that the responsible dictator is exactly of type $\widehat{\theta}$ (this would follow from blindly applying equation (5)). In perfect Bayesian equilibrium, however, the international community realizes that $\widehat{x}_1$ could also be set by a dictator of type $\theta \in \left[\overline{\theta},\widehat{\theta}\right)$, who is trying to mimic a dictator of type $\widehat{\theta}$ so as to unlock the amnesty option (the "Terminator effect"). Consequently, the rational expectation is $\mathbb{E}\left\{\theta|\overline{\theta} \leq \theta \leq \widehat{\theta}\right\} < \widehat{\theta}$ so 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.

At this point, we know that $\overline{\theta}$ should solve (13), while $\widehat{\theta}$ 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 $\mathbb{E}_2^{IC}\{\theta|\widehat{x}_1\} = (\overline{\theta}+\widehat{\theta})/2$ and condition (9) implies:

$$\widehat{\theta} = 2(\gamma+\phi)\Lambda(\phi) - \overline{\theta} \tag{15}$$

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

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

This completes the solution for this part of the model.

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

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.[20] 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).

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 $\theta < \overline{\theta}$. By definition of $\overline{\theta}$, 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:

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

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 $\underline{\theta} < \overline{\theta}$, 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 $p\left(\underline{\theta}\right)$ (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).

[20] We introduce the mass point at $\theta = 0$ to be able to capture the deterrence effect in equation (18) below, but this could also be done in many 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 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 atrocities, 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 atrocities). 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:

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

with:

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

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 atrocities given that they can't derive any utility from doing so. Consequently, the probability that a country ends up with a benevolent ruler increases 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 $\mathbb{E}\left\{\theta|\mathbf{1}_{\mathcal{B}} = 0\right\} = \mathbb{E}\left\{\theta|\theta \geq \underline{\theta}\right\}$, for which it holds that:

$$\frac{\partial \mathbb{E}\left\{\theta|\theta \geq \underline{\theta}\right\}}{\partial \phi} > 0, \tag{19}$$

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 Summarizing

Now that we have specified the setup, we can summarize our model by the following sequence of events:

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.

2. Given his θ, this dictator commits atrocities x_1 during period 1.

3. At the beginning of period 2, the international community forms the expectation $\mathbb{E}_2^{IC}\{\theta|x_1\}$ as in (14). It subsequently decides whether to make the dictator an amnesty offer (according to (6)) or not.

4. If an amnesty is offered and accepted, $x_2 = 0$ and the dictator gets V_A. If no amnesty-abdication deal is made, the dictator commits atrocities 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 atrocities that he has committed in the past. Otherwise, he is able to leave the stage unpunished (the 90 year old Zimbabwean president Robert Mugabe is well on his way to become an example in this category).

Denoting the international community's action at the beginning of period 2 by $s \in \{$*offer amnesty,punish*$\}$, one can define the equilibrium of our signalling game 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* $E\{\mathcal{L}|x_1\}$*.*

2. *The beliefs* $\mu(\theta|x_1)$ *held by the international community about the dictator's type* θ *are compatible with Bayes' rule.*

3. *Taking the international community's best response* $s^{eq}(x_1)$ *into account, a dictator of type* θ *decides to commit atrocities* $x_t^{eq}(\theta)$ *that maximizes his utility.*

Eventually, our model 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 atrocities to make up for that fee.

- The "bad dictators" with $\underline{\theta} \leq \theta < \overline{\theta}$. 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 $\theta \geq \widehat{\theta}$.

- The "mimicking dictators" with $\overline{\theta} \leq \theta < \widehat{\theta}$. This is an interesting group, as they choose to modify their first-period behavior by "investing" in committing first-period atrocities, 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 force the international community to become time-inconsistent.

- The "evil dictators" with $\theta \geq \widehat{\theta}$. 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 $\underline{\theta} < \overline{\theta} < \widehat{\theta}$ - 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 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).[21] 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 $\partial\overline{\theta}/\partial\phi > 0$.

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

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

[21] Since 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.

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

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

Consequently, the implicit function theorem implies:

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

■

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

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

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

Proof. Differentiating (15) gives:

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

since $\overline{\theta} < \widehat{\theta}$. ■

This states that the location of the threshold type $\widehat{\theta}$ 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 furthermore implies that when ϕ is higher, dictators of type $\theta < \widehat{\theta}$ will have to put in a greater mimicking effort (and commit more atrocities) if they want to unlock the amnesty option.

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

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

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

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

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

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

Using the Lambert W function gives:

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

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

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

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

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 $\left[\overline{\theta}, \widehat{\theta}\right)$-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 $\partial\widehat{\theta}/\partial\phi > \partial\overline{\theta}/\partial\phi$, which implies that the mimicking group grows in size when ϕ increases (this explains why arrow 4 points up).

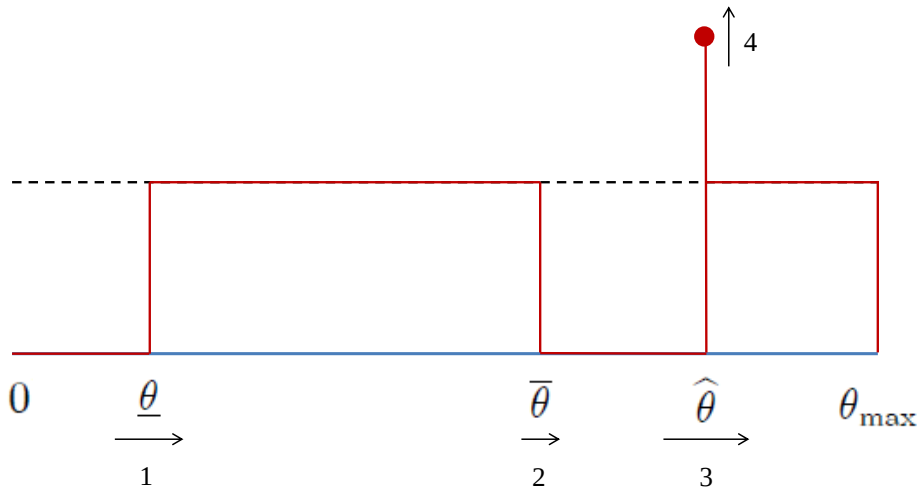

Figure 1: Graphical illustration of the setup when θ follows a uniform distribution

Using this "effective" period 1 distribution for θ, its expected value at the beginning of period 1 is given by:

$$\mathbb{E}_1^{\dagger}\{\theta\} = \frac{1-F(\underline{\theta})}{b+1-F(\underline{\theta})}\left[\int_{\underline{\theta}}^{\overline{\theta}} \theta f(\theta)d\theta + \left[F(\widehat{\theta}) - F(\overline{\theta})\right]\widehat{\theta} + \int_{\widehat{\theta}}^{\theta_{\max}} \theta f(\theta)d\theta\right], \tag{20}$$

where the †-superscript indicates that this is the expectation after entry and mimicking incentives have been taken into account. The first term in (20) captures the probability that a malevolent dictator takes office ($\Pr[\mathbf{1}_{\mathcal{B}} = 0]$; after all, with complementary probability this will be a benevolent dictator with $\theta = 0$), while the term in square brackets represents $\mathbb{E}_1^{\dagger}\{\theta|\mathbf{1}_{\mathcal{B}} = 0\}$.

Finally, there is also a "disciplining effect" for those dictators with $\theta \in \left[\underline{\theta}, \overline{\theta}\right)$ or $\theta \in [\widehat{\theta}, \theta_{\max}]$. 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 atrocities:

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 $\theta \in \left[\underline{\theta}, \overline{\theta}\right)$ or $\theta \in [\widehat{\theta}, \theta_{\max}]$ 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 atrocities 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 $(\widehat{\theta} - \overline{\theta})$) increases with ϕ (Proposition 4). 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.

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 $\theta \in \left[\underline{\theta}, \overline{\theta}\right)$ or $\theta \in [\widehat{\theta}, \theta_{\max}]$, the effect is good, whereas it is bad if the incumbent is of type $\theta \in \left[\overline{\theta}, \widehat{\theta}\right)$. This contributes to the aforementioned polarization in the behavior of dictators following an increase in ϕ.

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

Table 1: Effects brought about by an increase in ϕ

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. The examples of Joseph Kony and Bosco Ntaganda (who both worsened their behavior after the ICC increased its commitment to prosecute them by unsealing their arrest warrants) seem to corroborate this claim.

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 $\theta \in \left[\underline{\theta}, \overline{\theta}\right)$ or $\theta \in [\widehat{\theta}, \theta_{\max}]$ they will improve their behavior by the disciplining effect, but if they are of type $\theta \in \left[\overline{\theta}, \widehat{\theta}\right)$, the "Terminator effect" will cause them to commit more atrocities in response to increased commitment of the international community to ex post punishment.

If one believes that dictators with $\theta \in \left[\overline{\theta}, \widehat{\theta}\right)$ are relatively more numerous,[22] it would have been better to give the ICC only jurisdiction over dictators that chose to enter *after* its establishment.[23] In that case, the malign "Terminator effect" would 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 Swedish Foreign Minister Carl Bildt seems to be aware of this. Although Sweden is party to the Rome Statute of the ICC, 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'".[24] We would add that Assad also has an option to "fight his way to amnesty" (recall David Cameron's quotation in the Introduction and Idi Amin's example (among many others)) - making it even more likely that violence in Syria would surge if the ICC were to indict Assad.

[22] Note 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 $\left[\underline{\theta}, \overline{\theta}\right)$-group is somewhat more fragile, as it requires both $\mathcal{E}$ being low enough and $\Delta < \widehat{\theta}$ (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.

[23] The 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 $\theta \in [\overline{\theta}, \widehat{\theta})$, he will be subject to the "Terminator effect" and he will worsen his behavior in response to an increase in ϕ.

[24] See http://www.thelocal.se/45642/20130116.

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 this may invoke a strategic response from regulatees which worsens the original problem.

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".[25]

Corresponding considerations apply to kidnapping. Although all kidnappers typically threaten to kill the victim unless a ransom is paid, this is not always their true intention.[26] Consequently, authorities face an inference problem on the type of the kidnappers, while kidnappers have an incentive to signal that they are of the worst possible type (as that maximizes the probability that their demands will be met). One of the most fitting examples is that of John Paul Getty III. 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 kidnappers recognized the problem and broke the stale mate 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.

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

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

Interestingly, the comparative statics predicted by our model are 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 seemed to be more gruesome and last 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 (receiving the first ear was no longer enough to force time-inconsistency).

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.[27]

Our model can also be applied to non-violent strategic behaviour, such as a hunger strike to force time-inconsistency.[28] In 1942, Mahatma Gandhi intensified his demand for Indian independence. Gandhi was soon arrested and responded by going on a hunger strike. British officials initially took an uncompromising stance: the Viceroy of India at the time, Lord Linlithgow, said he was "strongly in favour of letting Gandhi starve to death" rather than release him from prison. However, when Gandhi's health actually started deteriorating, officials changed their mind and released him after all. Sir Stafford Cripps, a member of Britain's Cabinet, said that Gandhi was "such a semi-religious figure that his death in our hands would be a great blow and embarrassment to us". Soon after, India achieved its independence.

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

[27] See http://www.theguardian.com/world/2003/aug/04/westafrica.qanda.

[28] Hunger strikers typically threaten to starve themselves to death unless their demands are met. As with kidnappers, however, this does not always seem to be their true intention: Scanlan, Stoll and Lumm (2008: 299) found that only 6 percent of all hunger strikes over the last century resulted in the death of the protester, while their demands were not met in at least 24 percent of all cases (usually, demands are only met in a partial way in response to which hunger strikers often end their actions already).

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.[29] 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.[30]

Finally, one may also want to consider a variant of the model where the dictator has 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 atrocities 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 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 12), we leave it for future work.

Relatedly, 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.

[29] Note 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.

[30] In 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.

6 Conclusion

The punishment strategy for malevolent dictators is complicated by a time-consistency problem: whereas authorities always want to threaten rulers with a harsh penalty for committing atrocities ex ante, there are strong incentives to forget about these threats while atrocities 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 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* atrocities 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 atrocities 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 with the dictator), than the pay-off that it expects to receive from sticking to its earlier promises to try the dictator. This gives rise to a "mimicking dictator group" which consciously chooses to worsen its behavior in the first period. 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 atrocities.

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".

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 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.

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 (2010), "The Logic of Political Violence", *Quarterly Journal of Economics*, 126 (3), pp. 1411-1445.

Besley, T. and T. Persson (2011), "Fragile States and Development Policy", *Journal of the European Economic Association*, 9 (3), pp. 371-398.

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.

Egorov, G. and K. Sonin (2005), "The Killing Game: Reputation and Knowledge in Non-Democratic Succession", CEPR Discussion Paper No. 5092.

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.

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.

Scanlan, S.J., L.C. Stoll, and K. Lumm (2008), "Starving for Change: The Hunger Strike and Nonviolent Action, 1906–2004", *Research in Social Movements, Conflicts and Change*, 28, 275-323.

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.

Schwarz, M. and K. Sonin (2008), "A Theory of Brinkmanship, Conflicts, and Commitments", *Journal of Law, Economics, and Organization*, 24 (1), pp. 163-183.

Shavel, S. (1993), "An Economic Analysis of Threats and Their Illegality: Blackmail, Extortion, and Robbery", *University of Pennsylvania Law Review*, 141 (5), pp. 1877-1903.

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

Wintrobe, R. (1998), *The Political Economy of Dictatorship*, New York: CUP.

Zimbardo, P. (2007), *The Lucifer Effect: Understanding How Good People Turn Evil*, New York: Random House.